International Journal on New Trends in Education and Their Implications
January 2016 Volume: 7 Issue: 1 Article: 03 ISSN 1309-6249

COMPARISON OF INSTRUCTOR AND SELF-ASSESSMENTS ON PROSPECTIVE TEACHERS' CONCEPT MAPPING PERFORMANCES THROUGH GENERALIZABILITY THEORY

Assist. Prof. Dr. Göksu GÖZEN
Mimar Sinan Fine Arts University
İstanbul- TURKEY

Assist. Prof. Dr. Kaan Zülfikar DENİZ
Ankara University
Ankara- TURKEY

Abstract

The purpose of the research is to examine, according to generalizability theory, a) the consistency between instructor assessments and self-assessments on concept mapping performance of 100 secondary school prospective teachers who attended Pedagogical Formation Certificate Program at Mimar Sinan Fine Arts University in 2014-2015 academic year, b) their severity-leniency behaviors in these assessments, and c) the difficulty levels of the performance criteria used in these assessments. Generalizability study was carried out by creating a *p x c x r* (p: person, c: criterion, r: rater) pattern obtained through scoring of the designed concept maps by the prospective teachers and instructor on the same assessment form. The following were found out based on the findings; - instructors and prospective teachers exhibited equal severity-leniency in scoring both throughout the assessment criteria and by comparison, - performance criteria were distributed across the different difficulty levels, - there was no consistency between prospective teachers' self-assessments and the instructor assessments, and – self-assessments of prospective teachers were more positive compared to the instructor assessments.

Keywords: Instructional material design; concept map; performance assessment; self-assessment; generalizability theory.

INTRODUCTION

In line with the competencies targeted in the 21st century teaching processes; it has become obligatory to implement principles such as learning through experience, reading, listening, gaining oral and written skills, providing solutions to problems through scientific approaches, making proof-based inferences and generalizations through research and analysis. The most important factor for the students to exhibit the aforementioned skills, undoubtedly, is the teachers who will enable the student to gain those skills. Therefore the literature focusing on teacher education is expanding on daily basis. As Avalos (2011) mentioned, the core of the recent scientific research on the professional development of teachers is not only concerned to teachers' transforming their knowledge into practice for the benefit of students' growth but also teachers' learning processes, particularly their gaining knowledge and experience in the instructional methods and techniques as well as domain-specific and differential strategies and tools of assessment and evaluation to judge the quality of education.

Over the last three decades, the most common used assessment tools have been objective tests (i.e. with multiple choice items) in almost all education systems over the world as a means of measuring and monitoring the quality of education. Stecher (2010) and Chen & Brown (2013) stated that the nature of these tools do not reflect the nature of performance in the real world, therefore they are not well suited to judging students' ability to express points of view, marshal evidence, and display other advanced skills since they have not focused primarily on the higher-order thinking and performance skills. In a similar way, Darling-Hammond & Adamson (2010) reported the evidence, which suggests that the nature and format of the assessments affects the depth of knowledge and types of skills developed by students, and that performance assessments are better suited to assessing high level, complex thinking skills. Thus, within the studies comparing objective and essay exams with several kinds of alternative strategies for educational assessment and evaluation (i.e. İngeç, 2009; Moreira, 2006; Rafferty & Fleschner, 2010) , particularly based on performing association, inference and interpretation skills, concept mapping is proposed as a viable pedagogical tool for meaningful learning and understanding in almost every grades of education. It is defined as a way to represent knowledge schematically through establishing the most prominent and most useful cross-links (relationships) between several concepts, which involves what Bloom (1956) identified as high levels of cognitive performance, namely evaluation and synthesis of knowledge (Edmondson, 2000). This indirect method of observation is also supposed to be effective in reaching the goal of turning out people who can think and generate (İngeç, 2009; Novak & Cañas, 2007; Strautmane, 2012).

Literature contends that a teacher's knowledge of concept map-based instruction and/or assessment influence how their students perceive the instructional content and execute students' creation of acceptable concept maps to present their own way of learning. Thus, teachers need to understand the educational function of concept mapping in relation to the nature and quality of the graphical structures of such practices and in terms of how these structures impact and/or effectuate learning. Accordingly, Subramaniam & Esprívalo Harrell (2015) stated that teachers who are skilled concept mappers are able to (1) understand and apply the operational terms to construct a hierarchical/non-hierarchical concept map; (2) identify the legitimacy of the constructed concept map by verifying its graphical structure and its educational utility; and (3) determine the inherent 'good' and 'poor' qualities of the resulting graphical structure to reiterate the 'good' qualities and to coach and provide feedback to alleviate 'poor' qualities. Behind numeorus studies focusing on teachers' knowledge and competencies to design concept maps or other several instructional materials (e.g. Novak & Cañas, 2008; Strautmane, 2012; Yelken & Alıcı, 2008; Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, 2005), in the present study this subject is handled from a different viewpoint, only which is established as a principle focus in limited number of studies (Cronbach, Linn, Brennan, & Haertel, 1997; Dochy, Segers, & Sluijsmans; 1999; Jimenez-Snelson, 2010; McClure, Sonak, & Suen, 1999; Plummer; 2008; Yin & Shavelson, 2008), and the consistency between the self-assessments of prospective teachers' and assessments performed by their instructor on their concept mapping performances is investigated in conjunction with raters' severity or leniency behaviours in the assessments and the difficulty of the criteria used in the assessments. Hence, apart from having the necessary knowledge and skills to contemplate qualified teaching and assessment materials that serve different purposes and meet different needs, teachers should have high order cognitive and affective skills of self-assessment, which, as a matter of fact, is a prerequisite for possession of qualified teaching competency. Individuals, of course, attain the vision and responsibility to see their deficiencies in their work or in the process of gaining the necessary technical and practical knowledge and skills related with that work and to compensate these deficiencies only through evaluating themselves objectively. From this point of view, the present study also takes prospective teachers' self-assessment of their utilization level of the technical information learned in the process of designing the concept maps as an important dimension.

It is desired that all the assessment tools used and developed in all scientific research have a high reliability. As for performance based assessment, in which students must construct an answer, produce a product, or perform an activity rather than choosing among pre-determined options, reliable scoring becomes more important. Different variability sources mingled with the assessments and the interaction between these different sources are quite important from reliability standpoint (Brennan 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In this context, historically, reliability issues in psychology and education have been addressed principally using Classical Test Theory (CTT), which postulates that an observed score can be decomposed into a true score (T) and a single, undifferentiated random error term (E).On the other hand, Generalizability Theory (G-Theory) that handles different sources of error and their interaction together and simultaneously and that liberalizes CTT as stated by Brennan (2001) is one of the useful methods not only in identifying the validity and reliability of different assessment tools (multiple choice tests, performance assessment tools, etc.), but also in making comparisons to see the consistency between the evaluations when there are more than one rater involved in the assessment (Atılgan, 2005; Güler, 2009, 2011; Nunally, 1982;

Nunally & Bernstein, 1994; Volpe, McConaughy & Hintze, 2009; Yelboğa, 2012; Yılmaz-Nalbantoğlu &Gelbal, 2011). The present study examined, based on the G-Theory, a) the consistency between instructor assessments and prospective teachers' self-assessments on the concept maps the prospective teachers prepared, b) severity-leniency behaviors in the assessments, and c) difficulty levels of the criteria used in the assessments.

METHOD

Research Design

In CTT, reliability coefficients having different meanings are obtained through different reliability methods for the same assessment. Considering these different measures related with the reliability estimation, Cronbach, Nageswari, & Gleser (1963) suggested the Generalizability Theory (G-Theory) both as an extension of and as a flexible alternative to CTT. This conceptual and statistical framework, for evaluating the dependability (reliability) of behavioral measurements (e.g., a test score) evaluates all sources of variance, e.g. rater, time, items, setting etc., in other words, all possible sources of error (in G-Theory, sources of variation are referred to as *facets of measurement*) that may occur in an assessment process together and simultaneously and therefore tests the generalizability of the sample drawn from the universe of admissible observation to the universe (Brennan, 2001; Sanders, 2014; Webb, Shavelson & Haertel, 2006). The detailed information on every source of variance and their interaction makes G-Theory the recommended (Atılgan, 2005; Nunally & Bernstein, 1994) approach which enables making inter- and cross- comparisons in assessments that involve more than one rater.

Individuals (persons) were defined as object of measurement, and not considered sources of variance as in many assessments due to the fact that they are the targets of the assessment activity and process, and also that their variance is natural and systematic. On the other hand, each assessment criteria for the concept maps as a teaching tool and raters (instructor and prospective teachers making self-assessment) were considered a source of variance (facet) that can have an impact on individuals' universe scores (these scores correspond to true scores in CTT). Each prospective teacher and instructor prepared and assessed their material according to the same 10 criteria. However, considering that these 10 criteria were picked randomly among the universe of admissible observation that can be used to measure the concerned feature and that the raters were randomly sampled among the universe of raters who are able to assess the performance of the prospective teachers on these criteria, it may be suggested that the research has a crossed two-facet random-effects design of G-Theory. Accordingly, the research pattern is symbolized as *p x c x r*, prospective teachers as *p* (persons), each assessment criteria *c* (criterion) and raters *r* (raters).

Variance components in G–Theory are derived from many sources like systematic variance of individuals that are the objects of measurement, multiple variance sources and the corresponding interaction between them (Crocker & Algina, 1986; Goodwin, 2001). These probable sources of variance are estimated together and simultaneously based on the variance analysis (ANOVA). In this study, in order to determine what portion of the total variance in the results arises from which source or the interaction of facets, ANOVA equations were used in line with *p x c x r* pattern and seven variance components listed below were obtained;

- three main effects – persons (*p*), criteria (*c*) and raters (r),
- three common impacts - person-criterion (*p x c*), person-rater (*p x r*) and criterion-rater (*c x r*), and
- remainder effect (*p x c x r*, e).

Instead of focusing on a specific measurement result or the score observed, G-Theory focuses on generalization of measurement results to the sample universe which is wider than a specific sample; more specifically, it concentrates on the effect of different dimensions of the universe on the test scores. In G-Theory, this effect is shown with generalizability coefficients similar to reliability coefficients in actual score model. Thus, instead of "reliability" which is one of the main concepts of CTT, "generalizability" which is a broader and flexible term in G-Theory is used (Güler, 2009, 2011; Anıl & Büyükkıdık, 2012). G-Theory enables estimating the reliability levels based on different sources of variance, i.e. test-retest reliability, internal consistency, reliability among raters etc. through one study and also facilitates reliability estimation of not only relative decisions as in CTT, but also of absolute decisions which focus on the level of an individual's performance independent of others' performance (cf. domain-referenced interpretations) (Shavelson & Webb, 1991; Yin & Shavelson, 2008).

Thereby different reliability coefficients, G and Phi (�) (dependability), are produced based on two different decision making types which are relative and absolute. In the present study, both coefficients were used to obtain the indicators of reliability among raters (instructor assessments and prospective teachers' self-asssessments).

Participants

The participants were composed of 100 prospective secondary school teachers, who were attending the Pedagogical Formation Certificate Programme at Mimar Sinan Fine Arts University in the academic year 2014-2015, of whom 23 are male ve 77 are female, and one educator who was their instructor in the spring semester of this academic year and lecturing the course entitled "Instructional Technologies and Material Design", within which the prospective teachers designed the concept maps. The distribution of the prospective teachers by their specialties which constitute the basis for their assignment to the teaching profession are presented in Table 1.

Table 1: Distribution of prospective teachers by their specialties

Specialty	n
Turkish Filology	36
History	23
Visual Arts	14
Philosophy Group (Sociology, Pyschology and Philosophy)	11
Mathematics	10
Health Sciences	4
Physical Training and Sports	2
Total	100

Research Instrument and Data Collection

In this research, 100 prospective secondary school teachers were asked to draw a hierarchical concept map related to their teacher roles in the course entitled "Instructional Technologies and Material Design" after they were taught about the meaning, content, and construction of different kinds of concept maps and given several examples. Hierarchical concept maps that were designed by prospective teachers were scored by themselves and the instructor simultaneously and independently on a score sheet containing ten performance criteria in total with 0-1-2 scores. While scoring guidelines were being determined, criteria that required exhibition of minimal grammatical competence but good conceptual understanding (based on an effective visualing the sub-concepts that makes up a main concept) were taken into consideration. A moderation discussion that supports prospective teachers using and understanding the use of grade criteria was carried out prior to scoring. For each prospective teacher, two scoring series comprised of instructor assessments and prospective teachers' self-assessments were obtained following use of the score sheet. Criteria, scores and definitions used in the assessments are listed in Table 2.

Table 2: Criteria, scores and definitions used in the assessments of hierarchical concept mapping

Criteria	Scores and Definitions		
	2	1	0
1.Legibility and clarity	Main concept is easily distinguished from other concepts on the map.	Main concept, even if discernible, is not sufficiently distinguished from other concepts on the map.	Main concept within the concept graphical structures is indistinguishable.

2. Hierarchical arrangement	There is a hierarchical arrangement between the root concept and subordinate concepts and a coherent design can be observed between the concepts at the same level.	Even if there is a hierarchical arrangement between the root concept and subordinate concepts, a coherent design cannot be observed.	Hierarchical arrangement and coherent design cannot be observed between the root concept and subordinate concepts.
3.Relevance and Systemacity/Order	The root concept and subordinate concepts have relevant and systematic/orderly relations.	Some of the relations indicated between the root concept and subordinate concepts are not relevant and systematic/orderly.	Relevant and systematic/orderly relations cannot be established between the root concept and subordinate concepts.
4.Visual links	Visual links are established between concepts, directions of visual links are correctly set, and connections are defined with coherent verbs and conjunctions.	Although there are visual links between concepts, direction of them are not set, and unclear and incognizable statements are seen on some links.	Concepts maps only contain linking lines but lack direction, linking phrases, labelled lines and propositions.
5.Exampling	Examples are provided for each concept.	Examples are provided for some concepts.	No example was provided for the concepts.
6.Use of material	The content is enriched with the support of different visual materials (pictures, drawings, photographs, cartoons, designing with various materials, etc.).	Some part of the content is supported by different visual materials (pictures, drawings, photographs, cartoons, designing with various materials, etc.).	The content is not supported by different visual materials (pictures, drawings, photographs, cartoons, designing with various materials, etc.), only written elements are available.
7.Suitability	Conceptual graphical structures are completely suitable for the development and learning level of the target student group.	Although very few, concepts that are unsuitable for the development and learning level of student group are observed in the conceptual graphical structures.	Conceptual graphical structures are not suitable for the development and learning level of target student group.
8.Consistency	Conceptual graphical structures, are consistent with cognitive attainments available under the scope of the related the class/subject.	Some sub concepts in the conceptual graphical structures are not covered in the cognitive attainments available under the scope of the related class/subject.	Conceptual graphical structures, are not consistent with cognitive attainments available under the related the class/subject.
9.Design features	A map that is appropriate for all design features (color, size, highlight, etc.) to allow easier reading of conceptual graphical structures is prepared.	Some design features are ignored, and this decreases readability in some parts of the conceptual graphical structures.	Design features were ignored in such a way that it makes conceptual graphical structures unreadable.
10.Form and purpose of use	How and for what purpose (teaching, repetition, enhancement, practice, assessment and	How and for what purpose (teaching, repetition, enhancement, practice, assessment and evaluation,	How and for what purpose (teaching, repetition, enhancement, practice, assessment and

	evaluation, etc.) the concept map will be used is defined in detail.	etc.) the concept map will be used is superficially defined, and there are some vague areas.	evaluation, etc.) the concept map will be used is not defined.

There is more than one technique used in determining the reliability of the abovementioned tools and the like utilized in performance based assessments. Some of these techniques provide the consistency level between the raters over the total scores attained by individuals for a specific performance, while the others address approach each performance criterion separately. In the context of this study, both techniques were utilized to study the reliability of the tool provided in Table 2. As a basis of the reliability study, concept map development performance of 50 prospective teachers out of 100 who formed the study group was evaluated simultaneously and separately by the instructor responsible for the class in which concept maps were developed and a second instructor observing the class. Under the scope of the techniques that reveal the consistency level between the raters over the total performance scores, the correlation technique was used, and the level of the relationship between the two instructors' assessments on concept map development performances of 50 prospective teachers was found to be r_{xy}= 0.96 (p<.001). This value was considered to indicate the consistency/coherence between the assesssments of the two instructors on prospective teachers' performances. In order to obtain separate reliability indicators for each criterion, Cohen's Kappa formula (Krippendorff, 2004) was used, which is a coefficient that provides more precise information than simple percentage consistency calculation as it also takes consistency percentage obtained by chance into consideration and is one of the non-parametric statistic types used for categorical variables. Symmetrical assessment values obtained regarding the criteria on the score sheet are listed in Table 3.

Table 3: **S**ymmetric measures for assessment criteria

Criteria	**Measure of Agreement Kappa**	**p**
Legibility and clarity	0.315**	.002
Hierarchical arrangement	0.728***	.000
Relevance and Systemacity/Order	0.401***	.000
Visual links	1.000***	.000
Exampling	1.000***	.000
Use of material	0.791***	.000
Suitability	0.865***	.000
Consistency	1.000***	.000
Design features	0.739***	.000
Form and purpose of use	0.742***	.000

**p<.01 ***p<.001

Kappa consistency measure values presented in Table 3 show to what extent the consistency among the fixed number of raters is not random, and thus, suggest a high consistency among the raters if it is proximate to 1.00 (Reynold, Livinston & Wilson, 2006). According to this, each criterion used to evaluate the performance of prospective teachers allows both instructors to score objectively and nonfortuitously; in other words, it is suitable for producing a reliable scoring. On the other hand, the content validity of the assessment tool was tried to be ensured through the opinions of two measurement and evaluation experts.

Data Analysis

In the present study, analyses based on the consistency between the two instructors in determining the reliability of the tool used for evaluating the prospective teachers' performance in designing hierarchical concept maps were conducted using SPSS 20.00 software. In analyses under the scope of G-Theory to investigate the consistency between instructor assessments and prospective teachers' self-assessments, the statistical software EduG 6.1 was used.

FINDINGS

Instructor assessments and prospective teachers' self-assessments were examined correlatively, and the findings of the generalizability study conducted to obtain information about scoring consistency, assessment behavior and difficulty of performance criteria are provided in Table 4.

Table 4: Estimated variance components and percentages for prospective teachers' instructional concept map designs by ANOVA

Source*	Sum of Squares	df	Mean Square	Estimated Variance Component	Total Variance (%)
p	93.69	99	0.95	0.003	0.7
c	85.37	9	9.49	0.032	7.5
r	32.26	1	32.26	0.029	6.8
p x c	379.33	891	0.43	0.134	31.2
p x r	61.24	99	0.62	0.046	10.7
c x r	25.32	9	2.81	0.027	6.2
p x c x r, e	141.18	891	0.16	0.158	37.0

* p: person, c: criteria, r: rater, e: error

Even though the analyses in generalizability studies are based on random-effects factorial ANOVA, as also suggested by Shavelson & Webb (1991) and Brennan (2001), this concept has nothing to do with the hypothesis test. Therefore, there are no F and p values in Table 4.

When Table 4 is analyzed, it is observed that the variance component of 0.003 estimated for person (p) main effect has the smallest share of variance and that it explains only 0.7% of the total variance. This variance component that is for the universal scores suggests that the individuals do not differ systematically from each other in terms of the characteristics assessed, in other words, the prospective teachers do not differ in performance of designing concept maps and that they exhibit similar performance levels. This variance component estimated for individuals is the universal score variance which corresponds to the actual score variance in CTT and therefore, the value is desired to be greater.

The value of the estimated variance component value for criteria (c) used in evaluating the prospective teachers' hierarchical concept mapping performance is 0.032 which explains 7.5% of the total variance. Accordingly, it may be suggested that difficulty levels of some criteria differ from others, in other words, there are criteria that have difficult and easily achieved/fulfilled contents. Based on this finding, the criteria that are the most difficult and the easiest to achieve in terms of the assessments made by the instructor and the prospective teachers were examined separately, and the breakdown of the assessments conducted with scores 0, 1 and 2 for each criterion is shown in Table 5.

Table 5: Breakdown of criteria scores based on instructor assessments and prospective teachers' self-assessments

Criteria	Score Distribution for Instructor Assessments (n)			Total	Score Distribution for Prospective Teachers' Self-Assessments (n)			Total
	0	1	2		0	1	2	
Legibility and clarity	2	2	96	100	1	1	98	100
Hierarchical arrangement	13	20	67	100	-	6	94	100
Relevance and Systemacity/Order	6	8	86	100	-	-	100	100
Visual links	7	3	90	100	8	4	88	100
Exampling	25	27	48	100	16	12	72	100

Use of material	38	25	37	100	12	17	71	100
Suitability	9	18	73	100	5	4	91	100
Consistency	10	14	76	100	9	8	83	100
Design features	28	17	55	100	3	2	95	100
Form and purpose of use	4	7	89	100	4	1	95	100

When Table 5 is analyzed, “Use of Material” and “Design Features” were observed to be the most difficult criteria to achieve based on the instructor assessments where respectively 38 and 28 prospective teachers scored 0 (zero). Accordingly, the easiest criteria to achieve are “Legibility and Clarity” and “Visual Links” where respectively 96 and 90 prospective teachers obtained full scores (2 points). On the other hand, based on their self-assessments, the most difficult criteria the prospective teachers mentioned that they hard time achieving were “Exampling”, and consistently with the instructor assessments “Use of Material” where respectively 16 and 12 prospective teachers evaluated themselves with 0 (zero) points. The criteria that they achieved the most easily were “Relevance and Systemacity/Order”, and again consistently with the instructor assessments “Eligibility and Clarity” where respectively all prospective teachers (100 individuals) and 98 prospective teachers evaluated themselves with full scores (2 points). However, it is important to note here that “Exampling” and “Use of Material” criteria, which were considered as two of the most difficult criteria by both the instructor and the prospective teachers, were scored 2 full points by the instructor for less than 50% of the prospective teachers as can easily be observed from a review of the row, while these criteria were scored 2 full points by 70% of the prospective teachers in self-assessment. Difficulty levels appear to differ from each other throughout the criteria and at the same time this difference is also seen during the independent assessments made by instructors and prospective teachers reciprocally.

Raters (*r*) main effect shows whether or not the severity-leniency levels of raters scoring all the individuals differ. Table 4 shows that the estimated value for this variance component is 0.029 and that the total variance explained by this component is 6.8%. This value being proximate to zero indicates that the instructor and the prospective teachers themselves treat the prospective teachers with equal severity-leniency.

Persons-criteria common effect (*p x c*) shows whether or not a specific individual’s relative position (performance criteria) changes from one subject to another. The estimated value for this variance component is 0.134 and total variance explained by this component is 31.2%. Hence, the component based on the common interaction ranks second as per the total percentage it explains, in other words it has a significant place in the ranking due to its magnitude. Correspondingly, it is suggested that relative positions (performance criteria) of prospective teachers may differ from one criterion to another.

Persons-raters common effect (*p x r*) shows whether or not the raters and prospective teachers scored more severely-leniently by comparison. Table 4 shows that the estimated value for this variance component is 0.046 and total variance percentage explained this component is 10.7%. Accordingly, it may be suggested that raters’ scores may differ from one individual to another and, in other words, that individuals rated high by a rater is rated low by the other. Hence, in addition to what is presented in Table 4, when instructor assessments and prospective teachers’ self-assessments are thoroughly examined, it is found that prospective teachers’ assessments ($\bar{X}_{PT} =$ 18.29) are more positive compared to those of the instructor ($\bar{X}_{I} =$ 15.75), and that the difference [$t_{(198)}$=6,421, p<.001] between the assessments is significant. Hence, in parallel to this finding, Yılmaz-Nalbantoğlu’s study (2012), based on the findings of the t test used to compare students’ self-assessment scores to those of the rater, suggested that students had a significant tendency for a relatively more positive self-assessment.

Criteria-rater common effect (c x r) shows whether or not raters’ scores are stable from one criterion to another. In fact the value calculated for this variance component is 0.027 and that the total variance explained by this component is 6.2%. The component value being proximate to zero may suggest that the scoring made by both the instructor and the prospective teachers are stable to a great extent even though there is a minor difference from one criterion to another.

Remainder effect variance (*p x c x r*, e) are comprised of persons-criteria-rater common effect and random errors. Table 4 shows that the highest variance value of 0.158 belongs to this component with which 37.0% of the total variance is explained. The important place of the explained variance in the rank order serves as an indicator of the magnitude of the remainder effect. The magnitude of this effect not only means that the difference between the individuals' scores is caused by the criteria and the raters, but also shows the possibility of variability due to the factors other than the ones in the research design. Hence, in addition to interpreting the main and common effects based on the relative magnitudes of variances obtained by G-study, the present study conducted a reliability estimation in respect of instructor assessments and prospective teachers' self-assessments on hierarchical concept map performances based on G and Phi (⍰) coefficients, and G coefficient was found to be 0.06, and Phi (⍰) coefficient, which produces a stricter value, was found to be 0.05. In other words, it may be suggested that the magnitude of remainder effect variance is caused by the inconsistency between instructor assessments and prospective teachers' self-assessments on concept maps. Hence, while the observations based on Table 5 explain this finding to some extent, the correlation analysis used to obtain an extra measurement of the aforementioned consistency showed a low level of relationship (r_{xy}= 0.24, $p<.05$) between the instructor assessments and prospective teachers' self-assessments. It is possible to come across studies with similar findings in the literature; for example Longhurst & Norton (1997) also found a correlation coefficient value (0.43), which was not high, in their research on the correlation between instructor and student scorings.

DISCUSSION AND CONCLUSION

Under the scope of the present study, a generalizability study was carried out by creating *p x c x r* (p: person, c: criterion, r: rater) pattern through scoring of the prospective teachers' concept maps by the instructor and the prospective teachers themselves. Analysis based on estimated variance values revealed the following:

- In terms of their performance on concept map design, prospective teachers are homogeneously dispersed.
- Difficulty levels of the criteria used in the assessment vary. However, while the difficulty levels of the criteria vary throughout the criteria, this variance is also reciprocally observed in the instructor's and the prospective teachers' independent assessments.
- An equal severity-leniency is observed in both the assessments made by the instructor for all prospective teachers and the self-assessments made by the prospective teachers.
- Prospective teachers' performance measurements differ from one criterion to another; in other words some prospective teachers were found to exhibit high performance in some criteria and low performance in some other criteria.
- Scores given to concept map design performances of the prospective teachers vary based on the rater, and self-assessments are more positive compared to the instructor assessments.
- Raters were found to be consistent (stable) in scoring in respect of the criteria in the scoring sheets.
- The difference between the prospective teachers' performances was found to have resulted from the criteria and the rater.

Another indicator that this difference was caused particularly by the raters is the low G and Phi (⍰) coefficients (respectively 0.06 and 0.05) and the low correlation coefficient (0.24) between the scores of the raters. Considering the reliability evidence obtained based on the Kappa coefficient for each criterion of the scoring key that was initially found to be fit for reliable scoring based on the results obtained with regard to the fact that two different raters performed consistent scoring, one of the possible and the most important reasons why we cannot observe consistency between the raters is believed to the prospective teachers' incapability to make a self-assessment which requires an objective self-evaluation behavior. On the other hand, as Ewing and Everett (2015) stated, from a realistic point of view, it is stipulated that the instructors have overall responsibility for the final agreed grade from a quality assurance perspective. On the opposite, people grading themselves usually tend to move out of the grade band. However, it is also known that self-assessment skills can be acquired and applied competently in adulthood if they are experienced as of childhood, and that especially where the education system is not supportive of the development of such skills starting from childhood, individuals assess themselves either as very low or very high performer when they become an adult. Considering that acquiring these skills began to be emphasized following the revisions after 2005 and were not specifically emphasized among the main targets of our national Turkish education and training programs before this date, it is a usual to see the inconsistency between prospective teachers' self-assessments and the instructor assessments on prospective teachers' concept map design performances. The drawback here is that the prospective teacher who is incapable of self-assessment will be incompetent in defining and compensating own deficiencies related with the technical and applied information on his/her related field. Besides, another reason for the difference between the instructor assessments and the prospective teachers' assessments throughout the criteria may be that the prospective teachers may not have comprehended the content of assessment criteria on concept map design performances well enough or that the prospective teachers may have not paid due care to the importance of aforementioned self-assessment practice. For example "Use of Material" criteria that increases the readability and attractiveness of the concept maps and the number of sense organs in learning was found to have one of criteria with the highest difficulty level (the most difficult) as mutually determined by the instructor and the prospective teachers. But on the other hand, this common opinion differs at a very important point; while more than 70% of the prospective teachers defined the "Use of Material" as the criterion that is achieved with the highest performance level, the instructor stated that only 37% of the group was able to achieve this criterion at high performance. Based on the above, it is suggested that the initiatives supporting the development of self-assessment skills that will help teachers and prospective teachers define and compensate their deficiencies may also assist them with acquiring sufficient technical knowledge and skills that are the foundation for developing quality teaching and assessment materials that support acquisition of high order mental skills such as concept maps.

BIODATA AND CONTACT ADDRESSES OF AUTHORS

Assist.Prof.Dr. Göksu GÖZEN is a full-time academic staff of Educational Measurement and Evaluation in the Department of Educational Sciences at Mimar Sinan Fine Arts University. She received BA (2000), and MA (2002) degrees in Educational Measurement and Evaluation at Hacettepe University, and PhD (2007) degree in the same area at Ankara University. Her research interests include developing measurement and assessment tools, test scoring methods, teaching and assessing higher order thinking skills (i.e. creative thinking and problem solving skills), project-based learning, and performance assessment. She has also published in the area of ethics in educational and psychological testing process..

Assist.Prof.Dr. Göksu GÖZEN
Mimar Sinan Fine Arts University
Department of Educational Sciences
Cumhuriyet Mah. Silahşör Cad. No:71
34380 Bomonti, Şişli, İstanbu-l TURKEY
E-Mail: goksu.gozen@msgsu.edu.tr

Assist.Prof.Dr. Kaan Zülfikar DENİZ is a full-time Assistant Professor at Ankara University, Ankara, Turkey. He received BS (2001), and MS (2003) degrees in Measurement and Evaluation at Hacettepe University , Ankara, Turkey. He has completed his PhD dissertation (2008) at the Measurement and Evaluation, Ankara University, Turkey. His research interests include Scale development and measurement of affective properties. In the last 13 years he has worked as a full-time academic staff at Ankara University. He is a founding director of Ankara University Examination Management Center (ASYM).

Assist.Prof.Dr. Kaan Zülfikar DENİZ
Ankara University

Institute of Educational Sciences
06590 Cebeci, Ankara- TURKEY
E. Mail: zlfkrdnz@yahoo.com

REFERENCES

Anıl, D. ve Büyükkıdık, S. (2012). Genellenebilirlik kuramında dört facetli karışık desen kullanımı için örnek bir uygulama. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 3*(2), 291-296.

Atılgan, H. (2005). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama (Generalizability theory and a sample application for inter-rater reliability). *Eğitim Bilimleri ve Uygulama (Educational Sciences and Practice)*, *4*(7), 95-108.

Avalos, B. (2011). Teacher professional development in *Teaching and Teacher Education* over ten years. *Teaching and Teacher Education*, *27*(1), 10-20.

Bloom, B. S. (1956). Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain. New York: David McKay Co Inc.

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

Chen, J., & Brown, G. T. L. (2013). High-stakes examination preparation that controls teaching: Chinese prospective teachers' conceptions of excellent teaching and assessment. *Journal of Education for Teaching: International Research and Pedagogy*, *39*(5), 541–556. doi:10.1080/02607476.2013.836338

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Javanovich College Publishers, USA.

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

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. Educational and Psychological Measurement, *57*(3), 373–399.

Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology, 16*, 137-163.

Darling-Hammond, L. & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher education*, *24* (3), 331-350.

Edmondson, K. M. (2000). Assessing science understanding through concept maps. In Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (Eds.). Assessing science understanding: A human constructivist view (pp. 15-40). San Diego, CA: Academic Press.

Ewing, A. & Everett, S. (2015). Combining self-, peer- and tutor-assessment. Outside the Box Assessment & Feedback Practices, *1*(1), 9.

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

Güler, N. (2009). Genellenebilirlik kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması (Generalizability theory and comparison of the results of G and D studies computed by SPSS and GENOVA packet programs.). *Eğitim ve Bilim (Education and Science)*, *34*(154), 93-103.

Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması (The comparison of reliability according to generalizability theory and classical test theory on random data). *Eğitim ve Bilim (Education and Science)*, *36*(162), 225-234.

İngeç, Ş. K. (2009). Analysing concept maps as an assessment tool in teaching physics and comparison with the achievement tests. *International Journal of Science Education*, *31*(14), 1897-1915.

Jimenez-Snelson, L. (2010). Estimating the reliability of concept map ratings using a scoring rubric based on three attributes of propositions. Doctor of Philosofy Dissertation, April 2010. Brigham Young University, Department of Instructional Psychology and Technology. BYU ScholarsArchieve. Available at: http://scholarsarchieve.byu.edu/etd

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411-433.

Longhurst, N., & Norton, L. S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation*, *23* (4), 319-330.

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), 475-492.

Moreira, M. (2006). Concept mapping: An alternative strategy for evaluation. *Assessment & Evaluation in Higher Education*, *10*(2), 159-168.

Novak, J. D. & Cañas, A. J. (2007). Theoretical origins of concept maps, how to construct them, and uses in education. *Reflecting Education*, *3*(1), 29-42.

Novak, J. D. & Cañas A. J. (2008). The theory underlying concept maps and how to construct and use them. Technical Report IHMC CmapTools 2006-01 Rev 01-2008, Florida Institute for Human and Machine Cognition. Available at: http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf

Nunally, J. C. (1982). Reliability of measurement*. Encylopedia of Educational Research*, (5th edition). Editor H.E. Mitzel. New York.

Nunally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd edition). New York: Mc-Graw-Hill.

Plummer, K. J. (2008). Analysis of the psychometric properties of two different concept-map assessment tasks. Doctor of Philosofy Dissertation, April 2008. Brigham Young University, Department of Instructional Psychology and Technology. BYU ScholarsArchieve. Available at: http://scholarsarchieve.byu.edu/etd

Rafferty, C. D. & Fleschner, L. K. (2010). *Concept mapping: A viable alternative to objective and essay exams*. Reading Research and Instruction, *32*(3), 25-34.

Sanders, P.F.(2014). Generalizability theory: Estimation. Wiley StatsRef: Statistics Reference Online: http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat06695

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

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Strautmane, M. (2012). Concept map-based knowledge assessment tasks and their scoring criteria: An overview. In Concept maps: Theory, methodolgy, technology. Proceedings of the Fifth International Conference on Concept Mapping, Valletta, Malta, 2012.

Subramaniam, K., & Esprívalo Harrell, P. (2015). An analysis of prospective teachers' knowledge for constructing concept maps. *Educational Research*, *57*(3), 217-236. DOI:10.1080/00131881.2015.1050845.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay (Eds.), Handbook of Statistics, Vol. 26 (pp. 81-124).

Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of classroom behavior problem and on-task scores from the direct observation form. *School Psychology Review*, *38*(3), 382-401.

Yelboğa, A. (2012). Dependability of job performance ratings according to generalizability theory. *Education and Science*, *37*(163), 157-164.

Yelken, T. Y., & Alıcı, D. (2008). Öğretmen adaylarının hazırladıkları performansa dayalı değerlendirme materyallerine ilişkin görüşlerinin ve değerlendirmelerinin incelenmesi. *Jounal of Qafqaz University*, *24*, 222-235.

Yılmaz-Nalbantoğlu, F ve Gelbal S. (2011). İletişim becerileri istasyonu örneğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi,* 41, 509-518.

Yılmaz-Nalbantoğlu, F. (2012). İletişim becerileri istasyonundan elde edilen öz değerlendirme ve puanlayıcı değerlendirmelerinin karşılaştırılması. *Eğitim ve Öğretim Araştırmaları Dergisi*, 2, 357-363.

Yin, Y., & Shavelson, R. J. (2008) Application of generalizability theory to concept map assessment research. *Applied Measurement in Education, 21*(3), 273-291.

Yin, Y., Vanides, J., Ruiz-Primo, M. A., Ayala, C. C., & Shavelson, J. R. (2005). A comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching*, *42*(2), 166–184.