Evaluating the reliability of a framework for mathematical activities using generalizability theory

Ali Bozkurt[1], Mehmet Fatih Özmantar [2] and Sibel Tutan Teskin[3]

[1]Gaziantep University, Faculty of Education, Gaziantep, Turkey; alibozkurt@gantep.edu.tr

[2]Gaziantep University, Faculty of Education, Gaziantep, Turkey; ozmantar@gantep.edu.tr

[3]Ministry of Education, İstanbul, Turkey; sibell27@outlook.com

Abstract

This study investigates the measurement reliability of the Framework for Mathematical Activities (FfMA), developed to assess the quality of activity scripts and their implementations. Utilizing Generalizability Theory, the measurement reliability of scores obtained from the FfMA tool was determined. Data were collected based on a descriptive survey model. In this context, four activity scripts and classroom implementation videos of these texts were requested from each of 20 middle school mathematics teachers. The data were scored independently by three raters using the FfMA tool. The scores obtained from the raters were analyzed using the EDUG 6.1e program. Findings indicate that the measurement reliability of the FfMA tool is considered reliable with a value of 0.78, and it does not fall below 0.70 even in scenarios with the minimum number of raters (2) in D studies. These coefficients suggest that the FfMA tool consistently measures the quality of mathematical activities.

Key words: Mathematical activity script, mathematical activity implementation, generalizability theory, reliability.

Introduction

In mathematics education, activity-based teaching holds the potential to influence students' mathematical success and prepare them to be independent and democratic thinkers (Noreen & Rana, 2019). However, to realize this potential, there is a need for high-quality activity scripts and implementations. A review of the literature reveals various frameworks for assessing the quality of mathematics instruction, such as the Mathematical Quality of Instruction [MQI] (LMTP, 2011), Teaching for Robust Understanding [TRU] (Schoenfeld, 2013), Classroom Assessment Scoring System [CLASS] (Pianta & Hamre, 2009), and Framework for Teaching [F*f*T] (Danielson, 2013). These frameworks are generally designed to determine the overall quality of a lesson and do not provide a detailed evaluation specifically focused on the quality of the instructional activities. On the other hand, studies specifically concerning the mathematical dimensions of activities often limit their focus to particular aspects like cognitive demand, purpose, and materials (Stein & Smith, 1998). However, considering only limited, and often isolated, aspect such as cognitive demand or purpose is insufficient for a comprehensive evaluation, as the quality of a mathematical activity scripts and implementation is influenced by numerous components. In order to fill this gap, Framework for Mathematical Activities (F*f*MA) (Bozkurt et al., 2023) was developed to be used as a feedback tool and to evaluate the quality of activity text and activity applications.

Framework for Mathematical Activities (F*f*MA)

F*f*MA is an assessment tool that can be used to determine the quality of activity scripts and implementations by considering them separately. Based on these assessments, it is intended to be used in such a way as to provide users with feedback on the strengths and improvement areas of the

mathematical activity text and the performance of the activity implementation. In order for F*f*MA to have a functional use as an assessment tool, concrete indicators and observable criteria were taken as a basis. In this way, F*f*MA is intended to serve for reliable scoring as well as valid results. The products that F*f*MA evaluates are the activity text and its implementation. The activity script is a concrete tool produced as a document found in various sources or prepared by the teacher himself/herself and has observable qualities. Implementation, on the other hand, occurs in the real classroom environment based on the interaction between the student-teacher-content triad and has observable characteristics. The dimensions and components of F*f*MA are illustrated in Figure 1.

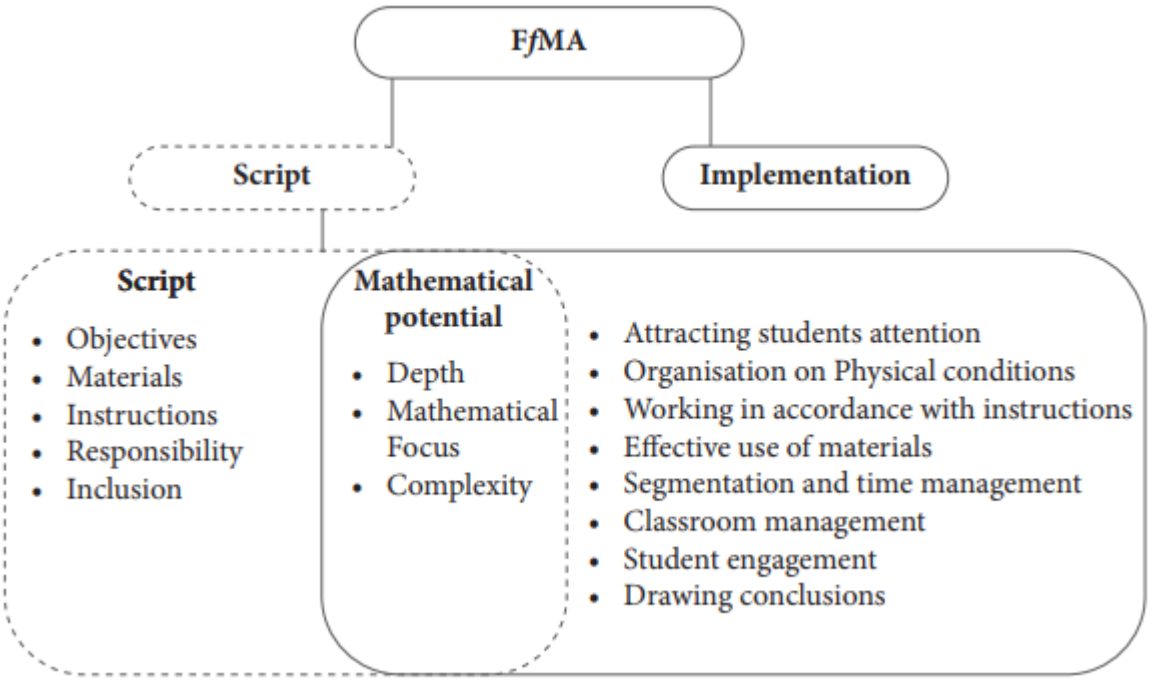

Figure 1. Dimensions and components of F*f*MA (Bozkurt et al, 2023)

As illustrated in Figure 5.1.1, the activity script has a total of 8 components and the implementation has a total of 11 components. Both dimensions, the activity text and the implementation, include an evaluation in terms of mathematical potential. Mathematical potential includes components related to determining the mathematical quality of the activity script and the implementation.

This study examines the measurement reliability of the Framework for Mathematical Activities (F*f*MA) (Bozkurt et al., 2023), which offers a broader perspective for evaluating activity texts and implementations. Utilizing the approach of Generalizability Theory, the study aims to determine the measurement reliability and generalizability of scores derived from the F*f*MA tool.

Method

This study stems from a project which was initiated, funded by TUBITAK (Project Number: #119K773). Spanning two years from 2020 to 2022 (Bozkurt et al., 2022). Within the scope of this study, the reliability study of the F*f*MA developed by the expert researchers in this project was conducted. In the literature, different theories have been developed to test the reliability of measurements obtained from a measurement tool: Generalizability Theory, Classical Test Theory, Multivariate Rash Model, Item Response Theory, etc. These theories differ according to the purpose of use, limitations, and how the measurement results are used (Brennan, 2001). Generalizability studies are organized to determine the source of variability from which measurement errors occur with a single analysis. Generalizability Theory, the chosen method for testing the reliability of F*f*MA,

is a statistical theory based on the analysis of variance (ANOVA). It is particularly useful in measurements involving different sources of error, allowing for the estimation of errors stemming from these sources and their interactions. Generalizability studies are divided into two types: G (generalization) studies and D (decision) studies. In G studies, the aim is to determine the variance components that affect reliability, as well as to generalize the measurement taken from the sample to the larger population (Brennan, 2001). D studies provide data for researchers to identify candidates for selection-placement, compare experimental groups, and investigate the relationship between at least two variables (Arterberry et al., 2012).

Data were collected from different schools and different grade levels during one semester in the 2022-2023 academic year. Four activity scripts and video records of their classroom implementations were collected from each of 20 middle school mathematics teachers. Participants were selected through purposive sampling method (Rai & Thapa, 2015). The teachers were asked to design activities or adapt an existing activity and implement it in their classrooms. They were also instructed to video record their implementations in actual settings. The decision regarding which lesson to record and when was left to the discretion of the teachers. The sample size was within the range of similar G-theory models and observational measurements (e.g., minimum 8) (Hill et al., 2012). This approach aligns with the methodological emphasis on the quality and applicability of the data rather than just the quantity, ensuring a more targeted and relevant evaluation of the F*f*MA tool's reliability.

The activity scripts and in-class video recordings of the implementation of these scripts were analyzed by the researchers. At the end of the analyzes, it was seen that in some videos, it was not possible to obtain healthy data on the student-teacher-content triad. These video recordings and the activity scripts used in these implementations were excluded from the evaluation. After this preselection process, 65 activity scripts and corresponding implementation videos were selected for further analysis. The data were independently scored by three raters using the F*f*MA. The components in the F*f*MA are graded on 4 score types (0: Very low; 1: Low; 2: Medium; 3: High). The scores that can be produced in an evaluation with F*f*MA are in the range of 0-24 points as minimum-maximum for the 8-component activity script dimension. For the 11-component activity implementation dimension, the minimum-maximum score range is 0-33 points. The scores obtained from the raters were then analyzed using the EDUG 6.1e program. In this framework, a G (generalizability) study was conducted following the pattern of "*activity script (a) x component interaction effect (c) x rater (r)*" and "*implementation (u) x component interaction effect (c) x rater (r)*". This study involved analyzing variance values for main and interaction effects. Subsequently, a D (decision) study was conducted to calculate the G coefficients for the reliability of the scores.

Findings

Generalizability of F*f*MA's Activity Script Dimension

The G (generalizability) study conducted for evaluating the activity scripts in F*f*MA revealed that the variance component attributed to the activity text source explained 9.7% of the total variance. This indicates that the components in the script tool are capable of distinguishing between different components of the activity script. The variance component estimated for the activity script-component-rater (*a x c x r*) or residual (unobserved or unintended) effect was 0.01736 and this effect explained 47.8% of the total variance. The second-largest source of variance was the interaction between component and rater (*c x r*), accounting for 37.1%. This variance indicates variability in the ratings given by different raters to different components. The interaction between activity script and component (*a x c*), contributing 5.4% to the variance, suggests that the difficulty levels and qualities of the components do not vary significantly from one script to another. Other sources of variance (*c, r, a x r*) were found to contribute zero or near-zero to the total variance, indicating their minimal impact on the overall variability in this context.

The D (decision) studies, which varied the number of raters and components in different scenarios, demonstrated that an increase in the number of raters leads to an improvement in reliability parameters. Based on the relative error variance, the G coefficient was found to be 0.78. In scenarios where the number of raters was three, and the number of components was eight or more, the G coefficients exceeded 0.80. It was also determined that even when the number of raters was reduced to two, the reliability parameters did not fall below 0.70. These findings highlight the robustness of the F*f*MA tool's reliability across different scenarios, emphasizing that even with a reduced number of raters, the tool maintains a satisfactory level of measurement reliability. This consistency in reliability, regardless of the number of raters, underscores the effectiveness of the F*f*MA framework in providing dependable evaluations of mathematical activity scripts.

Generalizability of F*f*MA's Activity Implementation Dimension

The G (generalizability) study focused on the evaluation of activity implementations within the F*f*MA framework revealed that the variance component attributed to the implementation of activities accounted for 5.7% of the total variance. This indicates that the components of the measurement tool are effective in distinguishing between the components of activity implementations. The most significant source of variance was the interaction between implementation, component and rater (*i x c x r*), accounting for 41%. This high level of variance suggests that the scores for the components of the activity implementations varied significantly due to the interaction effect and/or random errors, more than by common effects, from one rater to another and from one component to another. The second highest source of variance was the interaction between component and rater (*c x r*), accounting for 32%. This indicates that there is variability and inconsistency in the ratings given by different raters to different components of activity implementations. Interestingly, the variance component attributable to the raters alone (r) explained 0% of the total variance. This can be interpreted as an indication of the raters providing consistent scores, highlighting their reliability and uniformity in evaluating the activity implementations.

D studies were conducted by creating scenarios with varying numbers of raters and components. According to the results derived from the relative error variance, the G coefficient was found to be 0.78. It was observed that when the number of components increased while the number of raters remained constant, the G coefficient also increased. This suggests that the reliability parameters are affected by the variance in errors, which fluctuate depending on the number of raters and components involved. Even in scenarios where the number of components was reduced, the reliability parameters did not fall below 0.70, even when the number of raters was as low as two. This finding underscores the robustness of the F*f*MA tool's reliability, demonstrating that it maintains a significant level of accuracy and consistency across various testing conditions. It highlights that the F*f*MA is a reliable tool for evaluating mathematical activities, providing dependable results even with variations in the number of raters and components used in the assessment.

Discussion

According to the findings of the study, the generalizability (G) coefficient of the reliability of the measurements obtained with the F*f*MA tool was found to be 0.78. Even in decision studies where the minimum number of raters (two) was used, the reliability coefficient did not fall below 0.70. In the study where the number of raters was 3, G coefficients were above 0.80 in scenarios where the number of components was 8 or more. These coefficients indicate that the F*f*MA tool can produce consistent assessments in different scenarios (Brennan, 2001; Shavelson & Webb, 1991). This demonstrates that F*f*MA measures the quality of mathematical activities in a consistent and reliable way.

Another finding of the study is the residual variance effect, which has the highest source of variance. The raters' contribution to the shared variance is zero, indicating their consistency in scoring. This

finding shows that the quality of the activities varies from component to component and from rater to rater, and is more affected by the common effect and/or random errors. A similar finding was found in the study conducted by Solano-Flores (2006). In this study, G theory was used to estimate the amount of measurement error in the sources of variance of the instrument dealing with psychometric approaches to testing English language learners. Two groups were measured by giving questions in different dialects. The largest measurement error observed was due to the interaction of student, item and code (residual variance/random variance). The high residual variance may be due to various sources of variance not included in the design as well as the interaction effect (Uzun et al., 2018). Moreover, the decision studies reveal that reducing the number of raters does not significantly compromise the reliability of the tool. In fact, it was found that as the number of raters and components increases, so does the reliability coefficients. This finding is significant as it demonstrates the robustness of the F*f*MA tool in different evaluation contexts, ensuring that it remains a reliable measure for assessing the quality of mathematical activities, even with variations in the number of raters and components used.

Some of the coefficients obtained from the research were found to be low, which could be attributed to the closeness of the scores derived from the activities included in the study. This is because G coefficients are negatively affected by group homogeneity. According to Generalizability Theory, G and phi coefficients are calculated by relating universe variance to observed variance (Brennan, 2001). In other words, as the similarity between groups increases, the variance value decreases, which in turn could explain why the calculated G coefficients are not found to be high. A similar situation was observed in the study by Ozbası and Arcagok (2021), where they examined students' projects in a fully crossed design (*j x p x i*) within the framework of generalizability theory, considering jurors, projects, and items. In their research, the homogeneity among the projects led to a decrease in variance, thus affecting the generalizability coefficients in a similar manner. This underlines the impact of group homogeneity on the reliability measures in generalizability studies, suggesting that while the FfMA tool is reliable, the nature of the activities and the scoring patterns can influence the overall variance and, consequently, the generalizability coefficients.

As a result, considering the studies conducted, the reliability of this measure obtained from F*f*MA is acceptable. As a result of the decision studies, it was observed that the G coefficient increased more in scenarios where the number of raters was increased while the number of components was kept constant. This study contributes to the literature in terms of determining the factors affecting the reliability of the scores obtained from the evaluation of the quality of activity design and implementation in mathematics education through G theory. Based on our findings, it is suggested to provide teachers with guidance on how to evaluate and score mathematical activities. This can include training or resources that clarify the assessment process within the framework of the F*f*MA tool. Furthermore, to enhance the reliability of assessments, it could be beneficial to involve peer and expert evaluations of activity scripts and implementations across different student groups. By allowing for a diverse range of assessments, it would be possible to analyze the measurement results within the framework of Generalizability Theory more comprehensively. Such initiatives could not only improve the accuracy of evaluations but also provide valuable insights into the effectiveness of mathematical activities in diverse educational settings.

Acknowledgment

This research project was supported by Scientific and Technological Research Council of Turkey (TÜBİTAK 1001 no: 119K773)

References

Arterberry, B. J., Martens, M. P., Cadigan, J. M. & Smith, A. E. (2012). Assessing the dependability of drinking motives via generalizability theory. *Measurement and Evaluation in Counseling and Development, 45(4), 292-302.*

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

Bozkurt, A., Özmantar M.F., Agaç, G., & Güzel, M. (2022). Developing an Evaluation and Feedback Tool for Determining the Quality of Mathematical Task and Implementation. *Project Report, TÜBİTAK 1001 119K773.*

Bozkurt, A., Özmantar M. F., Agaç, G. & Güzel, M. (2023). *A Framework for Evaluating Design and Implementation of Activities for Mathematics Instruction.* Ankara: Pegem Academy.

Brennan, R. L. (2001). *Generalizability theory*. USA: Springer-Verlag New York Inc.

Danielson, C. (2013). *The Framework for teaching evaluation instrument,* 2013 edition. The Danielson Group.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64.

Karasar, N. (2008). *Scientific Research Method.* Ankara: Nobel Publishing House.

LMTP (Learning Mathematics for Teaching Project), (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14, 25-47

Noreen, R., & Rana, A.M.K. (2019). Activity-Based teaching versus traditional method of teaching in mathematics at elementary level. *Bulletin of Education and Research*, *41*(2), 145-159.

Ozbası, D. & Arcagok, S. (2021). Examining student projects with Generalizability Theory, *Journal of Theory and Practice in Education, 17*(2), 69-78. doi: 10.17244/eku.1024532

Pianta, R.C. & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119.

Rai, N., & Thapa, B. (2015). A study on purposive sampling method in research. *Kathmandu: Kathmandu School of Law, 5*(1), 8-15.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record, 108*(11), 2354-2379.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM,45*(4), 607–621.

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

Stein, M. K., & Smith, M. S. (1998). Mathematical tasks as a framework for reflection: From research to practice. *Mathematics teaching in the middle school, 3*(4), 268-275.

Uzun, N. B., Alıcı, D., & Aktas, M. (2018). Reliability of the analytic rubric and checklist for the assessment of story writing skills: G and decision study in generalizability theory. *European Journal of Educational Research, 8(1), 169-180.*