Eurasian Journal of Educational Research, Issue 66, 2016, 211-228

Using Generalizability Theory to Examine Different Concept Map Scoring Methods

Bayram CETIN[*]
Nese GULER[**]
Rabia SARICA[***]

Suggested Citation:

Cetin, B., Guler, N & Sarica, R. (2016). Using generalizability theory to examine different concept map scoring methods.*Eurasian Journal of Educational Research*, 66, 212-228
http://dx.doi.org/10.14689/ejer.2016.66.12

Abstract

Problem Statement: In addition to being teaching tools, concept maps can be used as effective assessment tools. The use of concept maps for assessment has raised the issue of scoring them. Concept maps generated and used in different ways can be scored via various methods. Holistic and relational scoring methods are two of them.

Purpose of the Study: In this study, the reliability of the concept map scores, which were made by the students and which were scored by different teachers using different scoring methods (holistic and relational), will be discussed in terms of G theory.

Methods: The research was performed during the fall semester of the 2010-2011 academic year, between December and January. Concept maps created by thirty-six students were scored by three different teachers who played roles as raters. Data were obtained from four different concept maps that were generated by each student.

Findings and Results: In focusing on the size of the variance estimates according to holistic scoring methods, while the student component (objects of measurement) accounts for one of the largest percentages of the

[*] Dr., Faculty of Education, Gazi University, Ankara, Turkey, bcetin27@gmail.com
[**] Corresponding author: Dr., Faculty of Education, Sakarya University, Sakarya, Turkey, gnguler@gmail.com
[***] Dr., Ministry of Education, Ankara, Turkey, rabiasarica@gmail.com

variance (20%), the main effects of the task and the raters account for about 14% and almost 0% of the total variance, respectively. The difficulty level of tasks did not differ so much from student to student, and there is a scoring agreement among raters. Using the holistic scoring method, G and Φ coefficients were calculated as 0.63 and 0.57, respectively, depending upon the four tasks and three raters. In terms of relational scoring, the student component (object of measurement) accounts for 10% of the variance, the main effect of the task accounts for a very significant percentage of the variance (56%), and the main effect of the raters does not demonstrate any variance. G and Φ coefficients calculated over the four tasks and three raters in the study were .63 and .34, respectively.

Conclusions and Recommendations: According to the results of this study, Phi coefficient was higher in the concept map study in which the holistic scoring method was used. In this study, tasks represented a significant variance component for both scoring methods. This may be interpreted to mean that the levels of difficulty for the tasks differed according to the students using both methods. In each of the scoring methods, the variance related to the raters was found to be zero, which may result in the interpretation that raters scored the maps consistently.

Keywords: Generalizability theory, rater effect, scoring concept maps, scoring methods

Introduction

Concept maps, which allow the visualization of concepts and show the relations between the concepts, are used to organize and present information in a graphical way. Generally, the concepts are written into the circles and square-like shapes, and the relationships between these concepts are shown by the use of arrows (Canas & Novak, 2006). Concept maps are an alternative method used to detect whether students understand a topic; through concept maps, students learn how to bridge the gap between learning issues and establish a meaningful learning. Also, it is an effective teaching strategy that involves active participation of students, which, in turn, gives students responsibility for their own learning (Kaptan, 1998; Nakhleh, 1994; cf. Kaya, 2003).

The basis of the concept map depend on Ausubel's (1962) meaningful learning. Novak (2010) stated that the theoretical basis of the concept map was established after the publication of Ausubel's Assimilation Theory of Meaningful Learning in 1963. According to Novak (2010), the key idea in Ausubel's theory is the distinction between rote learning and meaningful learning. In meaningful learning, the individual learns to apply knowledge to solve problems faced in real life, and to become adept at bringing information to the new learning. In short, it can be expressed as the ability to establish a relationship between prior and new learning. Information which is learned meaningfully becomes more permanent and serves to

solve the original problem, while allowing one to incorporate future learning along with creative thinking. An effective and economical method of providing meaningful learning in concept mapping studies has confirmed this idea (Novak, 2010).

The origin of the concept map depends on Novak and his research team's studies, set out in the 1970s at Cornell University, in a teaching process of 12 years, following changes in the methods through which students were introduced to science concepts (Misdates, 2009). Novak and Gowin's (1984) studies have been effective for recognition of concept maps all over the world (Ahlberg, 2004). Novak (2010) specified that they had been trying to determine why some students experience deep, meaningful learning while others develop just a superficial understanding.

Graphical maps of the concept in which information is schematized in a hierarchical structure are utilized in many different disciplines, especially in education, for different purposes, by both teachers and students at every stage of learning--in preparation of exams, various evaluation studies and course reviews (Kaptan, 1998; Kaya, 2003; Ingec, 2008). Novak (2001) suggested that concept maps can be used for educational purposes as well as for evaluation purposes. Additionally, the use of multiple-choice tests is not a necessity. Even in the context of national achievement exams over time, these tools may be used as effective assessment tools (cf. Kaya & Kılıc, 2004). Using concept maps in education for the purpose of evaluation of student achievement is very important in terms of revealing shortcomings related to learning, as they enable us to learn whether students understand topics correctly. Concept maps play a very central role in understanding a student's knowledge structure, mistakes and misconceptions on given subjects (Sahin, 2002). As hierarchical, two-dimensional diagrams showing how information is organized, concept maps are accepted as a valid means of evaluation and research, primarily in mathematics and science fields. In addition, it is noted that this technique may be used as a tool of both preliminary assessment and final assessment with regard to revealing, strengthening and consolidating information (Allen, 2006).

The first step to be taken before using concept maps as a means of scoring and evaluation is to assure that teachers have earned the required qualifications to use them. After providing adequate training to teachers and making sure that they have the necessary competence, concept maps can be effectively used as tools for evaluation. Additionally, scoring maps belong to students who have not gained convenient knowledge and skills about visualizing what they have learned, starting them with figures and making meaningful connections, potentially leading to incorrect assessment of the student. In such a case, it could be difficult to determine the student's deficiency resulting from a subject area or a lack of understanding of technique

Using concept maps as a tool for assessment has brought the issue of scoring them to the agenda. In order to use this method for the purpose of assessment, teachers need to understand rating methods very well. Concept maps generated and used in different ways can be scored using varied methods. McClure, Sonak and

Suen (1999) appraised the comparative point reliability of six different concept map scoring methods by calculating a generalization coefficient for each method. These six different scoring methods are holistic, holistic with criteria map, relational, relational with criteria map, structural and structural with criteria map.

In the holistic scoring method, concept maps are taken as a whole. Taking into account students' reflections on their learning with related concepts on the map, and the existence of the related concepts on the map, they are evaluated with points on a scale of 1 to 10. Sonak and Suen (1999) developed a relational scoring method, adopting a technic discovered by McClure and Bell (1990). The relational scoring method is based on the separate grading of propositions. The proposition of the relationship between the two concepts is indicated using a labelled arrow. The total score of the map is calculated by collecting the scores given to each of the propositions, and each proposition is scored on a point scale of 0-3, based on whether it is correct (McClure, Sonak & Suen, 1999).

The structural scoring method is developed by Novak and Gowin (1984). In this method of scoring, propositions, hierarchy, examples and cross-links are scored. According to this method, the total score is calculated by giving 1 point for each correct proposition, 5 points for the current levels of hierarchy, 10 points for accurate and meaningful cross-links where propositions are valid and 1 point for each sample (Nakiboglu & Ertem, 2010).While the structural scoring method focuses on organization of the hierarchical structure of the concept maps, the relational scoring method is based on the quality of each individual component of the map (West, Park, Pomeroy & Sandoval, 2002).

Modified forms of previously described holistic, relational and structural scoring methods include holistic with criteria map, relational with criteria map and structural with criteria map scoring methods. In these methods, maps are scored based on a concept map developed by an expert group on the subject, as well as on the criteria (McClure, Sonak & Suen, 1999). Although technical characteristics of concept maps become critical when used as tools for evaluation, the means through which to evaluate reliability and validity of the scores obtained is not always clear (Yin & Shavelson, 2008). Measuring instruments such as those used in scientific studies to produce reliable results are desired.

Generalizability (G) theory is a statistical theory based on variance analysis developed by Cronbach and his colleagues (1972). This theory provides for the assessment of reliability by bringing a different perspective to the concept (Shavelson & Webb, 1991 cf. Deliceoglu, 2009). G theory purports to generalize points obtained by means of specific measuring instruments to a larger universe of their sample (Guler, 2009). G theory provides for the calculation of a single reliability coefficient by incorporating all mistakes coming from all sources of variability at the same time, and additionally examining sources of mistakes individually, with interactions specified with the theory itself (Brennan, 2001; Tasdelen, Kelecioglu & Guler, 2010; Srikaew, Tanghanakanond & Kanjanawasee, 2015). If scores received by one of the students are considered an example of the universe of the concept map scores (under

varying conditions; for example, the task, response format, scoring methods and so on), then scoring of concept maps can be examined within the scope of G theory. In this respect, one of the reasons for using G theory is that there are many sources of errors in scoring of concept maps, and classical test theory cannot overcome the sources of these errors effectively (Yin & Shavelson, 2008)**.** Ruiz-Primo and Shavelson (1996) emphasized that the scoring of concept maps can lead to different error sources like concepts, propositions, task type, response formats, conditions, raters. Thus, using G theory is especially appropriate in this kind of research (cf. Yin & Shavelson, 2008). Additionally, many studies have investigated the inter-rater reliability of concept map scoring using G theory. For instance:

Kaya Uyanik and Guler (2016) conducted a study to demonstrate that G theory is preferable to classical test theory while investigating the reliability of concept map measurement results. The G and Phi coefficients were computed. Taking the results of the research into consideration, it may be recommended that the G and D studies based on G theory should be performed when determining the reliability of measurement results in which different sources of variability such as concept maps are available; this approach presents detailed and explanatory results with one single analysis, in contrary to classical test theory.

Canbazoglu Bilici, Dogan and Erduran Avci (2015) investigated the use of concept maps as an alternative assessment tool in Science and Technology courses. For this purpose, they used structural and relational scoring methods to evaluate the concept maps. Using the scores given by two raters, Pearson correlation and generalizability coefficients were calculated to determine inter-rater reliability. The results of Pearson correlation demonstrated that there were strong and statistically significant correlations between the raters for both scoring methods. Using generalizability theory, G coefficients were calculated and results suggest that both concept map scoring methods are valid and reliable.

Erduran Avci, Unlu and Yagbasan (2009) conducted a study to analyze the concepts of a 7th grade science course. They used concept maps as an assessment tool. The two raters scored student concept maps, and G theory was used to investigate the reliability. G coefficient was calculated as .97. In addition to G theory, Pearson moment multiplication correlation coefficient of inter-rater was calculated and was found to be .99 ($p < .01$). They stated that, according to these results, it can be said that the evaluation was reliable and valid.

Because G theory can be chosen, especially in cases in which there is more than one active source of variability, many raters exist or measurement is performed more than one occasion (Guler, 2011; Lakey, 2016). G theory was preferred to use for determining reliability. So in this study, reliability of scores of concept maps, which were made by students and which were scored by different teachers, will be discussed in terms of G theory. Two different concept map scoring methods are used within the scope of this research. These are holistic and relational scoring methods. Using just two scoring methods for concept maps can be seen as one of the constraints of the research.

Method

Study Group

The research was performed during the fall semester of the 2010-2011 academic year, between December and January. Participants consisted of thirty-six seventh-grade students whose ages ranged from 12 to 14, attending Ataturk Elementary School, Osmaniye, Turkey. Twenty-one of them were male, and fifteen of them were female. Information about the study group is also provided in Table 1.

Table 1.

Information about the Study Group

	Frequency	%
Female	15	42
Male	21	58
Total	36	100
Age Average	12.56	

Raters' Characteristics

In the study, concept maps created by students were scored by three different teachers who played roles as raters. Two of the raters were Science and Technology teachers, and the other was one of the researchers. Among the raters, two of them were female and one of them male. Teaching experience of the raters was 20, 16 and 5 years, respectively. The necessary training on concept maps and methods of scoring was provided by the researchers to the teachers. Science and Technology teachers stated that they benefited from this method, and there are some activities at the end of the guide books that they shared with their students.

Data Collection Tool

Data were obtained from four different concept maps that were used as data collection tools. The concept maps used in this study are related to a "*force and motion*" unit. Students had learned the topics of springs, force energy and power in actions, simple machines, and their concept maps related to these topics were scored. In the first of these concept maps, students created the concept map by themselves. In the second, some concepts were provided to students, and they were asked to build propositions and connections. In the third scenario, students chose missing concepts and connection sentences in the concept maps from the given alternatives. On the last concept map, students were asked to transfer to a concept map their knowledge about the topic before training. Teachers studied these concept maps together and examined the course books and necessary resources to make sure all of these topics were addressed, and they agreed on how to ask questions about the concept maps. For all of these reasons, structured and semi-structured concept maps were preferred.

Results

In this study, 36 students' proficiency with creating concept maps was scored through two different scoring methods by three raters. The scores obtained from these scoring methods were analyzed separately according to G theory using SPSS (Musquash & O'Connor, 2006), and the results and interpretation are explained below.

Analysis of Scores Obtained from Holistic Scoring Method According to G Theory

Students (s) in this study were the objects of the measurement, the concept maps were the sources of other variables tasks (t) and raters (r) were the facets of this study. In this study, students were responsible for creating all of the concept maps, and then all of the concept maps created by students were scored by raters via the holistic scoring method. Thus, the research design of this study is a fully crossed (s x t x r) design. According to this design, the results related to the estimated variance components are provided below in Table 2.

Table 2.

Analysis of Variance Results and Variance Component Estimates for Students, Tasks of Concept Maps, Raters and Their Interactions

Source of variance	*SS*	*df*	*MS*	*Variance Component Estimates*	*Percentage of Total Variance Estimates*
s	239.44	35	6.84	.360	.204
t	91.16	3	30.39	.240	.135
r	6.48	2	3.24	.002	.001
st	239.18	105	2.28	.616	.348
sr	46.69	70	0.67	.590	.034
tr	15.97	6	2.66	.062	.035
str	90.20	210	0.43	.430	.243

In Table 2, both the key elements of ANOVA table and the variance component estimates are observed. Because G theory focuses on the size of the variance component estimates, and not the statistical significance of the facets or their interactions, Table 2 does not include the significance test results (Goodwin and Goodwin, 1991). In addition, percentages of each variance component as part of the total variance appear in the last column of the table. Four sources of variation are relatively large compared to the others. The variance component for students, which indicates the variance for a student mean score over tasks and raters, accounts for about 20% of the total variance. This result demonstrates that students systematically differed in their level of proficiency at creating concept maps. A second significant component is tasks, which accounts for about 14% of the total variance. This relatively large component of the main effect of tasks indicates that tasks differed in difficulty level; some tasks were harder than others. A third significant component, students by task interaction, which accounts for about 35% of

the variance, shows that some students created some concept maps well and other students created other concept maps well. A fourth large component (24%), residual error, indicates a large student-by task-by-rater interaction, unmeasured sources of variation, or both. This value indicates that a substantial proportion of the variability is due to facets not included in the study and/or random error. According to G theory, this interaction variance value should be as low as possible.

The components of variance due to the rater effect and its interactions were relatively small. The main effect for rater (.001), the interaction between students and raters (.034), and the interaction between raters and tasks (.035) were near zero. These results demonstrate that raters similarly scored student concept maps. The implication of the small rater effect for future similar research is that single raters can provide dependable ratings. As a result, and as seen in Table 2, as an advantage of G theory, researchers can see very clearly which resources affect the total variance (Guler, 2009). In G theory, the coefficient of G equivalent reliability coefficient in classical test theory is calculated. The coefficient of G is calculated using the equation provided below;

$$G - coefficient = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{st}^2}{n_t} + \frac{\sigma_{sr}^2}{n_r} + \frac{\sigma_{str}^2}{n_t n_r}}$$

In G theory, in contrast with the classical test theory, Phi coefficient can also be calculated in the circumstance of certain assessment. In this calculation, tasks, raters and all interactive variance components are taken as parts of certain variance. The greater denominator is calculated by adding these to the denominator of the coefficient of Phi. Thus, when the obtained coefficient gets smaller, phi coefficient--called reliability coefficient--is calculated this way;

$$\Phi - coefficient = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_t^2}{n_t} + \frac{\sigma_r^2}{n_r} + \frac{\sigma_{st}^2}{n_t} + \frac{\sigma_{sr}^2}{n_r} + \frac{\sigma_{tr}^2}{n_t n_r} + \frac{\sigma_{str}^2}{n_t n_r}}$$

In this study, G and Φ coefficients were calculated as 0.63 and 0.57 and depended on four tasks and three raters. As can be understood from the equation, raters raised the reliability further. The low number of tasks in this study causes the reliability coefficient to be at a low level. In G theory, similar calculations to Spearman-Brown in classical test theory are possible. By means of this formula, when it is possible to change the number of items only in one test in classical test theory, G and Φ coefficient depend on the changing level of sources of variability which can be calculated with the D Study in the G theory. G and Φ coefficients in cases of changing number of raters in circumstances of certain number of tasks are provided below in Table 3.

Table 3.

G and Φ coefficients of D Studies (n_t: 4)

Raters	1	2	**3***	4	5
G-coeff.	.53	.60	**.63**	.65	.66
Φ-coeff.	.48	.54	**.57**	.58	.59

(*The number of raters in the study)

As seen in Table 3, an increased number of raters raise the reliability coefficient, but not so much. Therefore, raising the number of raters provides a positive contribution. In Table 4 below, G and Φ coefficients were calculated with number of raters settled as a constant and number of tasks as a variable.

Table 4.

G and Φ coefficients of D Studies (n_r: 3)

Tasks	4*	8	12	16	20
G-coeff.	.63	.76	.81	.84	.86
Φ-coeff.	.57	.71	.77	.81	.83

(*The number of tasks in the study)

As seen in Table 4, the increasing number of tasks raises the reliability. Therefore, if it is not possible to raise number of raters, and if it is possible to raise number of tasks, reliability increases. As can be seen in Table 3, twice the number of tasks raises the reliability coefficient by 0.07 when other circumstances are held as a constant. Therefore, in similar concept maps, using more tasks constitutes the study. In addition to Tables 2 and 3, Figure 1 clearly shows how increasing the number of tasks and raters affects the G and Phi coefficients simultaneously. According to Tables 3 and 4, together with Figure 1, it can be said that the number of tasks being increased should be more effective than increasing the number of raters.

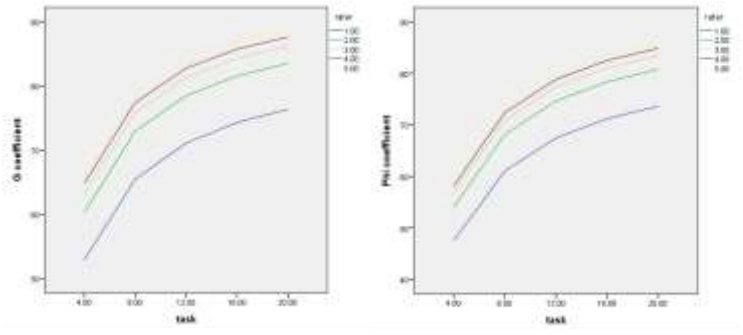

Figure 1. G and Phi coefficients for different number of tasks and raters

Analysis of Scores Obtained from Relational Scoring Method According to G Theory

The students (s) in the relational scoring method are the measurement object, just as in the holistic scoring method; the concept maps of other sources of variability in the tasks (t) and raters (r) are facets of the study. However, all of the students were responsible for creating concept maps, and these concept maps were scored by all raters together using the relational scoring method. Hence, this study is also a fully crossed (s x t x r) design. The patterns obtained by the analysis of variance and the generalizability results of the following components are provided in Table 5.

Table 5.

Analysis of Variance Results and Variance Component Estimates for Students, Tasks of Concept Maps, Raters and Their Interactions

Source of variance	*SS*	*df*	*MS*	*Variance Component Estimates*	*Percentage of Total Variance Estimates*
s	279.96	35	7.99	.421	.103
t	767.78	3	255.93	2.307	.564
r	1.56	2	.78	.000	.000
st	239.18	105	2.86	.813	.199
sr	299.89	70	0.51	.024	.006
tr	25.79	6	4.29	.108	.026
str	87.54	210	0.42	.417	.102

In Table 5, both key elements of ANOVA table and the variance component estimates are observed. When the results of Table 5 are compared to those of Table 2, similar findings can be seen. The variance component for students, which indicates the variance for a student mean score over tasks and raters, accounts for about 10% of the total variance. This result demonstrates that students systematically differed in their level of proficiency with creating concept maps. A second significant component is tasks, which accounts for about 56% of the total variance. This relatively large component of the main effect of tasks indicates that tasks differed in difficulty level; some tasks were more difficult than others. A third significant component, students by task interaction, which accounts for about 20% of the variance, shows that the relative standing of students in creating concept maps differed across tasks. A fourth large component (10%), residual effect, suggests a large student-by-task-by-rater interaction, unmeasured sources of variation, or both. The components of variance due to the rater effect and its interactions were relatively small. The main effect for raters was zero, and the interaction between students and raters and the interaction between raters and tasks were near zero (.006 and .026, respectively). Overall, more of the variability comes from tasks than from raters. These results show that raters similarly scored student concept maps. The implication of the small rater effect for future similar research is that a single rater can provide dependable ratings.

G and Φ coefficients calculated over the four tasks and three raters for this design were .63 and .34, respectively. Although one of the highest variances was among students as measurement objects, task main effect variance and its interactions with other effect variances were higher than for student main effect, which results in a decrease in value of the coefficient of Φ, adding this highest variance to the denominator in calculation of Φ. This study of concept maps used the scoring method in Table 6 below. The number of tasks is held as a constant, and in case of changing number of raters, estimated coefficient values are given in G and Φ.

Table 6.

G and Φ coefficients of D Studies (n_t: 4)

Raters	1	2	**3***	4	5
G-coeff.	.56	.61	**.63**	.64	.65
Φ-coeff.	. 31	.319	**.336**	.339	.342

(* the number of raters in the study)

As shown in Table 6, increasing the number of raters increases the value of the coefficient of Φ. For this reason, it can be noted that in the case of circumstances where more raters work, this can contribute to an increase in the coefficient Φ. The following Table 7 shows the estimated values of G and Φ in the circumstances in which the number of raters is held as a constant and the number of tasks changes.

Table 7.

G and Φ coefficients of D Studies (n_r: 3)

Tasks	4*	8	12	16	20
G-coeff.	.63	.77	.83	.86	.88
Φ-coeff.	.336	.501	.598	.663	.709

(*The number of task in the study)

Increasing the number of tasks increases the reliability value, as can be seen in Table 7. For this reason, if it is not possible to raise the number of raters in the study, increasing the number of tasks may contribute to the study. In addition to Tables 6 and 7, in Figure 2 it can be observed clearly how increasing the number of tasks and raters affects the G and Phi coefficients simultaneously. As seen in Tables 6 and 7, together, and Figure 2, it can be concluded that increasing the number of tasks should be more effective than increasing the number of the raters.

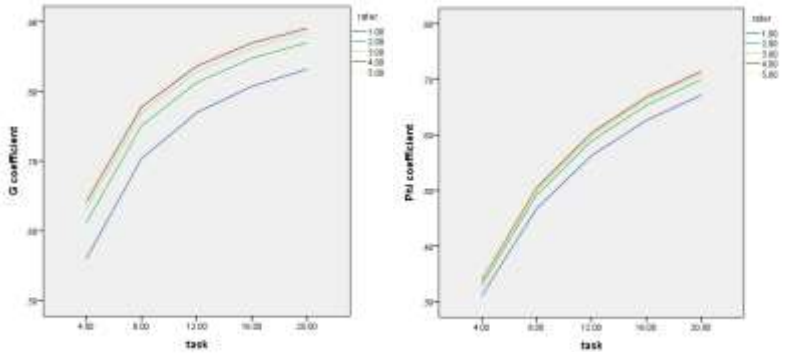

Figure 2. *G and Phi coefficients for different number of tasks and raters*

Discussion and Conclusion

According to the results of this study, G and Phi coefficients were higher in the concept map study in which the holistic scoring method was used, and estimated residual variance component (sxtxr) calculated using the relational concept map scoring method was higher. The proportion of task variance is 20% in the study in which the holistic scoring method was used, and the task variance component calculated using the relational scoring method accounted for about 56% of the total variance in scores. This may be interpreted as a result of the levels of difficulty of the tasks differing according to individuals when using the relational scoring method. In each of the scoring methods, the variance related to the raters was found to be almost zero, which may mean that raters scored the maps consistently in both scoring methods. On the basis of these results, it is suggested that holistic scoring method be used in evaluating concept map studies. In cases where the relational scoring method is used, it is advisable to make the students practice creating concept maps, offer more explanation to the raters and provide more details about scoring methods. In addition, according to the results of both scoring methods and based on high residual variance, it is recommended that students take a source of error in other external factors (environment, a measurement tool, test manager, etc.) in creating concept maps. Since the G coefficients are similar for both scoring methods, and the Phi coefficient is higher for the holistic scoring method than for the relational scoring method, if the aim of the study is to make an absolute decision, the holistic scoring method is recommended.

For future similar studies, it can be suggested that more tasks and fewer raters be used for reliable results. In this study, the "Force and Motion" unit in a Science and Technology course is discussed. The concept maps on different courses in different subjects and whether they provide reliable and valid results can be researched. In addition, the studies which include different and more sources of variability besides the sources of variability of the tasks and the raters in this study may be recommended.

References

Ahlberg, M. (2004).Varieties of Concept Mapping. *Concept Maps: Theory, Methodology, Technology.* Proc. of the First Int. Conference on Concept Mapping. A. J. Cañas, J. D. Novak, F. M. González, Eds. Pamplona, Spain 2004.

Allen, B.D. (2006). *Concept Map Scoring: Empirical Support for A Truncated Joint Poisson and Conway-Maxwell-Poisson Distribution Method.* Paper Presented at the 32nd Annual Meeting of The New England Mathematical Association of Two Year Colleges, Manchester.

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

Canbazoglu Bilici, S., Dogan, A. & Erduran Avci, D. (2015). Using concept maps as an alternative assessment tool and investigation by comparing with multiple choice tests. *Kastamonu University, Journal of Kastamonu Education*, 23(3), 1031-1046.

Cañas, A. J., & Novak, J. D. (2006). *Re-Examining the Foundations for Effective Use of Concept Maps.* Paper presented at Proc. of the Second Int. Conference on Concept Mapping, Costa Rica.

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

Deliceoglu, G. (2009). *The comparison of the reliabilities of the soccer abilities' rating scale based on the classical test theory and generalizability.* Unpublished Doctoral Dissertation, Ankara University, Ankara.

Erduran Avci, D., Unlu, P. & Yagbasan, R. (2009). Using concept maps as a method of assessment in work-energy subject. *Journal of Applied Sciences*, 9(3), 427-439.

Guler, N. (2011). The comparision of reliability according to generalizability theory and classical test theory on random data. *Education and Science*, 36, 162, 225-234.

Guler, N. (2009). Generalizability theory and comparison of the results of G and D studies computed by SPSS and GENOVA Packet Programs. *Education and Science*, 34(154), 93-103.

Ingec, S. K. (2008). Using concept maps as an assessment tool in physics education. *Hacettepe University Journal of Education, 35*, 195-206.

Kaptan, F. (1998). The use of concept map technique in science education. *Hacettepe University Journal of Education, 14*, 95-99.

Kaya, O. N. & Kilic, Z. (2004). *Student-Centered Reliability, Concurrent Validity and Instructional Sensitivity in Scoring of Students' Concept Maps in a University*

Science Laboratory. Poster presented at 18th International Conference on Chemical Education "Chemistry Education for the Modern World", Istanbul.

Kaya, O. N. (2003). An alternative way of assessment in education: Concept maps. *Hacettepe University Journal of Education, 25*, 265-271.

Kaya Uyanik, G., & Guler, N. (2016). Investigation of concept map scores' reliability: Example of crossed mixed design in generalizability theory. *Hacettepe University Journal of Education*, 31(1), 97-111.

Lakey, B. (2016). Understanding the P X S Aspect of Within-Person Variation: A Variance Partitioning Approach. Frontiers in Psychology. Doi: 10.3389/fpsyg.2015.02004.

McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity and logistical practicality. *Journal of Research in Science Teaching*, 36(4), 475-192.

Misdates, V., M. (2009). Concept mapping in introductory physics. *Journal of Education and Human Development, 3*(1), 1-6.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS Programs for generalizability theory analysis. *Behavior Research Methods*, 38 (3), 542-547.

Nakiboglu, C., & Ertem, H. (2010). Comparison of the structural, relational and proposition accuracy scoring results of concept maps about atom. *Journal of Turkish Science Education,* 7(3), 60-77.

Novak J.D. (2010). Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations. *Journal of e-Learning and Knowledge Society, 6*(3), 21 - 30.

Novak, J.D. & Gowin, R. (1984). *Learning how to learn*. New York: Cambridge University Pres.

Novak, J. D., & Cañas, A. J. (2008). The Theory Underlying Concept Maps and How to Construct and Use Them, *Technical Report IHMC* Cmap Tools 2006-01 Rev 01-2008, Florida, Institute for Human and Machine Cognition, 2008, available at: http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf

Ruiz- Primo, M .A, & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, 33(6), 569-600.

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

Srikaew, D., Tangdhanakanond, K., & Kanjanawasee, S. (2015). English speaking skills assessment for grade 6 thai students: an application of multivariate generalizability theory. *Scientific Publications*. http://dx.doi.org/10.7220/2345-024X.16.3

Sahin, F. (2002). A research on usage of concept maps as an evaluation tool. *Pamukkale University Journal of Education, 11*(1), 17-32.

Tasdelen, G., Kelecioglu, H., & Guler, N. (2010). A comparison of scores obtained by nedelsky ve angoff cutting score procedures with generalizability theory. *Journal of Measureument and Evaluation in Education and Psychology*, 22-28.

Yin, Y., & Shavelson, R. J. (2008). Application of generalizability theory to concept map assessment research. *Applied Measurement in Education,* 21, 273–291.

West, D.C, Park, J.K., Pomeroy, J.R., & Sandoval, J. (2002). Concept mapping assessment in medical education: A comparison of two scoring systems. *Medical Education, 36*, 820–826.

Kavram Haritalarının Puanlanmasında Puanlayıcı ve Puanlama Yöntemi Etkisinin Genellenebilirlik Kuramıyla İncelenmesi

Atıf:

Cetin, B., Guler, N & Sarica, R. (2016). Using generalizability theory to examine different concept map scoring methods.*Eurasian Journal of Educational Research*, 66, 212-228
http://dx.doi.org/10.14689/ejer.2016.66.12

Özet

Problem Durumu: 1970'lerde ortaya konan kavram haritaları, bilginin hiyerarşik bir düzen içerisinde şematize edilerek görselleştirilmesini sağlayan grafiksel araçlardır.Kavram haritaları eğitimde bir konudaki kavramlar arasındaki ilişkinin daha açık, anlamlı öğrenilmesini sağlamaya yardımcı olabilecek araçlardır. Novak (2001), kavram haritalarının öğretim amaçlı kullanılabildiği gibi değerlendirme amaçlı da kullanılabileceğini, çoktan seçmeli testlerin kullanılmasının bir zorunluluk olmadığını ve hatta zamanla ulusal başarı sınavlarında bu araçların etkili bir değerlendirme aracı olarak kullanılabileceğini belirtmiştir (Akt: Kaya ve Kılıç, 2004). Kavram haritalarının eğitimde değerlendirme amaçlı olarak kullanılması, öğrencilerin konuyu anlayıp anlamadıklarını göstermesi ve öğrenme ile ilgili eksiklerini ortaya çıkarması açısından çok önemlidir. Kavram haritaları, öğrencinin

bilgi yapısını, konuyla ilgili yanılgılarını ve yanlış anlamalarını belirlemede oldukça fonksiyonel bir işleve sahiptir (Şahin, 2002). Kavram haritalarının değerlendirme aracı olarak kullanılması bunların puanlanması konusunu gündeme getirmiştir. Bu yöntemin değerlendirme amaçlı olarak kullanılabilmesi için öğretmenler tarafından puanlama yöntemlerinin çok iyi bilinmesi gerekmektedir. Farklı şekilde oluşturulan ve kullanılan haritalar farklı yöntemlerle puanlanabilmektedir. Bu yöntemlerden iki tanesi bütüncül ve ilişkisel puanlama metotlarıdır. Bütüncül puanlama yönteminde kavram haritaları bir bütün olarak ele alınır, öğrencilerin kavramlarla ilgili öğrenmelerini haritaya yansıtabilmeleri ve ilgili kavramların haritada yer alması göz önünde tutularak 1-10 arasında bir puanla değerlendirilir. İlişkisel puanlama yöntemi önermelerin ayrı ayrı puanlanması temeline dayanmaktadır. Önerme iki kavram arasındaki ilişkinin etiketlenmiş bir ok aracılığıyla gösterilmesi olarak tanımlanır. Haritanın toplam puanı, ayrı önermelerin her birine verilen puanların toplanmasıyla bulunmaktadır ve her bir önerme doğru olup olmadıklarına göre 0-3 arasında bir puan almaktadır (McClure, Sonak ve Suen,1999). Kavram haritası, değerlendirme aracı olarak kullanıldığında teknik özellikleri kritik hale gelmesine rağmen, elde edilen puanların güvenirlik ve geçerliliğinin nasıl değerlendirileceği her zaman net değildir (Yin ve Shavelson, 2008). Genellenebilirlik (G) kuramı, temeli varyans analizine (ANOVA) dayanan güvenirliğin değerlendirilmesini sağlayan, Cronbach ve arkadaşları (1972) tarafından geliştirilen, güvenirlik kavramına farklı bir bakış açısı getiren istatistiksel bir kuramdır (Shavelson ve Webb, 1991 Akt; Deliceoğlu, 2009). Öğrencilerden birinin aldığı puan kavram haritası puanlarının evreninden bir örnek olarak düşünülürse (değişen bütün koşullar altında örneğin; görev, cevap formatı ve puanlama metotları vb.) kavram haritalarının puanlanması G kuramı kapsamında incelenebilir. Ruiz-Primo ve Shavelson, (1996) kavram haritası puanlamasının; kavramlar, önermeler, görev tipi, cevaplama formatları, durumlar, puanlayıcılar ve puanlama yöntemleri gibi farklı hata kaynakları içerdiğinden, bu tür araştırmalarda G kuramının kullanılmasının bilhassa uygun olduğunu belirtmiştir (Akt: Yin ve Shavelson, 2008).

Araştırmanın Amacı: Bu çalışmada, farklı öğretmenler tarafından puanlaması yapılan öğrencilerin oluşturduğu kavram haritalarının puanlarının güvenirlikleri G kuramı açısından ele alınacaktır. Bu araştırma kapsamında kavram haritası puanlama yöntemlerinden ikisi kullanılmıştır. Bunlar; bütünsel (holistik) puanlama ve ilişkisel puanlama yöntemleridir. Kavram haritalarının puanlanmasında sadece bu iki yöntemin kullanılabilmiş olması araştırmanın sınırlılıklarından biri olarak görülebilir.

Araştırmanın Yöntemi: Araştırma, Osmaniye ili Merkez Atatürk İlköğretim okulunda 7.sınıfta öğrenim görmekte olan 15'i kız, 21'i erkek olmak üzere 36 öğrenci ile gerçekleştirilmiştir. Araştırma 2010-2011 eğitim-öğretim yılı güz dönemi Aralık-Ocak ayları içerisinde gerçekleştirilmiştir. Araştırma kapsamında öğrencilerin yapmış olduğu kavram haritalarını üç farklı öğretmen puanlamışlardır. Veriler, veri toplama aracı olarak kullanılan dört farklı kavram haritasından elde edilmiştir. Bu çalışmada kullanılan haritalar Kuvvet ve Hareket ünitesiyle ilgilidir.

Araştırmanın Bulguları: Çalışmada 36 öğrencinin dört kavram haritası oluşturabilme düzeyleri iki farklı puanlama yöntemiyle üç puanlayıcı tarafından puanlanmıştır.

Her bir puanlama yöntemine göre elde edilen puanlar G kuramına göre ayrı ayrı analiz edilmiş ve elde edilen sonuçlar yorumlanmıştır.

Bütünsel puanlamada, çalışmada yer alan öğrenciler (s) ölçmenin objesi olup, diğer değişkenlik kaynakları olan kavram haritaları görevleri (t) ve puanlayıcılar (r) da çalışmanın yüzey (facet)lerini oluşturmaktadır. Bu çalışmada tüm öğrenciler tüm kavram haritalarını oluşturmakla sorumlu olduklarından ve tüm puanlayıcılar tarafından bütünsel puanlama yöntemiyle puanlandıkları için çalışma tümüyle çaprazlanmış (s x t x r) desenden oluşmaktadır. Genellenebilirlik analiziyle elde edilen varyans bileşenlerine ilişkin sonuçlara göre,en büyük değişkenlik kaynaklarından birinin öğrenciler olduğu görülmüştür (gerçek varyans). Diğer ana etkiler olan görev, toplam varyansı açıklayan en büyük bileşenlerden biri olurken (yaklaşık %14), puanlayıcı bileşeni toplam varyansın açıklanmasına nerdeyse hiç bir katkıda bulunmamaktadır (%001). Etkileşimlere baktığımızda öğrenci-görev bileşeni toplam varyansın yaklaşık %35'ini açıklarken, görev-puanlayıcı etkileşimi toplam varyansın çok küçük bir kısmını açıklamaktadır (%034). Üçlü etkileşimin, bir başka deyişle artık etkisinin, toplam varyansdaki payı ise %24'tür. G kuramına göre, artık etkisine ilişkin varyans değerinin olabildiğince küçük olması istenir. Bu değer, puanlardaki değişimin çalışmada yer almayan farklı değişkenlik kaynaklarına bağlı ortaya çıkmış olabileceğinin sinyalini vermektedir. G kuramında, klasik test kuramındaki güvenirlik katsayısına karşılık gelebilecek G katsayısı hesaplanmaktadır. G kuramında, klasik test kuramından farklı olarak bir de mutlak değerlendirmenin söz konusu olduğu durumlar için ayrıca Phi katsayısı (reliability coefficient) da hesaplanabilmektedir. Yukarıdaki eşitliklere dayalı olarak, çalışmada yer alan dört görev ve üç puanlayıcı üzerinden hesaplanan G ve Φ katsayısıları sırasıyla .63 ve .57 olarak bulunmuştur.

İlişkisel puanlama yönteminde de aynı desen kullanılmış ve yine en büyük değişkenlik kaynaklarından birinin öğrenciler olduğu görülmüştür (%10). Görev ana etki bileşeni, toplam varyansı açıklayan en büyük bileşen olurken (yaklaşık %56), puanlayıcı bileşenin toplam varyansın açıklanmada bir payı bulunmamaktadır (%000). Diğer taraftan ikili etkileşimlere bakıldığında öğrenci-görev, öğrenci-puanlayıcı ve görev-puanlayıcı etkileşimleri sırasıyla yaklaşık %20, %0 ve %03 olarak elde edilmiştir. Buradan anlaşılacağı üzere, kavram haritalarında yer alan görevlerin zorluk düzeyleri öğrenciler için farklılık gösterirken, öğrencilerin ve görevlerin puanlanması puanlayıcıdan puanlayıcıya farklılık göstermemektedir. Üçlü etkileşimler artık etki olarak isimlendirilir ve eğer çalışmada, ölçme sonuçları güvenilir ise artıklara ait olan bu değerin olabildiğince küçük olması istenir. İlişkisel puanlama yönteminin kullanılarak elde edilen puanlar üzerinden bulunan artık etki varyansı toplam varyansın %10'unu açıklamaktadır. Elde edilen bu varyans değeri, puanlardaki değişimin çalışmada yer almayan farklı değişkenlik kaynaklarına bağlı ortaya çıkmış olabileceğinin sinyalini vermektedir. Çalışmada yer alan dört görev ve üç puanlayıcı üzerinden ilişkisel puanlama yöntemi için hesaplanan G ve Φ katsayısıları sırasıyla .63 ve .34 olarak bulunmuştur.

Araştırmanın Sonuç ve Önerileri: Elde edilen sonuçlara göre, her ik puanlama yöntemi için G katsayısı aynı bulunmuşken, Phi katsayısı bütünsel puanlama yönteminin kullanıldığı kavram haritası çalışmasında daha yüksek bir değere sahiptir. Bu sonuçlara dayanarak mutlak kararkların alınması amaçalanan kavram haritası

çalışmalarında, bütünsel puanlama yöntemini kullanmak önerilebilir. İlişkisel puanlama yönteminin kullanılacağı durumlarda ise öğrencilerin kavram haritalarını oluşturmada daha fazla pratik yapması ve puanlayıcılara puanlama konusunda daha fazla açıklama yapılması ve puanlama ölçütlerinin daha ayrıntılı verilmesi önerilebilir. Ayrıca, her iki puanlama yöntemiyle elde edilen sonuçlara göre, artık varyansın yüksek çıkmasına dayalı olarak, öğrencilerin kavram haritası oluşturulmasında hata kaynağı olabilecek diğer dış etkenlerin (ortam, ölçme aracı vb.) de dikkatlice kontrol altına alınması gerektiği önerilmektedir.

Anahtar sözcükler: Genellenebilirlik kuramı, puanlayıcı etkisi, kavram haritalarının puanlanması, puanlama yöntemleri.