PEGEM EĞİTİM ve ÖĞRETİM DERGİSİ

PEGEM JOURNAL OF EDUCATION and INSTRUCTION

Article Type:
Research Paper

Original Title of Article:
A comparison of the reliability of the Solo- and revised Bloom's Taxonomy-based classifications in the analysis of the cognitive levels of assessment questions

Turkish Title of Article:
Değerlendirme sorularının bilişsel düzeylerinin tespitinde Solo ve revize edilmiş Bloom Taksonomisi'ne dayalı sınıflamaların güvenilirliklerinin karşılaştırılması

Author(s):
Mustafa ILHAN, Melehat GEZER

For Cite in:
İlhan, M., & Gezer, M. (2017). A comparison of the reliability of the Solo- and revised Bloom's Taxonomy-based classifications in the analysis of the cognitive levels of assessment questions. *Pegem Eğitim ve Öğretim Dergisi, 7*(4), 637-662, http://dx.doi.org/10.14527/pegegog.2017.023

Makale Türü:
Özgün Makale

Orijinal Makale Başlığı:
A comparison of the reliability of the Solo- and revised Bloom's Taxonomy-based classifications in the analysis of the cognitive levels of assessment questions

Makalenin Türkçe Başlığı:
Değerlendirme sorularının bilişsel düzeylerinin tespitinde Solo ve revize edilmiş Bloom Taksonomisi'ne dayalı sınıflamaların güvenilirliklerinin karşılaştırılması

Yazar(lar):
Mustafa ILHAN, Melehat GEZER

Kaynak Gösterimi İçin:
İlhan, M., & Gezer, M. (2017). A comparison of the reliability of the Solo- and revised Bloom's Taxonomy-based classifications in the analysis of the cognitive levels of assessment questions. Pegem Eğitim ve Öğretim Dergisi, 7(4), 637-662, http://dx.doi.org/10.14527/pegegog.2017.023

PEGEMAKADEMİ

Pegem Eğitim ve Öğretim Dergisi, 7(4), 2017, 637-662

www.pegegog.net

A Comparison of the Reliability of the SOLO- and Revised Bloom's Taxonomy-Based Classifications in the Analysis of the Cognitive Levels of Assessment Questions

Mustafa İLHAN [*a], **Melehat GEZER**[a]

[a]Dicle University, Ziya Gökalp Faculty of Education, Diyarbakır/Turkey

Article Info

DOI: 10.14527/pegegog.2017.023

Article History:
Received 06 March 2017
Revised 03 August 2017
Accepted 15 August 2017
Online 05 September 2017

Keywords:
SOLO Taxonomy,
Revised Bloom's Taxonomy,
Generalizability Theory.

Article Type:
Research paper

Abstract

This study aims to compare the reliability of SOLO and revised Bloom's taxonomy-based (RBT) classifications in the determination of the cognitive levels of the assessment questions. The data were obtained by three experts' examination of the assessment questions in the Seventh Grade Social Studies Textbook published by the Ministry of National Education and Eight Grade Science and Technology Textbook published by Yıldırım Publishing in 2015. The collected data were analyzed in a crossed design whereby the assessment questions were the object of measurement and the experts were the facet on the basis of the generalizability theory. It was found that the variance percentage of the main effect of the assessment question was found to be higher in the SOLO-based classifications. The variance component related to the experts' main effect and the residual variance values were found to be higher in the RBT than in the SOLO taxonomy. In conclusion, compared to the RBT, the SOLO-based classifications were found to have higher G and Phi coefficients. These results indicate that SOLO taxonomy seems to be more reliable than RBT in the determination of the cognitive level of assessment questions.

Değerlendirme Sorularının Bilişsel Düzeylerinin Tespitinde SOLO ve Revize Edilmiş Bloom Taksonomisi'ne Dayalı Sınıflamaların Güvenirliklerinin Karşılaştırılması

Makale Bilgisi

DOI: 10.14527/pegegog.2017.023

Makale Geçmişi:
Geliş 06 Mart 2017
Düzeltme 03 Ağustos 2017
Kabul 15 Ağustos 2017
Çevrimiçi 05 Eylül 2017

Anahtar Kelimeler:
SOLO Taksonomisi,
Revize edilmiş Bloom Taksonomisi,
Genellenebilirlik Kuramı.

Makale Türü:
Özgün makale

Öz

Bu araştırmada, değerlendirme sorularının bilişsel düzeylerinin tespitinde SOLO ve revize edilmiş Bloom taksonomisine (REBT) dayalı sınıflamaların güvenirlikleri karşılaştırılmıştır. Araştırmanın verileri, 2015 yılında Milli Eğitim Bakanlığı tarafından basılan Yedinci Sınıf Sosyal Bilgiler Ders Kitabı ile Yıldırım Yayıncılık tarafından basılan Sekizinci Sınıf Fen ve Teknoloji Ders Kitabındaki değerlendirme sorularının üç uzman tarafından SOLO taksonomisi ve REBT'nin düzeylerine göre incelenmesiyle elde edilmiştir. Elde edilen veriler, değerlendirme sorularının ölçme objesi ve uzmanların yüzey olarak yer aldığı çapraz bir desen ile genellenebilirlik kuramına göre analiz edilmiştir. Çalışmada, değerlendirme sorusu ana etkisine ait varyans yüzdesi SOLO taksonomisine göre yapılan sınıflamalarda daha yüksek bulunmuştur. Uzman ana etkisine ilişkin varyans bileşeni ile artık varyans değerleri ise REBT'de SOLO taksonomisine kıyasla daha yüksek çıkmıştır. Son olarak, REBT ile karşılaştırıldığında, SOLO taksonomisine dayalı sınıflandırmalara ait G ve Phi katsayılarının daha yüksek olduğu saptanmıştır. Bu sonuçlar, değerlendirme sorularının bilişsel düzeylerinin belirlenmesinde SOLO taksonomisinin REBT'ye göre daha güvenilir bir model olduğunu göstermektedir.

[*] **Author/Yazar:** mustafailhan21@gmail.com

Introduction

The main aim of measurement and evaluation studies is to determine whether the learning objectives have been met and if met, to what extent the objectives have been achieved. Therefore, in measurement and evaluation studies, the learning outcomes to be evaluated should be in line with the curriculum objectives. The maintenance of this unity depends on the clarity of the learning objective definitions and statement of them in the form of observable performances. Taxonomies are used in order to define the learning objectives clearly and to transform them into observable learning outcomes. Taxonomies can be defined as the classification of things from simple to complex and in stages as prerequisites of one another. In the area of curriculum development, a taxonomy is the classification of expected behaviours from easy to complex, concrete to abstract in a hierarchical manner (Sönmez, 2004). In the literature, there are various taxonomies, such as Bloom's, SOLO, Haladyana, Marzano, Fink and Dettmer's. Of them, SOLO and Bloom's taxonomy stand out as the two most widely used taxonomies (Arı, 2013).

SOLO Taxonomy

The SOLO taxonomy, which is defined as the structure of observed learning outcomes, was developed by Biggs and Collis in 1982. This taxonomy is widely used for evaluating the students' performances (Biggs, 1979) and analysing the learning activities and assessment questions (Fensham & Bellocchi, 2013; Gezer & İlhan, 2014; Gezer & İlhan, 2015) in different disciplines such as geography, mathematics, foreign languages and science (Braband & Dahl, 2009; Jones, Collis, & Watson, 1993; Lucas & Mladenovic, 2009; Pegg & Tall, 2004). The SOLO taxonomy has a five-level structure. The levels that form the taxonomy and their basic features are shown in Table 1.

Table 1.
The SOLO Taxonomy's Levels and Their Basic Features[*].

Levels	Features
Pre-structural	Previous learning about the topic is inaccurate or nothing has been learned.
Uni-structural	Focuses on one aspect of the studied topic.
Multi-structural	Two or more aspects of the studied topic are understood; however, links cannot be formed between them.
Relational	Different aspects of the topic studied are related and thus, a consistent whole is obtained.
Extended Abstract	Reasoning beyond the existing knowledge and reaching generalizations. Knowledge is transferred to different areas.

Consistency, association and multi-level thinking increase and more meaningful learning is achieved. ↓

[*](Çetin & İlhan, 2016)

As Table 1 shows, the SOLO taxonomy comprises five levels: pre-structural, uni-structural, multi-structural, relational and extended abstract (Biggs, 1996; Biggs & Collis, 1982; Biggs & Tang, 2007). All the levels are taken into consideration in the evaluation of student performance, but the pre-structural level is not considered in the classification of the curriculum components. For, in contrast to the pre-structural level where no learning exists, all of the curriculum components correspond to a learning type, including even the lower levels. Therefore, the pre-structural level is not taken into consideration in the analysis of the cognitive levels of the learning outcomes and assessment questions. Analyses are performed on the basis of uni-structural, multi-structural, relational or extended abstract levels.

Bloom's Taxonomy

Bloom's taxonomy was devised by Bloom in 1956 in order to classify learning objectives in the cognitive field and was revised by Anderson and Krathwohl (Anderson & Krathwohl, 2001; Krathwohl, 2002). The terms, Bloom's taxonomy and revised Bloom's taxonomy (RBT), are used in studies to distinguish between the original and the updated forms. This study uses the revised form of Bloom's taxonomy developed by Anderson and Krathwohl (2001) and will use its acronym, RBT. The RBT has two dimensions: knowledge and cognitive process (Näsström, 2008; Turgut & Baykul, 2012). The differentiation depends on the research question. In the knowledge dimension, the critical question is: "what do the students know?", while in the cognitive process dimension, it is "how students think?" (Demirel, 2012; Anderson & Krathwohl, 2001). The knowledge dimension of the RBT comprises four categories: factual, conceptual, procedural and metacognitive. The cognitive process dimension can be ordered from simple to complex as: remember, understand, apply, analyse, evaluate and create (Anderson & Krathwohl, 2001; Krathwohl, 2002). This study compares the RBT and SOLO taxonomies, and since the SOLO taxonomy has only cognitive dimensions, only the cognitive process aspects of the RBT will be considered. The knowledge dimension will not be taken into consideration. The levels comprising the cognitive process dimensions of the RBT and their basic features are shown in Table 2.

Table 2.
The Levels of the RBT and their Features[*].

Levels	Features
Remember	Retrieving knowledge from long term memory
Understand	Understanding oral, written and graphic communication and giving related examples
Apply	Using and applying the appropriate operation in a given situation
Analyse	Dissecting the material into its components, determining the relationship between the parts and the general structure.
Evaluate	Making decisions and judgments based on criteria and standards
Create	Forming an original product and bringing pieces together in order to form a consistent whole

The levels are listed from simple to complex.

[*](Krathwohl, 2002)

As Table 2 shows, the levels of the cognitive process dimension of the RBT are: remember, understand, apply, analyse, evaluate and create. Of these, remember, understand and apply form the lower cognitive processes, while analyse, evaluate and create levels form the upper cognitive processes (Krathwohl, 2002).

A Comparison of SOLO and Bloom's Taxonomies

An analysis of the related literature reveals that a lot of studies have used the SOLO and Bloom's taxonomies (Alsaadi, 2011; Bağdat & Anapa-Saban, 2014; Başol, Balgalmış, Karlı, & Öz, 2016; Çalışkan, 2011; Dindar & Demir, 2006; Göçer & Kurt, 2016; Gökler, Aypay, & Arı, 2012; Özdemir & Göktepe-Yıldız, 2015; Peter & Alberto, 2013; Tarman & Kuran, 2015; Zorluoğlu, Kızılaslan, & Sözbilir, 2016); however, few have compared the two. Studies comparing SOLO and Bloom's taxonomies have yielded inconsistent results. For example, Hattie and Purdie (1994) found the SOLO taxonomy to be more reliable than Bloom's taxonomy, whereas Chan, Tsui, Mandy and Hong (2002) found Bloom's taxonomy to be more reliable than the SOLO taxonomy. The lack of consistent data on the reliability of the SOLO and Bloom's taxonomies necessitates further research on this issue.

The Aim and Significance of the Research

This study aims to compare the reliability of SOLO- and Bloom's taxonomy-based classifications in the determination of the cognitive levels of the assessment questions. The reliability of SOLO- and RBT-based classifications will be compared using generalizability theory to detect which of the taxonomies yields more reliable results. The comparison and contrast of different models with the same purpose is one of the key functions of science (Doğan, 2002). Since there is more than one theory or models serving the same or similar purposes, in order to advance scientific knowledge, the existing models should be compared and contrasted, and their strengths and weaknesses should be examined (Atılgan, 2004). This study aims to contribute to science by comparing two widely used education taxonomies. In addition, through the comparison of the two taxonomies, which of the taxonomies is more useful for achieving a common understanding between the experts in the analysis of the cognitive levels of assessment questions will be understood. In this respect, this study is expected to have practical implications.

This study differs from the previous literature in various aspects. Hatti and Purdie's (1994) study was the first study to compare the SOLO and Bloom's taxonomies. In their study, 30 teachers were given 19 multiple-choice questions and were asked to examine the cognitive levels of these questions in accordance with the SOLO and Bloom's taxonomies (Hattie & Purdie, 1994). In their study, 15 of the teachers examined the questions on the basis of the SOLO taxonomy, while the remaining 15 teachers used Bloom's taxonomy. It was found that when compared with the Bloom's taxonomy, the SOLO taxonomy yielded more consistent results among the teachers and it was argued that SOLO was more reliable than Bloom's taxonomy. This study differs from Hattie and Purdie (1994) in the following ways. First, in Hattie and Purdie (1994) there were 19 multiple-choice questions from the same discipline. However, classifications based on SOLO and Bloom's taxonomies may yield different results depending on the discipline and question types (open-ended, matching, completion, true-false). Therefore, the use of questions from different disciplines is important for the comparison of the two taxonomies. This study aims to contribute to the existing literature by contrasting SOLO and RBT in two disciplines: social studies and science and technology. In addition, in Hattie and Purdie (1994), different teachers were involved in the analysis of the assessment questions in terms of the SOLO and Bloom's taxonomies. Therefore, the finding that the SOLO taxonomy was found to be of higher reliability than Bloom's taxonomy could also be related to the employment of different teachers for the classifications. In this study, the classifications related to SOLO and Bloom's taxonomies will be carried out by the same experts, making it possible to attribute any differences between the two taxonomies to the taxonomies, not the experts.

Chan et al. (2002) also compared the reliability of the SOLO and Bloom's taxonomies. Their study examined inter-rater reliability with rubrics based on the SOLO and Bloom's taxonomies. The analysis showed that the inter-rater reliability of Bloom's taxonomy was higher than that of SOLO. This study differs from Chan et al. (2002) in the following ways. First, in Chan et al. (2002), the SOLO and Bloom's taxonomies were not used in the classification of curriculum components. The reliability of the rubrics based on these taxonomies was examined. Therefore, it is unknown whether the findings related to the reliability of the two taxonomies in Chan et al. (2002) will be relevant to studies which determine the cognitive levels of the assessment questions. The fact that the reliability of SOLO and RBT will be based on the analysis of the cognitive levels of the assessment questions distinguishes this study from that of Chan et al. (2002).

Finally, both Hattie and Purdie (1994) and Chan et al. (2002) regarded the experts and the raters as the only source of errors. In Hattie and Purdie (1994), the reliability between the raters was analysed on the basis of a simple percent agreement coefficient. In Chan et al. (2002), Pearson's correlation coefficient was used for the estimation of the inter-rater reliability. Unlike these studies, this study will investigate the reliability of the taxonomies using generalizability (G) theory. Therefore, random errors as well as expert related errors will be determined. In this way, in addition to explaining the inter-rater reliability of the two taxonomies, information related to susceptibility to random errors will also be

provided. Since G theory is used, how an increase or decrease in the number of experts will influence the reliability values of the two taxonomies will also be determined. In this way, it will answer the questions of how many experts are needed to reach optimal reliability values in the determination of cognitive levels of the assessment questions or if an increase in the number of experts will increase reliability. For these reasons, this study is expected to make significant contributions to the literature.

Method

Research Design

The study adopted the descriptive research design. Descriptive research seeks answers to the questions: What? and How? These studies are based on the description of a situation as it is without intervention (Kumar, 2008). Depending on the aim of the descriptive research, questionnaires, observation, interviews or document analysis can be used (Anderson & Arsenaut, 2005; Erkuş, 2011). This study used document analysis. The first stage of this technique, also known as document review, is to collect materials related to the research topic (Karasar, 2009). The second stage is the analysis of the collected documents using specific criteria (Yıldırım & Şimşek, 1999). In document analysis, very different document types can be analysed. These include written materials such as books, journals, newspapers, diaries, official documents, and statistics as well as audio-visual records such as films, videos, or photos (Cansız-Aktaş, 2014).

Data Collection

The data were obtained by three experts' examination of the assessment questions in the Seventh Grade Social Studies Textbook published by the Ministry of National Education and the Eight Grade Science and Technology Textbook published by Yıldırım Publishing in 2015. In the social studies textbook, 98 assessment questions (18 true-false questions [18.37%], 11 matching questions [11.22%], 43 multiple-choice questions [43.88%] and 26 short-answer or essay questions [26.53%]) were analysed. As for the science and technology textbook, 198 assessment questions were analysed. Of the assessment questions, 41 were true-false questions (20.71%), 45 were matching questions (22.73%), 76 were multiple-choice questions (38.38%), and 36 (18.18%) were short-answer or essay type questions.

Data Coding

Three experts (two educational curriculum and instruction experts, one measurement and evaluation expert) carried out the analysis of the cognitive levels of the assessment questions. The experts classified the assessment questions in the social studies and science and technology textbooks on the basis of the SOLO and RBT taxonomies. The basic features of the SOLO and RBT taxonomy levels and the indicator verbs were used for these classifications. The indicator verbs are shown in Table 3.

Table 3 does not show the pre-structural level of the SOLO taxonomy. This is because there is no learning on the pre-structural level, and there are thus no corresponding indicator verbs (Çetin & İlhan, 2016). Therefore, in the analysis of assessment questions, the prestructural level was not taken into consideration. The classifications were based on the uni-structural, multi structural, relational and extended abstract levels.

Table 3.
SOLO Taxonomy and RBT Indicator Verbs.

	SOLO Taxonomy[1]
Uni-structural	Memorize, recite, define, name, label, match, recall, quote, draw, portray, imitate, recognize
Multi-structural	Classify, list, discuss, show, select, do basic calculations, tell, report, separate, summarize
Relational	Apply, compare, predict, differentiate, organize, analyse, calculate the relationships between X and Y, interpret, review and rewrite, structure, transform, infer
Extended Abstract	Hypothesize, generalize, construct, reflect, form a theory, invent, develop an original product, transfer to a different area, prove, solve from first principles
	RBT[2]
Remember	Name, list, label, recall, recite, cite, count, number, select, draw, portray, define
Understand	Add, explain, clarify, demonstrate, match, guess, explain the similarities and differences
Apply	Calculate, solve, organize, follow the steps, transfer by organizing, draw an outline
Analyse	Separate, summarize, define in broad terms, analyse, prioritize, simplify, categorize, test, optimize, infer
Evaluate	Evaluate, critique, prove, judge, defend, propose
Create	Organize, arrange, compose, design, reorganize, adapt, unify, discover, build, construct

[1] (Biggs & Tang, 2011), [2] (Thompson, 2012)

Data Analysis

After the assessment questions in the textbooks were classified by three experts using SOLO and RBT, the data were analysed using G theory. G theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is an extension of the classical test theory (Brennan, 2003; Huang, 2009). G theory statistics make use of ANOVA for separating students' test scores from the components contributing to the variance (Renkl, 1993). To clarify, in G theory analyses, the variance of the students' test scores is expected to be separated into different types of variances: variance resulting from the differences in the ability levels of students, variance related to different variability sources like raters and items, variance in the interaction between students and different variability sources, variance between the interaction of different variability sources, variance due to unexplained sources and random error. ANOVA is used for this purpose (Taylor & Pastor, 2013). However, in G theory, unlike traditional ANOVA practices, the size of the variance components is taken into consideration rather than the significance of the mean squares (Bell, 1985).

G theory can be understood on the basis of certain concepts. This specific terminology includes: the measurement object, facet, condition, crossed and nested design, generalizability (G) and decision (D) studies, generalizability and dependability coefficients. The measurement object is the source of variance targeted by the measurement study. In all measurement studies, differentiation of the

components forming the measurement object is expected. In other words, variance of the measurement object is expected to be high. In many measurement studies, the examinees are the object of measurement (Güler, Kaya-Uyanık, & Taşdelen-Teker, 2012). However, in this study, the case is a little different. Since this study aims to determine the cognitive levels of the assessment questions and to differentiate between questions of different cognitive levels, the assessment questions are the measurement object. In G theory, other sources of variance are facets, and the levels within each facet are called conditions. For example, if raters are a source of variance that may lead to error in measurement results, the rater is a facet and each of the raters (i.e., 1st, 2nd, 3rd, etc.) is a condition (Mushquash & O'Connor, 2006). In this study, the experts who analysed the cognitive levels of the assessment questions are the facet of the measurement, and since there were three experts, the number of conditions is three.

Generalizability theory statistics come in two categories: crossed designs and nested designs (Cardinet, Johnson, & Pini, 2010). In crossed design, all conditions of a facet influence all conditions of another facet. Designs where some of the conditions of any facet are observed by some conditions of another facet are called nested designs (Kumazawa, 2009). Crossed designs are notified by "x" while nested designs are shown with ":" (Brennan, 2003). For example, in a measurement operation aiming to evaluate individuals' mathematical performance, suppose that the participants are "p" and the items in the mathematics test are "i". Here, if all students answer the same items, the design is crossed and is shown as "p x i". For the same measurement operation, if each of the students is responsible for different items, the design is nested and is defined as "p : i" (Hathcoat & Penn, 2012). In this study, the assessment questions in the textbooks were classified in terms of their cognitive levels, and all of the questions (q) were analysed by three experts (e). Therefore, the G theory analyses were based on crossed "q x e" design.

G theory analyses have two separate, but related stages (Sun, Valiga, & Gao, 1997). The first are G studies, and the second, D studies (Shavelson & Webb, 1991). The aim of G studies is to predict the variance components related to the measurement population (Brennan, 2001). In other words, G studies are carried out in order to obtain information about the source of variance influencing the generalizability of the observations (Alkharusi, 2012). The second stage of G theory analyses is D studies. D studies use the predictions of the G studies (Webb, Shavelson, & Haertel, 2006). At this point, new observations are not made or data is not collected again (Murray, 1984). The aim of the D studies is to define error variance and predict the reliability coefficient (Stora, Hagtvet, & Heyerdahl, 2013) and determine how reliability will change if the number of item, rater or observation changes (Mabe, 2013). Namely, D studies aim to answer the question: How will the error and reliability of the measurement change if the number of conditions in a facet is increased or decreased? (Briggs & Wilson, 2007). D studies can assess the effectiveness of alternative designs so that error can be minimized, and reliability can be maximized (Webb et al. 2006). In this G study, the variances related to the measurement results were: *i)* variance related to different cognitive levels of the assessment questions, *ii)* variance related to experts analysing the cognitive levels and *iii)* residual variance from the interaction of the assessment questions and experts and random error. In the D study, how the number of experts influenced the dependability and reliability values was tested in alternative designs where the number of experts was one, two, four, five and six.

Another important concept related to the G theory is the difference between generalizability (G) and dependability (Phi) coefficients. In G theory, there are two types of decision making: relative and absolute, and G coefficient is used for relative evaluations, while the Phi coefficient is used for absolute evaluations (Atılgan, 2005). Therefore, in this study both G and Phi coefficients were reported for the reliability of the SOLO and RBT-based classifications. EduG software was used for G theory analyses.

Results

This section presented the findings. First, the data collected from the SOLO- and RBT-based classification of the assessment questions in the Seventh Grade Social Studies Textbook by three experts were analysed. The variance components predicted as a result of the crossed design (q x e) G study and their contributions to the total variance are shown in Table 4.

Table 4.
Variance Components Predicted by the SOLO and RBT-Based Classification of the Assessment Questions in the Social Studies Textbook.

Taxonomy	Source of Variance	Sum of Squares	df	Mean of Squares	Variance Component	Variance Percentage
SOLO	Assessment question (q)	187.21	97	1.93	.63	95.30
	Expert (e)	.01	2	.00	.00*	.00
	q x e, error	5.99	194	.03	.03	4.70
	Total	193.21	293			100
RBT	Assessment question (q)	409.73	97	4.22	1.26	71.60
	Expert (e)	11.07	2	5.53	.05	3.00
	q x e, error	86.93	194	.45	.45	25.50
	Total	507.73	293			100

* Negative (–.00028) variance components were assigned a value of zero.

According to Table 4, the rate of the variance component predicted for the assessment question main effect in the total variance was 95.30% in the SOLO-based classification and 71.60% in the RBT-based classification. These values show that in both the SOLO- and RBT-based classifications, the variance component that had the highest contribution to the total variance belonged to the assessment question main effect.

Following the assessment question main effect, the variance components predicted for the expert main effect were analysed. In the SOLO-based classification, the variance value of the expert main effect was found to be –.00028. In G theory, all variance components are expected to be positive (Cardinet et al. 2010). In spite of this, negative variance values can be obtained in practice, due to small sample size, using an inappropriate model, or the population value can be zero or close to zero (Chiu, 2001). When a negative but close to zero value is observed, it is recommended that the related variance component should be taken as zero. When a variance component that is negative and much lower than zero is obtained, the data set should be monitored for mistakes (Webb, Rowley, & Shavelson, 1988). In this study, since the negative variance calculated for the expert main effect was close to zero in the SOLO-based classification, a value of zero was assigned to the related variance, and the percentage of the variance components within the total variance was considered. As Table 4 shows, in the SOLO-based classification, the percentage of the expert-related variance in the total variance was calculated as .00%. In the RBT-based classification, 3.00% of the total variance could be explained by the expert main effect.

In G theory, the variance value of the interaction of the measurement object and all facets form the residual variance together with unexplained sources of variance (Shavelson & Webb, 1991). In this study, the assessment questions were the measurement object, and the experts were the only facet that was analysed. Therefore, the "q x e, error" variance referred to the residual variance. The variance rates in Table 4 indicate that the residual variance values were 4.70% in the SOLO-based classifications and 25.50% in RBT-based classifications.

After the prediction of the variance components, G and Phi coefficients were calculated for the data obtained by the SOLO- and RBT-based classification of the assessment questions in the social studies textbook by three experts. The G and Phi coefficients predicted for the original data are shown in Table 5 together with the alternative scenarios related to expert number.

Table 5.
G and Phi Coefficients Predicted by the SOLO and RBT-Based Classification of the Assessment Questions in the Social Studies Textbook in Alternative Decision Studies.

		Expert number					
Taxonomy	**G and Phi Coefficients**	**1**	**2**	**3**	**4**	**5**	**6**
SOLO	G	.95	.98	**.98**	.99	.99	.99
	Phi	.95	.98	**.98**	.99	.99	.99
RBT	G	.74	.85	**.89**	.92	.93	.94
	Phi	.72	.83	**.88**	.91	.93	.94

Table 5 shows that the G and Phi coefficients calculated in SOLO-based classifications of the cognitive levels of the assessment questions in the social studies textbook were found to be higher than those of the RBT-based classification. The findings of the alternative decision studies indicate that the effect of the changes in the expert number on G and Phi coefficients was smaller in the SOLO-based classifications than in the RBT-based classifications.

After data related to the detecting of the cognitive levels of the assessment questions in the social studies textbook were analysed by means of G theory, similar analyses were performed for the science and technology textbook. In this way, the limitations due to the use of assessment questions from a single course could be eliminated. The classification of the cognitive levels of the assessment questions in the science and technology course was analysed in a crossed design (q x e). The analyses had two steps: G and D stages. In the G study, the variance components and the percentage of each variance component in the total variance was calculated. These values are shown in Table 6.

Table 6.
Variance Components Predicted by the SOLO and RBT-Based Classification of the Assessment Questions in the Science and Technology Textbook.

Taxonomy	**Source of Variance**	**Sum of Squares**	**df**	**Mean of Squares**	**Variance Component**	**Variance Percentage**
SOLO	Assessment question (q)	251.83	197	1.28	.42	92.80
	Expert (e)	.04	2	.02	.00*	.00
	q x e, error	12.62	394	.03	.03	7.20
	Total	264.50	593			100
RBT	Assessment question (q)	285.60	197	1.45	.43	73.50
	Expert (e)	3.84	2	1.92	.01	1.50
	q x e, error	58.16	394	.15	.15	25.00
	Total	347.60	593			100

* Negative (-.00005) variance components were assigned a value of zero.

Table 6 shows that the predicted variance component for the assessment question main effect was 92.80% in the SOLO-based classification, with the highest contribution to the total variance. Similarly, in the RBT-based classification, the main effect of the assessment question had the highest contribution to the total variance, explaining 73.50% of the total variance. Table 6 shows the variance components predicted for the expert main effect. The rate of these variance components in the total variance was .00% in the SOLO-based classification and 1.50% in the RBT-based classification. Residual variance components explained 7.20% of the total variance in the SOLO-based classification and 25.00% of the total variance in the RBT-based classification.

D study results of the SOLO- and RBT-based classifications of the cognitive levels of the assessment questions in the science and technology textbook are in Table 7, which shows the G and Phi coefficients calculated for the original expert number (three) and alternative scenarios where the expert number is one, two, four, five and six.

Table 7.
Predicted G and Phi Coefficients for Data Obtained by the SOLO and RBT-based Classification of Science and Technology Course Assessment Questions.

Taxonomy	G and Phi Coefficients	Expert number					
		1	2	3	4	5	6
SOLO	G	.93	.96	**.97**	.98	.98	.99
	Phi	.93	.96	**.97**	.98	.98	.99
RBT	G	.75	.85	**.90**	.92	.94	.95
	Phi	.73	.85	**.89**	.92	.93	.94

As Table 7 shows, in the SOLO based classifications, the G and Phi coefficients were found to be equivalent (.97). In the RBT-based classification, G and Phi coefficients were found to be .90 and .89, respectively. The findings of the alternative D studies show that in SOLO-based classifications, the G and Phi coefficient for a single expert was .93 and as the number of experts increased, the G and Phi coefficients also increased, up to .99 for six experts. In the RBT-based classifications, the G and Phi coefficients of a single expert were found to be .75 and .73, respectively. As the number of experts increased, G and Phi coefficients also increased, and the highest increase was in the measurement conditions where the number of experts increases from one to two.

Discussion & Conclusion

In this study, the SOLO and RBT-based classifications of the cognitive levels of the assessment questions were analysed using G theory. The G study found that in both taxonomies, the variance components predicted for the main effect of the assessment question had the highest contribution to the total variance. The finding that the measurement object, that is, the assessment questions, had the highest contribution to the total variance was expected (Güler et al. 2012). Therefore, in the determination of cognitive levels of the assessment questions, no matter which taxonomy is used, a difference between the cognitive levels of the questions could be uncovered. However, it should be noted that the variance percentage of the assessment question in the total variance was higher in the SOLO-based classification than in the RBT. This difference points to the fact that the SOLO taxonomy is a more effective model in terms of the analysis of the cognitive levels of the assessment questions.

In both SOLO- and RBT-based classifications, the expert main effect had small contributions to the total variance. On this basis, it can be argued that, in the analysis of the cognitive level of the assessment questions, there are not significant differences between the experts' classification of the cognitive levels. A comparison of the variance components due to experts showed a slight difference between the two taxonomies. The agreement among the experts was higher in the SOLO based taxonomy. This finding is in line with Hattie and Purdie (1994). In Hattie and Purdie (1994), the SOLO taxonomy also yielded higher inter-rater reliability in the analysis of the cognitive levels of the multiple choice questions. The results of İlhan and Çetin (2016) and Çetin, Boran and Yazıcı (2014) on rater' views based on SOLO taxonomy support this study's findings. İlhan and Çetin (2016) analysed the teacher ratings of student responses to open ended questions on the basis of the SOLO based rubrics and the teachers' views of the rubrics. The teachers reported that the SOLO taxonomy had clear and intelligible levels. Similarly, in another study by Çetin et al. (2014), student responses to open-ended physics questions were rated by three raters on the basis of SOLO-based rubrics. All of the three raters reported that the objectivity of the SOLO taxonomy was high.

The results of Chan et al. (2002) differ from this study. Chan et al. (2002) analysed inter-rater reliability for SOLO and Bloom's taxonomy-based rubrics. They found that Bloom's taxonomy based rubrics had a higher correlation coefficient than SOLO-based rubrics. The difference between this study and Chan et al. (2002) could be attributed to unfamiliarity with the SOLO taxonomy and the difficulty in using it (Leung, 2000), although it has clear and intelligible levels. Since Bloom's is a more renowned

taxonomy among researchers and teachers (Arı, 2013), the same difficulty is not expected for Bloom's taxonomy. In this study, the indicator verbs corresponding to both taxonomies' levels were explained to the experts and possible effects related to unfamiliarity with the taxonomies were eliminated.

Analysis of residual variance percentages in G analysis found that in SOLO-based classifications, random error was low, and in RBT-based classifications, the residual variance explained a significant portion (1/4) of the total variance. Residual variance gives information about the random error influencing the measurement and is expected to be as close to zero as possible (Güler & Taşdelen-Teker, 2015). Therefore, unlike SOLO-based taxonomy, RBT is more open to random error. In other words, in RBT-based classifications, experts may assign the assessment questions to different levels in a non-systematic way and factors other than experts could influence the measurement results to a greater extent because RBT yields more random error than SOLO based classifications.

Finally, SOLO-based classifications were found to have higher G and Phi coefficients than those of RBT. In alternative D studies, it was calculated that the reliability coefficients obtained by a single expert in SOLO taxonomy could be produced by five or six experts in RBT. These results indicate that, in the analysis of cognitive levels of the assessment questions, the SOLO taxonomy yields more reliable and generalizable measures than RBT.

Limitations & Recommendations

This study makes the following contributions to the literature: it focuses on specific points not addressed by previous studies that compared SOLO and RBT, and instead of classical test theory, G theory was used. It should be noted that there a number of limitations to the study. The limitations of the study and further research recommendations can be listed as follows: first, this research is limited to the comparison of SOLO and RBT taxonomies. Studying different taxonomies (Haladyana, Marzano, Fink, Dettmer, MATH) for detecting the cognitive levels of curriculum components can be suggested; second, the assessment questions of the social studies and science and technology courses were used to compare SOLO and RBT, and using the assessment questions of two different courses contributes to the generalizability of its findings. The fact that social studies and science and technology courses have a broad-field design (the social studies includes the history and geography courses while biology, chemistry, and physics courses are integrated into general science) makes the findings obtained largely independent of any discipline. However, further related research should involve different disciplines such as mathematics and foreign languages.

Türkçe Sürüm

Giriş

Eğitim sürecinde yapılan ölçme değerlendirme çalışmalarının temel amacı, öğrenme hedeflerine ulaşılıp ulaşılmadığını belirlemek ve ulaşıldıysa söz konusu hedeflerin ne düzeyde gerçekleştiğini ortaya koymaktır. Dolayısıyla ölçme ve değerlendirme çalışmalarında dikkat edilmesi gereken temel husus, ölçülecek öğrenme çıktılarının programda yer alan hedefler ile uyumlu olmasıdır. Bu uyumun sağlanabilmesi, öğrenme hedeflerinin herkes tarafından aynı şekilde anlaşılmasını sağlayacak açıklıkta tanımlanmasına ve gözlenebilir performanslar şeklinde ifade edilmesine bağlıdır. Öğretim hedeflerinin açık bir biçimde tanımlanması ve gözlenebilir öğrenme çıktılarına dönüştürülmesinde taksonomilerden faydalanılmaktadır. Taksonomiler, varlıkların basitten karmaşığa ve birbirinin ön koşulu olacak biçimde aşamalı olarak sınıflandırılmasıdır. Program geliştirme alanında ise taksonomi; istendik davranışların kolaydan zora, somuttan soyuta ve hiyerarşik bir yapıda sıralanması anlamını taşımaktadır (Sönmez, 2004). Alanyazında farklı ders ve konu alanlarına yönelik olarak geliştirilmiş Bloom, SOLO, Haladyana, Marzano, Fink ve Dettmer gibi çok sayıda taksonomi bulunmakla birlikte, bunlardan özellikle SOLO ve Bloom araştırmacıların en sık kullandığı iki taksonomi olarak karşımıza çıkmaktadır (Arı, 2013).

SOLO Taksonomisi

Gözlenebilir öğrenme çıktılarının yapısı (structure of observed learning outcomes) olarak tanımlanan SOLO taksonomisi, 1982 yılında Biggs ve Collis tarafından ortaya konmuştur. Bu taksonomi; coğrafya, matematik, fen bilimleri, yabancı diller gibi farklı disiplinlerde (Braband & Dahl, 2009; Jones, Collis, & Watson, 1993; Lucas & Mladenovic, 2009; Pegg & Tall, 2004), hem öğrencilerin performanslarının değerlendirilmesinde (Biggs, 1979) hem de kazanımlar, öğrenme etkinlikleri ve değerlendirme sorularının bilişsel düzeylerinin incelenmesinde (Fensham & Bellocchi, 2013; Gezer & İlhan, 2014; Gezer & İlhan, 2015) sıklıkla kullanılan bir modeldir. SOLO taksonomisi beş düzeyli bir yapıya sahiptir. Taksonomiyi oluşturan düzeyler ile bu düzeylerin temel özellikleri Tablo 1'de sunulmuştur.

Tablo 1.
SOLO Taksonomisinin Düzeyleri ile Bu Düzeylerin Temel Özellikleri[*].

Düzeyler	Özellikleri
Yapı Öncesi	Üzerinde çalışılan konu ile ilgili öğrenilenler yanlıştır ya da herhangi bir şey öğrenilmemiştir.
Tek Yönlü Yapı	Üzerinde çalışılan konunun tek bir yönüne odaklanılır.
Çok Yönlü Yapı	Üzerinde çalışılan konunun iki ya da daha fazla yönü anlaşılır fakat parçalar arasında ilişki kurulamaz.
İlişkisel Yapı	Üzerinde çalışılan konunun farklı yönleri birbiri ile ilişkilendirilir, bu sayede tutarlı bir bütün elde edilir.
Soyutlanmış Yapı	Mevcut bilgilerin ötesinde akıl yürütülebilir ve genellemelere ulaşılabilir. Farklı bir alana transfer edebilme söz konusudur.

Tutarlılık, ilişkilendirme ve çok yönlü düşünme artmakta, daha anlamlı öğrenmeler gerçekleşmektedir.

[*](Çetin & İlhan, 2016)

Tablo 1'de görüldüğü üzere SOLO taksonomisi; yapı öncesi, tek yönlü yapı, çok yönlü yapı, ilişkisel yapı ve soyutlanmış yapı olmak üzere beş düzeyden oluşmaktadır (Biggs, 1996; Biggs & Collis, 1982; Biggs & Tang, 2007). Öğrencilerin performanslarının değerlendirilmesinde bu beş düzeyin tamamı esas alınırken, program öğelerinin sınıflandırılmasında yapı öncesi basamağı göz önünde bulundurulmamaktadır. Çünkü herhangi bir öğrenmenin mevcut olmadığı yapı öncesi düzeyinin aksine program öğelerinin tümü alt düzeyde de olsa bir öğrenmeye karşılık gelmektedir. Buna bağlı olarak

kazanımların ve değerlendirme sorularının bilişsel seviyeleri incelenirken yapı öncesi basamağı dikkate alınmamakta; analizler tek yönlü, çok yönlü, ilişkisel ve soyutlanmış yapı düzeyine göre gerçekleştirilmektedir.

Bloom Taksonomisi

Bloom taksonomisi, Bloom tarafından 1956 yılında bilişsel alana yönelik öğrenme hedeflerinin sınıflandırılması amacıyla geliştirilmiş, Anderson ve Krathwohl tarafından 2001 yılında revize edilmiştir (Anderson & Krathwohl, 2001; Krathwohl, 2002). Bu değişikliğin ardından, Bloom taksonomisinin orijinal formu ile güncellenmiş halinin birbirinden daha net bir biçimde ayırt edilebilmesi için yapılan çalışmalarda orijinal Bloom taksonomisi ve Revize Edilmiş Bloom Taksonomisi (REBT) şeklinde bir adlandırmaya gidilmiştir. Bu çalışmada Bloom taksonomisinin Anderson ve Krathwohl tarafından (2001) yenilenen formu esas alınmış ve çalışma boyunca güncellenen Bloom taksonomisini ifade etmek için REBT şeklinde bir adlandırma kullanılmıştır. REBT, bilgi ve bilişsel süreç boyutu olmak üzere iki ayrı boyutta ele alınmaktadır (Näsström, 2008; Turgut & Baykul, 2012). Bu ayrıma esas teşkil eden nokta ise yanıt aranan sorudur. Bilgi boyutunda öğrenciler *ne biliyor* sorusuna yanıt aranırken; bilişsel süreç boyutunda öğrenciler *nasıl düşünüyor* sorusu cevaplanmaya çalışılmaktadır (Demirel, 2012; Anderson & Krathwohl, 2001). REBT'nin bilgi boyutu, dört bilgi türünden oluşmaktadır. Bunlar olgusal, kavramsal, işlemsel ve üst bilişsel bilgidir. Bilişsel süreç boyutu ise basitten karmaşığa doğru hatırlama, anlama, uygulama, analiz, değerlendirme ve yaratma (oluşturma) şeklinde sıralanmaktadır (Anderson & Krathwohl, 2001; Krathwohl, 2002). Bu çalışmada REBT ile SOLO taksonomilerinin karşılaştırılması amaçlandığından ve SOLO taksonomisi sadece bilişsel süreç boyutundan meydana geldiğinden, araştırma kapsamında REBT'nin yalnızca bilişsel süreç boyutu üzerinde durulmuş; bilgi boyutuna ise yer verilmemiştir. REBT'de bilişsel süreç boyutunu oluşturan düzeyler ve bu düzeylerin temel özellikleri Tablo 2'de sunulmuştur.

Tablo 2.
REBT'nin Düzeyleri ile Bu Düzeylerin Temel Özellikleri[*].

Düzeyler	Özellikleri
Hatırlama	Uzun süreli bellekten bilgiyi geri çağırma.
Anlama	Sözlü, yazılı ve grafik iletişimi içeren öğretici mesajlardan anlam çıkarma ve konuya ilişkin farklı örnekler verme.
Uygulama	Verilen bir durumda uygun işlemi kullanma ve uygulama.
Analiz	Materyali bileşenlerine ayırma, parçaların birbiriyle ve materyalin genel yapısıyla nasıl bir ilişki içerisinde olduğunu belirleme.
Değerlendirme	Kriter ve standartlara dayalı olarak karar verme/yargıda bulunma.
Yaratma	Orijinal bir ürün oluşturma veya tutarlı bir bütün oluşturmak için parçaları bir araya getirme.

Düzeyler basitten karmaşığa doğru sıralanmaktadır.

[*](Krathwohl, 2002)

Tablo 2'de görüldüğü üzere REBT'yi oluşturan düzeyler; hatırlama, anlama, uygulama, analiz, değerlendirme ve yaratma şeklinde adlandırılmaktadır. Bu düzeylerden hatırlama, anlama ve uygulama alt bilişsel süreçleri; analiz değerlendirme ve yaratma basamakları ise üst bilişsel süreçleri meydana getirmektedir (Krathwohl, 2002).

SOLO ve Bloom Taksonomilerinin Karşılaştırılması

İlgili alanyazın tarandığında SOLO ve Bloom taksonomilerinin kullanıldığı çok sayıda araştırma (Örn.; Alsaadi, 2011; Bağdat & Anapa-Saban, 2014; Başol, Balgalmış, Karlı, & Öz, 2016; Çalışkan, 2011; Dindar & Demir, 2006; Göçer & Kurt, 2016; Gökler, Aypay, & Arı, 2012; Özdemir & Göktepe-Yıldız, 2015; Peter &

Alberto, 2013; Tarman & Kuran, 2015; Zorluoğlu, Kızılaslan, & Sözbilir, 2016) bulunduğu fakat bu taksonomilerin karşılaştırıldığı çalışmaların son derece sınırlı olduğu görülmektedir. SOLO ve Bloom taksonomilerinin karşılaştırılmasına dönük çalışmalar incelendiğinde ulaşılan sonuçlar arasında bir tutarlılık bulunmadığı anlaşılmaktadır. Örneğin, Hattie ve Purdie'nin (1994) araştırmasında SOLO taksonomisinin Bloom taksonomisine göre daha güvenilir olduğu şeklinde bir sonuca ulaşılmışken; Chan, Tsui, Mandy ve Hong (2002) tarafından yapılan çalışmada Bloom taksonomisinin SOLO taksonomisine kıyasla daha güvenilir sonuçlar ürettiği belirlenmiştir. Alanyazında SOLO ve Bloom taksonomilerinden hangisinin daha güvenilir sonuçlar verdiğine ilişkin net bir görüşün bulunmayışı bu konuda yeni araştırmalara ihtiyaç olduğunu düşündürmektedir.

Araştırmanın Amacı ve Önemi

Bu araştırmada, değerlendirme sorularının bilişsel düzeylerinin tespitinde SOLO taksonomisi ve REBT'ye dayalı sınıflamaların güvenirliklerinin karşılaştırılması amaçlanmaktadır. Bu amaç çerçevesinde; değerlendirme sorularının bilişsel düzeyleri için SOLO taksonomisi ile REBT'ye göre yapılan sınıflamaların güvenirliklerinin genellenebilirlik kuramına göre incelenmesi ve elde edilen sonuçların karşılaştırılıp hangi taksonominin daha güvenilir sonuçlar verdiğinin belirlenmesi hedeflenmektedir. Bilindiği üzere, benzer amaçlarla ileri sürülmüş değişik modellerin karşılaştırılması ve karşılaştırılan modeller arasındaki ortak noktaların veya farklılıkların saptanması bilimin temel işlevlerinden biridir (Doğan, 2002). Benzer veya aynı amaca hizmet etmesi için geliştirilmiş birden fazla kuram ya da model olabildiğinden bilimin ilerleyebilmesi için ileri sürülen farklı modellerin karşılaştırılması ve bu modellerin birbirlerine göre üstünlük/zayıflıklarının incelenmesi gerekmektedir (Atılgan, 2004). Bu bağlamda, eğitim alanında sıklıkla kullanılan iki taksonominin karşılaştırılmasının hedeflendiği bu çalışmanın bilimsel bir işlevinin olacağı öngörülmektedir. Buna ek olarak, iki kuramının karşılaştırılmasıyla değerlendirme sorularının bilişsel düzeylerinin tespitinde uzmanlar arasında ortak bir anlayış oluşturma konusunda hangi taksonominin daha işlevsel olduğu belirlenebilecektir. Bu yönüyle araştırmanın uygulamaya dönük katkılar sunması da beklenmektedir.

Bu araştırma çeşitli açılardan konu ile ilgili alanyazında bulunan önceki çalışmalardan farklılık göstermektedir. Hatti ve Purdie (1994) tarafından yapılan araştırma, SOLO ve Bloom taksonomilerinin karşılaştırılmasına yönelik alanyazında rastlanan ilk çalışmadır. Hattie ve Purdie'nin (1994) araştırmasında 30 öğretmene çoktan seçmeli 19 soru yöneltilerek bu soruların bilişsel düzeylerini SOLO ve Bloom taksonomisinin basamakları doğrultusunda incelemeleri istenmiştir. Çalışmada aynı soruları öğretmenlerden 15'i SOLO taksonomisine göre incelerken diğer 15 öğretmen Bloom taksonomisini temele alarak değerlendirmiştir. Araştırma sonucunda Bloom taksonomisi ile karşılaştırıldığında SOLO taksonomisinde öğretmenler arasındaki uyumun daha yüksek olduğu tespit edilmiş ve bu bulgu SOLO taksonomisinin Bloom taksonomisine göre daha güvenilir olduğu şeklinde yorumlanmıştır. Bu araştırmayı Hattie ve Purdie'nin (1994) çalışmasından farklı kılan özellikler şu şekilde açıklanabilir. İlk olarak Hattie ve Purdie'nin (1994) yaptığı araştırmada, tek bir disipline yönelik çoktan seçmeli 19 soruya yer verilmiştir. Ancak SOLO ve Bloom taksonomilerine göre gerçekleştirilen sınıflandırmaların çeşitli disiplinler ve değişik türdeki sorular (açık uçlu, eşleştirme, boşluk doldurma, doğru-yanlış) üzerinde farklı sonuçlar doğurabileceği düşünülmektedir. Bundan dolayı iki taksonominin karşılaştırılmasında farklı disiplinlere yönelik değişik türdeki soruların kullanılması önemli görülmektedir. Bu çalışmada SOLO taksonomisi ve REBT, Sosyal Bilgiler ve Fen Bilimleri dersi olmak üzere iki ayrı disiplin ile farklı türdeki değerlendirme soruları üzerinden karşılaştırıldığından araştırmanın alanyazına katkı sağlaması beklenmektedir. Ayrıca, Hattie ve Purdie (1994) tarafından yapılan çalışmada değerlendirme sorularının SOLO ve Bloom taksonomilerinin düzeylerine göre incelenmesinde farklı öğretmenler görev almıştır. Dolayısıyla Hattie ve Purdie'nin (1994) çalışmasında, SOLO taksonomisine ilişkin güvenirliğin Bloom taksonomisi için hesaplanan güvenirlikten daha yüksek çıkması, taksonomilerin kendisinden kaynaklanmış olabileceği gibi iki taksonomiye göre yapılan sınıflamalarda aynı öğretmenlerin görev almamasından da kaynaklanmış olabilir. Bu araştırmada ise SOLO ve Bloom taksonomisine ilişkin sınıflandırmalar aynı uzmanlar tarafından yapılacağından, iki taksonomi arasında bir fark belirlenmesi

durumunda bunun değerlendirme sorularının bilişsel düzeylerini inceleyen uzmanlardan değil taksonomilerin kendisinden kaynaklandığı daha net bir biçimde ortaya konulabilecektir.

SOLO ve Bloom taksonomilerinin güvenirliklerinin karşılaştırılmasına yönelik ikinci bir çalışma Chan vd. (2002) tarafından yapılmıştır. Chan vd. (2002) tarafından yapılan araştırmada SOLO ve Bloom taksonomisinin düzeyleri referans alınarak oluşturulmuş rubriklerde puanlayıcılar arası güvenirlik incelemiştir. Yapılan incelemede, Bloom taksonomisine dayalı rubriklerde puanlayıcılar arası güvenirlik SOLO temelli rubriklere kıyasla daha yüksek bulunmuştur. Bu araştırma ile Chan vd.'nin (2002) çalışmasının farklılaştığı noktalar şöyle sıralanabilir. Öncelikle Chan vd. (2002) tarafından yapılan çalışmada, SOLO ve Bloom taksonomisi program öğelerinin sınıflandırılmasında kullanılmamış, bu taksonomilere dayalı rubriklerin güvenirliklerinin incelenmesine odaklanılmıştır. Bundan dolayı, Chan vd.'nin (2002) çalışmasında SOLO ve Bloom taksonomilerinin güvenirliklerine ilişkin elde edilen bulguların bu taksonomilerin değerlendirme sorularının bilişsel düzeylerinin belirlenmesinde kullanıldığı çalışmalar için geçerli olup olmayacağı bilinmemektedir. SOLO taksonomisi ve REBT'nin güvenirliklerinin değerlendirme sorularının bilişsel düzeylerinin tespiti açısından incelenecek olması bu araştırmayı Chan vd.'nin (2002) çalışmasından farklı kılmaktadır.

Son olarak, hem Hattie ve Purdie (1994) hem de Chan vd.'nin (2002) yaptığı çalışmada hata kaynağı olarak sadece uzmanlar/puanlayıcılar alınmıştır. Hattie ve Purdie (1994) tarafından yapılan çalışmada değerlendirme sorularının bilişsel düzeylerini analiz eden uzmanlar arasındaki uzlaşma basit uyum yüzdesi ile incelenmiştir. Chan vd. (2002) tarafından yapılan araştırmada ise SOLO ve Bloom taksonomisine dayalı rubrikleri kullanarak değerlendirme yapan puanlayıcılar arasındaki güvenirliğin belirlenmesinde Pearson korelasyon katsayısından yararlanılmıştır. Söz konusu çalışmalardan farklı olarak bu araştırmada iki taksonominin güvenirlikleri genellenebilirlik (G) kuramına göre incelenmiştir. Dolayısıyla çalışmada uzman kaynaklı hataların yanı sıra tesadüfi hatalar da belirlenmiştir. Böylelikle araştırma bulguları, SOLO taksonomisi ve REBT'nin hangisinde uzmanlar arasındaki uyumun daha fazla olduğunu göstermesinin yanı sıra hangi taksonominin tesadüfi hatalardan etkilenmeye daha açık olduğu hakkında da bilgi sunmaktadır. Ayrıca çalışmada G kuramı kullanıldığından, araştırmada görev alan uzman sayısının arttırılması ya da azaltılması durumunda SOLO taksonomisine ve REBT'ye dayalı güvenirlik değerlerinin nasıl değiştiği belirlenmiştir. Bu sayede araştırma sonuçları, değerlendirme sorularının bilişsel düzeylerinin SOLO taksonomisine ve REBT'ye göre incelendiği çalışmalarda en uygun güvenirlik değerlerine en az kaç uzman ile ulaşılabileceği ya da uzman sayısının arttırılmasının iş gücünde yaratacağı kaybın güvenirlikte sağlayacağı kazanca değer olup olmadığı gibi sorulara cevap olabilmiştir. Sıralanan tüm bu hususlar sebebiyle, araştırmanın alanyazına önemli katkılarının olacağı tahmin edilmektedir.

Yöntem

Araştırma Deseni

Bu çalışma betimsel bir araştırma niteliğindedir. Betimsel araştırmalarda *ne* ve *nasıl* sorularına cevap aranır. Bu araştırmalar herhangi bir müdahaleye yer vermeksizin olanın olduğu gibi ortaya konulması esasına dayanır (Kumar, 2008). Betimsel çalışmalarda araştırmanın amacı doğrultusunda anket, gözlem, görüşme ya da doküman analizi gibi farklı yollarla toplanmış verilerden yararlanılabilmektedir (Anderson & Arsenaut, 2005; Erkuş, 2011). Bu çalışma kapsamında, verilerin elde edilmesinde doküman analizi tekniğinden faydalanılmıştır. Belge tarama şeklinde de isimlendirilen bu tekniğin ilk aşaması, araştırma konusu ile ilgili bilgi içeren materyallerin toplanmasıdır (Karasar, 2009). Toplanan belgelerin belli ölçütlere sahip oluş düzeyi bakımından incelenmesi ise doküman analizindeki ikinci aşamadır (Yıldırım & Şimşek, 1999). Doküman analizinde oldukça çeşitli belgeler incelenebilmektedir. Bu belgeler; kitap, dergi, gazete, günlük, resmi yayın ve istatistikler gibi yazılı materyaller olabildiği gibi konuyla ilgili film, video veya fotoğraflar gibi sesli/görüntülü kayıtlar da olabilmektedir (Cansız Aktaş, 2014).

Veri Kaynağı

Araştırmanın veri kaynağını; 2015 yılında Milli Eğitim Bakanlığı tarafından basılan Yedinci Sınıf Sosyal Bilgiler Ders Kitabı ile aynı yıl Yıldırım Yayıncılık tarafından basılan Sekizinci Sınıf Fen ve Teknoloji Ders Kitabındaki değerlendirme soruları oluşturmaktadır. Çalışma kapsamında Sosyal Bilgiler Ders Kitabından 18'i (%18.37) doğru-yanlış, 11'i (%11.22) eşleştirme, 43'ü (%43.88) çoktan seçmeli ve 26'sı (%26.53) yanıtı sınırlandırılmış veya uzun yanıtlı açık uçlu madde türünde olmak üzere toplam 98 değerlendirme sorusu incelenmiştir. Fen ve Teknoloji Ders Kitabından ise 198 değerlendirme sorusu analiz edilmiştir. Analizi yapılan değerlendirme sorularının 41'i (%20.71) doğru-yanlış, 45'i (%22.73) eşleştirme, 76'sı (%38.38) çoktan seçmeli ve 36'sı (%18.18) yanıtı sınırlandırılmış ya da uzun yanıtlı açık uçlu madde türündedir.

Verilerin Kodlanması

Çalışmada değerlendirme sorularının bilişsel düzeylerinin belirlenmesinde ikisi eğitim programları ve öğretim, biri ölçme değerlendirme alanından toplamda üç uzman görev almıştır. Uzmanlar, Sosyal Bilgiler ile Fen ve Teknoloji ders kitabındaki değerlendirme sorularını hem SOLO taksonomisinin hem de REBT'nin düzeylerine göre sınıflandırmıştır. Yapılan sınıflandırmalarda SOLO taksonomisinin ve REBT'nin düzeylerine ilişkin temel özellikler ile bu taksonomilerin düzeyleri için tanımlanan gösterge fillerden yararlanılmıştır. Sözü edilen gösterge fiillere Tablo 3'te yer verilmiştir.

Tablo 3.
SOLO Taksonomisi ve REBT'nin Gösterge Fiilleri.

	SOLO Taksonomisi[1]
Tek yönlü yapı	Ezberlemek, tanımlamak, adlandırmak, eşleştirmek, hatırlamak, aktarmak, çizmek, resmetmek, taklit etmek, tanımak.
Çok yönlü yapı	Sınıflandırmak, listelemek, tartışmak, göstermek, seçmek, basit hesaplamaları yapmak, anlatmak, rapor etmek, ayırmak, özetlemek.
İlişkisel yapı	Uygulamak, karşılaştırmak, tahmin etmek, ayırt etmek, organize etmek, analiz etmek, X ve Y gibi bilinmeyenler arasındaki ilişkileri hesaplamak, yorumlamak, gözden geçirip yeniden yazmak, yapılandırmak, çevirmek/dönüştürmek, çıkarımda bulunmak.
Soyutlanmış yapı	Hipotez kurmak, genellemelere varmak, oluşturmak, yansıtmak, bir teori ortaya koymak, icat etmek, orijinal bir ürün/çözüm yolu geliştirmek, farklı bir alana transfer etmek, ispatlamak.
	REBT[2]
Hatırlama	Adlandırmak, listelemek, hatırlamak, aktarmak, saymak, numaralandırmak, seçmek, çizmek, resmetmek.
Anlama	Eklemek, açıklamak, açıklığa kavuşturmak, göstermek, eşleştirmek, tahminde bulunmak, benzerlik ya da farklılıkları açıklamak.
Uygulama	Hesaplamak, çözmek, ihtiyaca göre düzenlemek, işlem basamaklarını takip etmek, düzenleyerek aktarmak, kabataslak çizmek.
Analiz	Ayırmak, özetlemek, ana hatlarıyla belirtmek, parçalayıp incelemek, öncelik sırasına koymak, sadeleştirmek, kategorilere ayırmak, test etmek, en uygun hale getirmek, çıkarım yapmak.
Değerlendirme	Değerlendirmek, kanıtlamak, yargıda bulunmak, savunmak, öneride bulunmak.
Yaratma	Organize etmek, bütün oluşturmak, tasarlamak, yeniden düzenlemek, uyarlamak, harmanlamak, keşfetmek, inşa etmek, kurmak.

[1] (Biggs & Tang, 2011), [2] (Thompson, 2012)

Tablo 3'te SOLO taksonomisinin yapı öncesi düzeyi yer almamaktadır. Bu durum, yapı öncesi düzeyde herhangi bir öğrenmenin olmaması ve dolayısıyla bu düzeye karşılık gelebilecek bir gösterge fiilin bulunmamasından kaynaklanmaktadır (Çetin & İlhan, 2016). Bu nedenle, çalışmada değerlendirme sorularının bilişsel seviyeleri incelenirken yapı öncesi düzeyi dikkate alınmamış; sınıflandırmalar SOLO taksonomisinin tek yönlü, çok yönlü, ilişkisel ve soyutlanmış yapı düzeylerine göre yapılmıştır.

Verilerin Analizi

Araştırmada; Sosyal Bilgiler ile Fen ve Teknoloji Ders Kitaplarındaki değerlendirme soruları üç uzman tarafından SOLO taksonomisi ve REBT'nin düzeylerine göre sınıflandırıldıktan sonra, elde edilen veriler G kuramına göre analiz edilmiştir. G kuramı (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), klasik test kuramının bir uzantısıdır (Brennan, 2003; Huang, 2009). G kuramı istatistiklerinde, öğrencilerin test puanlarına ilişkin varyansın değişkenliğe katkı sağladığı düşünülen bileşenlere ayrılması için ANOVA'dan yararlanılmaktadır (Renkl, 1993). Daha açık bir ifadeyle G kuramı analizlerinde öğrencinin test puanlarına ilişkin varyansın; gerçekten öğrencilerin yetenek düzeylerinin farklı olmasından kaynaklanan, puanlayıcı ve madde gibi farklı değişkenlik kaynaklarının etkisiyle oluşan, öğrenciler ile çeşitli değişkenlik kaynakları arasındaki etkileşim sonucunda ortaya çıkan, farklı değişkenlik kaynakları arasındaki etkileşimden dolayı oluşan, tanımlanamayan değişkenlik kaynaklarının etkisiyle açığa çıkan varyansa ve tesadüfî hatalara ayrılması amaçlanmaktadır. Bu amacı gerçekleştirmek için işe koşulan model ANOVA'dır (Taylor & Pastor, 2013). Ancak G kuramında geleneksel ANOVA uygulamalarından farklı olarak, kareler ortalamasının anlamlılığına değil; varyans bileşenlerinin büyüklüğüne odaklanılmaktadır (Bell, 1985).

G kuramının anlaşılabilmesi, kuramın temelinde yer alan bir takım kavramların bilinmesine bağlıdır. Ölçme objesi, yüzey, koşul, çapraz ve yuvalanmış desen, genellenebilirlik (*G)* ve karar (*K*) çalışmaları, genellenebilirlik ve güvenirlik katsayıları G kuramının kendine özgü terminolojisini oluşturan temel kavramlar arasında yer almaktadır. *Ölçme objesi*, ölçme çalışmasının hedefi durumundaki değişkenlik kaynağıdır. Tüm ölçme çalışmalarında ölçme objesini oluşturan bileşenlerin birbirinden etkili bir biçimde ayırt edilmesi; bir başka deyişle, ölçme objesine ilişkin varyansın yüksek olması istenmektedir. Pek çok ölçme durumunda sınava giren adaylar (bireyler) ölçmenin objesi konumundadır (Güler, Kaya-Uyanık, & Taşdelen-Teker, 2012). Ancak, bu araştırmada durum biraz farklıdır. Çalışmada değerlendirme sorularının bilişsel düzeylerinin belirlenmesi ve bilişsel düzeyleri farklı olan soruların birbirinden ayırt edilmesi amaçlandığından değerlendirme soruları ölçmenin objesini oluşturmaktadır. G kuramında ölçme objesi dışında ölçme sonuçlarında farklılaşmaya sebep olan değişkenlik kaynakları *yüzey* ve yüzeylere ait düzeylerin her biri *koşul* olarak tanımlanmaktadır. Örneğin, puanlayıcılar ölçme sonuçlarında hataya neden olabilen bir değişkenlik kaynağı ise puanlayıcı bir yüzey ve bu yüzeyde yer alan, birinci, ikinci… ve *n*. puanlayıcının her biri birer koşuldur (Mushquash & O'Connor, 2006). Bu çalışmada, değerlendirme sorularının bilişsel düzeylerini inceleyen uzmanlar ölçme işleminin yüzeyi durumundadır ve araştırmada üç uzman görev aldığından bu yüzeydeki koşul sayısı üçtür.

Genellenebilirlik kuramı istatistikleri, analizde kullanılan desen açısından *çapraz desenler* (crossed designs) ve *yuvalanmış desenler* (nested designs) olmak üzere iki kategoride incelenmektedir (Cardinet, Johnson, & Pini, 2010). Analizde yer alan herhangi bir yüzeyin bütün koşullarının bir başka yüzeyin bütün koşullarını etkilediği desenler *çapraz desen* olarak tanımlanmaktadır. Herhangi bir yüzeyin koşularından bazılarının, bir başka yüzeyin bazı koşullarınca gözlemlendiği desenler ise *yuvalanmış desen* olarak adlandırılmaktadır (Kumazawa, 2009). Çaprazlanmış desenler "x" ile gösterilirken; yuvalanmış desenler ":" ile temsil edilmektedir (Brennan, 2003). Örneğin, bireylerin matematik performansının değerlendirilmesinin amaçlandığı bir ölçme işleminde, bireyler "b" ve ölçme işleminde kullanılan matematik testindeki maddeler "m" ile ifade edilmiş olsun. Burada, tüm öğrencilerin aynı maddeleri cevaplandırmaları durumunda, çapraz bir desen söz konusu olmakta ve bu desen "b x m" olarak gösterilmektedir. Aynı ölçme işlemi için öğrencilerin her birinin farklı maddelerden sorumlu olması ise yuvalanmış desene karşılık gelmekte ve böyle bir desen " b : m" şeklinde tanımlanmaktadır (Hathcoat & Penn, 2012). Bu araştırmada, Sosyal Bilgiler ile Fen ve Teknoloji Ders Kitabındaki değerlendirme soruları

bilişsel düzeylerine göre sınıflandırılırken, soruların (s) tamamı araştırmada görev alan her üç uzman (u) tarafından da incelenmiştir. Dolayısıyla çalışmada G kuramı analizleri, "s x u" çapraz desenine göre gerçekleştirilmiştir.

G kuramı analizleri birbirinden ayrı fakat ilişkili iki adımdan meydana gelmektedir (Sun, Valiga, & Gao, 1997). G kuramı analizlerinin ilk adımını G, ikinci adımını ise K çalışmaları oluşturmaktadır (Shavelson & Webb, 1991). G çalışmalarının amacı, ölçme evreni ile ilgili olan varyans bileşenleri hakkında tahminler elde etmektir (Brennan, 2001). Bir başka deyişle G çalışmaları, gözlemlerin genellenebilirliğini etkileyen değişkenlik kaynaklarına ilişkin bilgi edinmek amacıyla gerçekleştirilir (Alkharusi, 2012). G kuramı analizlerinin ikinci adımını K çalışmaları oluşturmaktadır. K çalışmalarında, G çalışmasından elde edilen tahminler kullanılmaktadır (Webb, Shavelson, & Haertel, 2006). Bu adımda yeni gözlemler yapılması veya tekrar veri toplanması söz konusu değildir (Murray, 1984). K çalışmalarının amacı, hata varyansını tanımlayıp güvenirlik katsayısını tahmin ederek (Stora, Hagtvet, & Heyerdahl, 2013), madde, puanlayıcı ya da ölçme aracının uygulanma sayısının değiştirilmesi durumunda güvenirliğin nasıl değişeceğini belirlemektir (Mabe, 2013). Buna göre, K çalışmaları "Belirli bir yüzeydeki koşul sayısı arttırılırsa ya da azaltılırsa, ölçme işlemine farklı kaynaklardan karışan hatalar ve ölçme işleminin güvenirliği nasıl değişir?" sorusunu yanıtlamak amacıyla gerçekleştirilen bir çalışmadır (Briggs & Wilson, 2007). K çalışmaları sayesinde, ölçme işlemine karışacak hatayı minimum düzeye indirip güvenirliği maksimum düzeye çıkaracak alternatif tasarımların etkililiği değerlendirilir (Webb et al. 2006). Bu araştırmada G çalışmasıyla ölçme sonuçlarına ilişkin varyans; *i)* değerlendirme sorularının bilişsel düzeylerinin farklı olmasından kaynaklanan varyansa, *ii)* değerlendirme sorularının bilişsel düzeylerini inceleyen uzmanlardan kaynaklanan varyansa ve *iii)* değerlendirme soruları ile uzman etkileşiminin tesadüfi hatayla birlikte oluşturduğu artık varyansa ayrılmıştır. K çalışması kapsamında ise değerlendirme sorularının bilişsel düzeylerini inceleyen uzman sayısının bir, iki, dört, beş ve altı olduğu alternatif senaryolarda güvenirlik değerlerinin nasıl değiştiği test edilmiştir.

G kuramına ilişkin bir diğer önemli kavram genellenebilirlik (G) ve güvenirlik katsayısı (Phi) ayırımıdır. G kuramında, güvenirliği belirlemede bağıl ve mutlak olmak üzere iki tür karar vermenin söz konusu olduğu dikkate alınmakta ve bağıl değerlendirmeler için *G*, mutlak değerlendirmeler için Phi katsayısı hesaplanmaktadır (Atılgan, 2005). Dolayısıyla çalışmada değerlendirme sorularının bilişsel düzeyleri için SOLO taksonomisine ve REBT'ye göre yapılan sınıflamaların güvenirliği incelenirken hem G hem de Phi katsayı rapor edilmiştir. Araştırmada G kuramına ilişkin analizlerin tamamında EduG paket programı kullanılmıştır.

Bulgular

Bu bölümde araştırmada ulaşılan bulgulara yer verilmiştir. İlk olarak, Yedinci Sınıf Sosyal Bilgiler Ders Kitabındaki değerlendirme sorularının üç uzman tarafından SOLO taksonomisine ve REBT'ye göre sınıflandırılmasıyla elde edilen veriler analiz edilmiştir. Çaprazlanmış desen (s x u) ile gerçekleştirilen G çalışması sonucunda kestirilen varyans bileşenleri ve bu bileşenlerin toplam varyansı açıklama yüzdeleri Tablo 4'te sunulmuştur.

Tablo 4'teki bulgulara göre, değerlendirme sorusu ana etkisi için kestirilen varyans bileşeninin toplam varyans içerisindeki oranı SOLO taksonomisine göre yapılan sınıflamada %95.30, REBT'ye göre yapılan sınıflamada ise %71.60 olarak bulunmuştur. Bu değerler hem SOLO taksonomisine hem de REBT'ye göre yapılan sınıflamada toplam varyansa katkısı en yüksek olan varyans bileşeninin değerlendirme sorusu ana etkisine ait olduğunu göstermektedir. Değerlendirme sorusu ana etkisinin ardından uzman ana etkisi için kestirilen varyans bileşenleri incelenmiştir. SOLO taksonomisine göre yapılan sınıflamada uzman ana etkisine ilişkin varyans değeri –.00028 olarak bulunmuştur.

G kuramında teorik olarak tüm varyans bileşenlerinin pozitif olması beklenmektedir (Cardinet et al. 2010). Buna karşın uygulamada; örneklem yetersizliği, hatalı bir modelin kullanılması veya evren değerinin gerçekten sıfır ya da sıfıra yakın olması gibi sebeplerle negatif varyans bileşenleriyle karşılaşılabilmektedir (Chiu, 2001).

Tablo 4.
Sosyal Bilgiler Ders Kitabındaki Değerlendirme Sorularının SOLO Taksonomisine ve REBT'ye Göre Sınıflandırılmasıyla Elde Edilen Veriler için Kestirilen Varyans Bileşenleri.

Taksonomi	Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	Varyans Bileşeni	Varyans Yüzdesi
SOLO	Değerlendirme sorusu (s)	187.21	97	1.93	.63	95.30
	Uzman (u)	.01	2	.00	.00*	.00
	s x u, e	5.99	194	.03	.03	4.70
	Toplam	193.21	293			100
REBT	Değerlendirme sorusu (s)	409.73	97	4.22	1.26	71.60
	Uzman (u)	11.07	2	5.53	.05	3.00
	s x u, e	86.93	194	.45	.45	25.50
	Toplam	507.73	293			100

* Negatif (–.00028) varyans bileşenlerine sıfır değeri atanmıştır.

Varyans bileşenleri arasında, negatif ama sıfıra yakın bir değer gözlendiğinde ilgili varyans bileşeninin sıfır olarak alınması önerilmektedir. Negatif ve sıfırdan uzak bir varyans bileşeninin gözlenmesi durumunda ise veri setinde hata olup olmadığının kontrol edilmesi tavsiye edilmektedir (Webb, Rowley, & Shavelson, 1988). Araştırmada, SOLO taksonomisine dayalı sınıflamada uzman ana etkisi için hesaplanan negatif varyans sıfıra oldukça yakın olduğundan söz konusu varyans için sıfır değeri atanmış ve sonrasında hesaplanan varyans bileşenlerinin toplam varyans içerisindeki yüzdelerine bakılmıştır. Tablo 4'te görüldüğü üzere, SOLO taksonomisine göre yapılan sınıflamada uzmanlara ilişkin varyansın toplam varyans içerisindeki etkisi %.00 olarak saptanmıştır. REBT'ye göre yapılan sınıflamada ise toplam varyansın %3.00'ünün uzman ana etkisinden kaynakladığı tespit edilmiştir.

G kuramında ölçme objesi ile analizde işlem gören yüzeylerin tamamının etkileşimine ait varyans değeri, tanımlanamayan değişkenlik kaynaklarıyla birlikte artık varyansı meydana getirmektedir (Shavelson & Webb, 1991). Bu çalışmada, değerlendirme soruları ölçmenin objesi ve uzmanlar analizde işlem gören tek yüzey olduğundan "*s x u, e"* varyansı, artık varyansa karşılık gelmektedir. Tablo 4'teki artık varyans oranlarına bakıldığında SOLO taksonomisine dayalı sınıflandırmalarda %4.70 iken; REBT'ye göre yapılan sınıflamada %25.50 olduğu görülmektedir.

Varyans bileşenlerinin kestirilmesinden sonra, Sosyal Bilgiler Ders Kitabındaki değerlendirme sorularının üç uzman tarafından SOLO taksonomisinin ve REBT'nin düzeylerine göre analiz edilmesiyle elde edilen veriler için G ve Phi katsayıları hesaplanmıştır. Orijinal veriler için kestirilen G ve Phi katsayıları, değerlendirme sorularının bilişsel düzeylerinin belirlenmesinde görev alan uzman sayısının arttırılıp azaltılmasına ilişkin alternatif senaryolar karşısında hesaplanan G ve Phi katsayıları ile birlikte Tablo 5'te sunulmuştur.

Tablo 5.
Sosyal Bilgiler Ders Kitabındaki Değerlendirme Sorularının SOLO Taksonomisine ve REBT'ye Göre Sınıflandırılmasıyla Elde Edilen Veriler için Alternatif Karar Çalışmasında Kestirilen G ve Phi Katsayıları.

Taksonomi	G ve Phi Katsayıları	Uzman Sayısı					
		1	**2**	**3**	**4**	**5**	**6**
SOLO	G	.95	.98	**.98**	.99	.99	.99
	Phi	.95	.98	**.98**	.99	.99	.99
REBT	G	.74	.85	**.89**	.92	.93	.94
	Phi	.72	.83	**.88**	.91	.93	.94

Tablo 5'e göre; Sosyal Bilgiler Ders Kitabındaki değerlendirme sorularının bilişsel düzeyleri için SOLO temelli sınıflamalarda hesaplanan G ve Phi katsayıları, REBT'ye dayalı olarak yapılan sınıflamalara ait G ve Phi katsayılarından daha yüksek bulunmuştur. Alternatif karar çalışmalarına ilişkin bulgular, uzman

sayısında yapılan değişikliklerin G ve Phi katsayıları üzerinde oluşturduğu etkinin SOLO taksonomisine dayalı sınıflandırmalarda REBT esas alınarak yapılan sınıflandırmalara göre çok daha küçük olduğunu ortaya koymaktadır.

Sosyal Bilgiler Ders Kitabı'ndaki değerlendirme sorularının bilişsel düzeylerinin tespitine ilişkin veriler G kuramına göre analiz edildikten sonra, benzer işlemler Sekizinci Sınıf Fen ve Teknoloji Ders Kitabındaki değerlendirme sorularının bilişsel düzeylerinin sınıflandırılmasıyla elde edilen veriler üzerinden gerçekleştirilmiştir. Bu sayede SOLO taksonomisi ile REBT'den hangisinin daha güvenilir olduğuna ilişkin ulaşılacak sonuçlarda bilişsel düzeyleri incelenen değerlendirme sorularının tek bir derse ait olmasından dolayı oluşabilecek sınırlılıkların önüne geçilmesi hedeflenmiştir. Fen ve Teknoloji Ders Kitabındaki değerlendirme sorularının bilişsel düzeylerine yönelik olarak üç farklı uzman tarafından yapılan sınıflandırmalar çaprazlanmış desene (s x u) göre analiz edilmiştir. Analizler G ve K çalışması şeklinde iki adımda yürütülmüştür. Analizin ilk adımı olan G çalışmasında, varyans bileşenleri ve her bir varyans bileşeninin toplam varyans içerisindeki yüzdesi hesaplanmıştır. Hesaplanan değerler Tablo 6'da sunulmuştur.

Tablo 6.
Fen ve Teknoloji Ders Kitabındaki Değerlendirme Sorularının SOLO Taksonomisine ve REBT'ye Göre Sınıflandırılmasıyla Elde Edilen Veriler için Kestirilen Varyans Bileşenleri.

Taksonomi	Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	Varyans Bileşeni	Varyans Yüzdesi
SOLO	Değerlendirme sorusu (s)	251.83	197	1.28	.42	92.80
	Uzman (u)	.04	2	.02	.00*	.00
	s x u, e	12.62	394	.03	.03	7.20
	Toplam	264.50	593			100
REBT	Değerlendirme sorusu (s)	285.60	197	1.45	.43	73.50
	Uzman (u)	3.84	2	1.92	.01	1.50
	s x u, e	58.16	394	.15	.15	25.00
	Toplam	347.60	593			100

* Negatif (–.00005) varyans bileşenlerine sıfır değeri atanmıştır.

Tablo 6'daki bulgulara göre, değerlendirme sorusu ana etkisi için kestirilen varyans bileşeni SOLO taksonomisine dayalı sınıflamada %92.80'lik bir oranla toplam varyans içerisinde en yüksek paya sahiptir. Benzer şekilde, REBT kullanılarak yapılan sınıflamada değerlendirme sorusu ana etkisi %73.50'lik bir oran ile toplam varyans içerisindeki en büyük bileşeni temsil etmektedir. Tablo 6'da, değerlendirme sorusu ana etkisinin altında uzman ana etkisi için kestirilen varyans bileşenleri yer almaktadır. Bu varyans bileşenlerinin toplam varyans içerisindeki oranları, SOLO taksonomisine göre yapılan sınıflamada %.00; REBT'ye göre yapılan sınıflamada ise %1.50 olarak hesaplanmıştır. Artık (*s x u, e*) varyans bileşenleri, SOLO taksonomisine dayalı sınıflamada toplam varyansın %7.20'lik bir bölümünü oluştururken; REBT'ye dayalı sınıflamada toplam varyansın %25.00'lik bir kısmını meydana getirmektedir.

Fen ve Teknoloji Ders Kitabındaki değerlendirme sorularının bilişsel düzeylerinin SOLO taksonomisinin ve REBT'nin düzeylerine göre incelenmesiyle elde edilen veriler üzerinden gerçekleştirilen K çalışması sonuçları Tablo 7'de sunulmuştur. Tablo 7'de, çalışmadaki orijinal uzman sayısı (üç) için hesaplanan G ve Phi katsayıları; uzman sayısının bir, iki, dört, beş ve altı olduğu senaryolarda kestirilen G ve Phi katsayılarıyla bir arada verilmiştir.

Tablo 7'de görüldüğü gibi, üç uzmanın Fen ve Teknoloji Dersi değerlendirme sorularının bilişsel düzeyleri için SOLO taksonomisi kullanılarak yaptıkları sınıflamalarda G ve Phi katsayıları .97 değeri ile birbirine eşit bulunmuştur. REBT'ye dayalı sınıflamada ise G ve Phi katsayıları sırasıyla .90 ve .89 şeklindedir. Alternatif K çalışmalarına ilişkin bulgulara bakıldığında SOLO taksonomisine dayalı sınıflamalarda tek bir uzman için G ve Phi katsayılarının .93 olarak hesaplandığı, uzman sayısı arttırıldıkça G ve Phi katsayılarının da arttığı ve altı puanlayıcı için .99 olarak elde edildiği görülmektedir. REBT'ye

dayalı sınıflandırmalarda ise tek bir uzman için hesaplanan G ve Phi katsayıları sırasıyla .75 ve 73 olarak kestirilmiştir. Uzman sayısı arttıkça G ile Phi katsayılarının arttığı belirlenmiş, en büyük artışın ise uzman sayısının birden ikiye çıktığı ölçme koşullarında gözlendiği tespit edilmiştir.

Tablo 7.
Fen ve Teknoloji Ders Kitabındaki Değerlendirme Sorularının SOLO Taksonomisine ve REBT'ye Göre Sınıflandırılmasıyla Elde Edilen Veriler için Alternatif Karar Çalışmasında Kestirilen G ve Phi Katsayıları.

	G ve Phi	Uzman Sayısı					
Taksonomi	**Katsayıları**	**1**	**2**	**3**	**4**	**5**	**6**
SOLO	G	.93	.96	**.97**	.98	.98	.99
	Phi	.93	.96	**.97**	.98	.98	.99
REBT	G	.75	.85	**.90**	.92	.94	.95
	Phi	.73	.85	**.89**	.92	.93	.94

Tartışma ve Sonuç

Bu araştırmada, değerlendirme sorularının bilişsel düzeylerinin tespiti için SOLO taksonomisi ve REBT kullanılarak yapılan sınıflamalara ilişkin veriler G kuramına göre analiz edilmiştir. G çalışması sonucunda, her iki taksonomide de değerlendirme sorusu ana etkisi için kestirilen varyans bileşenlerinin toplam varyans içerisinde en büyük paya sahip olduğu belirlenmiştir. Araştırmada, değerlendirme soruları ölçmenin objesi konumunda olup ölçme objesi için kestirilen varyansın toplam varyans içerisinde en büyük paya sahip olması istenilen bir durumdur (Güler et al. 2012). Dolayısıyla değerlendirme sorularının bilişsel düzeylerinin belirlenmesinde SOLO taksonomisi ve REBT'den hangisi kullanılırsa kullanılsın soruların bilişsel düzeyleri arasındaki farklılıkların ortaya konulabildiği söylenebilir. Ancak araştırmada değerlendirme sorusu ana etkisine ait varyans yüzdesinin SOLO temelli sınıflamada REBT kullanılarak yapılan sınıflamaya kıyasla daha yüksek olduğu gözden kaçırılmamalıdır. Değerlendirme sorusu ana etkisine ait varyans yüzdeleri arasında SOLO taksonomisinin lehine bir fark olması, SOLO taksonomisinin değerlendirme sorularının bilişsel düzeylerinin tespitinde daha etkili bir model olduğuna işaret etmektedir.

Araştırmada değerlendirme sorularının bilişsel düzeyleri için hem SOLO temelli hem de REBT'ye göre yapılan sınıflamalarda uzman ana etkisinin toplam varyans içerisinde küçük bir yüzdeye karşılık geldiği saptanmıştır. Bu bulguya dayanarak değerlendirme sorularının bilişsel düzeylerinin tespitinde ister SOLO taksonomisi isterse REBT kullanılsın uzmanların soruları atadıkları bilişsel düzeyler arasında önemli farklılıkların bulunmadığı ifade edilebilir. Uzmanlardan kaynaklanan varyans bileşenleri karşılaştırıldığında iki taksonomi arasında küçük de olsa bir fark bulunduğu ve uzmanlar arasındaki uyumun SOLO taksonomisinde REBT'ye göre daha yüksek olduğu sonucuna varılmıştır. Bu bulgu Hattie ve Purdie (1994) tarafından yapılan çalışmanın sonuçları ile örtüşmektedir. Hattie ve Purdie'nin (1994) yaptığı çalışmada, çoktan seçmeli soruların bilişsel düzeyleri SOLO taksonomisine dayalı olarak belirlendiğinde uzmanlar arasındaki uyumun Bloom taksonomisinin kullanıldığı sınıflamalara göre daha yüksek olduğu saptanmıştır. İlhan ve Çetin (2016) ile Çetin, Boran ve Yazıcı'nın (2014) SOLO taksonomisine dayalı rubrikler hakkındaki puanlayıcı görüşlerini inceledikleri çalışmaların sonuçları da bu araştırmadan elde edilen bulguları destekler niteliktedir. İlhan ve Çetin (2016) açık uçlu sorulara verilen öğrenci cevaplarını SOLO taksonomisine dayalı rubrikler ile puanlayan öğretmenlerin bu rubrikler hakkındaki görüşlerini analiz etmiştir. Çalışmaya katılan öğretmenler SOLO'nun açık ve anlaşılır düzeylerden oluşan bir taksonomi olduğu yönünde görüş bildirmiştir. Benzer şekilde, Çetin vd. (2014) tarafından yapılan çalışmada öğrencilerin açık uçlu fizik sorularına verdikleri cevaplar üç öğretim elemanı tarafından SOLO temelli rubrikler yardımıyla puanlanmıştır. Çalışmada her üç öğretim elemanı da SOLO taksonomisinin objektifliği yüksek bir model olduğunu ifade etmiştir.

Chan vd. (2002) tarafından yapılan çalışmanın sonuçları ise bu araştırmada ulaşılan bulgulardan farklılık göstermektedir. Chan vd. (2002), SOLO ve Bloom taksonomilerine dayalı rubriklerde puanlayıcılar arası güvenirliği incelemiştir. Çalışmada Bloom taksonomisine göre oluşturulan rubriklerde

puanlayıcılar arası korelasyon katsayısının SOLO temelli rubriklere göre daha yüksek olduğu sonucuna ulaşılmıştır. Bu araştırmadan elde edilen bulgular ile Chan vd.'nin (2002) çalışmasında ulaşılan sonuçlar arasındaki fark, SOLO taksonomisi açık ve anlaşılır düzeylerden oluşmasına rağmen bu taksonomiye aşina olmayan kişilerin taksonomiye uygun puanlama yapmakta zorlanmasıyla (Leung, 2000) ilgili olabilir. Bloom ise araştırmacılar ve eğitimciler tarafından tanınırlığı daha yüksek bir taksonomi olduğundan (Arı, 2013) bahsi geçen durumun Bloom taksonomisi için ortaya çıkması daha zayıf bir ihtimal olarak görülmektedir. Bu araştırmada; uygulama öncesinde değerlendirme sorularının bilişsel düzeylerini inceleyen uzmanlara, her iki taksonominin düzeyleri ve düzeylere karşılık gelen gösterge fiiller ana hatlarıyla açıklanarak taksonomilerin uzmanlar tarafından tanınırlıklarının farklı olmasından dolayı oluşabilecek etkiler kontrol altına alınmaya çalışılmıştır.

G çalışmasında hesaplanan artık varyans yüzdeleri incelendiğinde SOLO taksonomisine dayalı sınıflandırmalarda tesadüfi hatanın düşük olduğu, REBT esas alınarak yapılan sınıflamalarda ise toplam varyansın dörtte birlik gibi önemli bir kısmının artık varyanstan oluştuğu tespit edilmiştir. Artık varyans, ölçme işlemine karışan tesadüfi hata miktarı hakkında bilgi vermekte ve sıfıra olabildiğince yakın olması istenmektedir (Güler & Taşdelen-Teker, 2015). Buna göre, SOLO taksonomisi ile karşılaştırıldığında REBT'nin tesadüfi hatalara daha açık bir model olduğu ifade edilebilir. Bir başka deyişle REBT kullanılarak gerçekleştirilen sınıflamalarda, uzmanların sistematik olmayan bir şekilde değerlendirme sorularını farklı düzeylere ataması ve/veya uzmanlar dışında başka faktörlerin de ölçme sonuçlarını etkilemesi daha olası bir durumdur. Yani değerlendirme sorularının bilişsel düzeylerinin tespitinde REBT kullanıldığında kaynağı bilinmeyen faktörlerden dolayı ölçme işlemine karışan hata miktarı SOLO temelli sınıflandırmalara göre daha yüksek çıkmaktadır.

Son olarak, SOLO taksonomisine dayalı sınıflandırmalar için hesaplanan G ve Phi katsayıları, REBT esas alınarak yapılan sınıflandırmalara ait G ve Phi katsayılarından daha yüksek bulunmuştur. Alternatif K çalışmalarında, SOLO taksonomisinde tek bir uzman ile elde edilen güvenirlik değerlerini REBT'nin yaklaşık beş ya da altı uzman ile üretebildiği saptanmıştır. Tüm bu sonuçlar, değerlendirme sorularının bilişsel düzeylerinin belirlenmesinde SOLO taksonomisinin REBT'ye göre genellenebilirliği ve güvenirliği daha yüksek ölçümler ürettiğini göstermektedir.

Sınırlılıklar ve Öneriler

SOLO taksonomisi ve REBT'nin karşılaştırılmasına yönelik alanyazındaki çalışmalarda ele alınmayan bazı noktalar üzerinde durması ve konuyla ilgili önceki çalışmalarda kullanılan klasik test kuramına dayalı yöntemlerden farklı olarak G kuramına göre yürütülmüş olması, bu araştırmayı alanyazına katkı sağlayacak özgün bir çalışma haline getirmektedir. Bununla birlikte, araştırmanın bir takım sınırlılıklarının bulunduğu göz ardı edilmemelidir. Çalışmanın sınırlılıkları ve bu sınırlılıkların aşılmasına yönelik ileri araştırma önerileri şu şekilde sıralanabilir: Öncelikle bu araştırma, SOLO taksonomisi ve REBT'nin karşılaştırılması ile sınırlandırılmıştır. Bu kapsamda, program öğelerinin bilişsel düzeylerinin belirlenmesi amacıyla geliştirilmiş olan farklı taksonomiler (Haladyana, Marzano, Fink, Dettmer, MATH) için benzer çalışmaların yapılması önerilebilir. İkinci olarak, bu araştırmada SOLO taksonomisi ve REBT'nin güvenirliklerinin karşılaştırılmasında Sosyal Bilgiler ile Fen ve Teknoloji dersine yönelik değerlendirme soruları kullanılmıştır. Araştırmada tek bir derse yönelik değerlendirme soruları yerine iki farklı derse ait değerlendirme sorularının kullanılmış olması, çalışmadan elde edilen bulguların genellenebilirliğine katkı sağlamaktadır. Sosyal Bilgiler dersinin Tarih ve Coğrafya, Fen ve Teknoloji dersinin Fizik, Kimya ve Biyoloji disiplinlerini içeren geniş alanlı bir tasarıma sahip olması da SOLO taksonomisi ile REBT'nin güvenirliğine ilişkin ulaşılan bulguları büyük ölçüde disiplinden bağımsız hale getirmektedir. Ancak yine de Matematik ve Yabancı Diller gibi çalışmaya dâhil edilmeyen farklı disiplinler üzerinde bu tür bir çalışmanın yapılması önerilebilir.

References

Alkharusi, H. (2012). Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education*, *2*(1), 184–196. http://dx.doi.org/10.5296/jse.v2i1.1227.

Alsaadi, A. (2011). A comparison of primary mathematics curriculum in England and Qatar: The SOLO taxonomy. *Proceedings of the British Society for Research into Learning Mathematics*, *21*(3). Retrieved from http://www.bsrlm.org.uk/wp-content/uploads/2016/02/BSRLM-IP-21-3-1.pdf.

Anderson, G., & Arsenaut, N. (2005). *Fundamentals of educational research*. London, UK: The Falmer.

Anderson, L.W., & Krathwohl, D.R. (2001). *A taxonomy for learning teaching and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Arı, A. (2013). Bilişsel alan sınıflamasında yenilenmiş Bloom, Solo, Fink, Dettmer taksonomileri ve uluslararası alanda tanınma durumları. *Uşak Üniversitesi Sosyal Bilimler Dergisi, 6*(2), 259–290. http://dx.doi.org/10.12780/UUSBD164.

Atılgan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. Unpublished doctoral dissertations, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Atılgan, H. (2005). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama*, *4*(7), 95-108.

Bağdat, O., & Anapa-Saban, P. (2014). İlköğretim 8. sınıf öğrencilerinin cebirsel düşünme becerilerinin SOLO taksonomisi ile incelenmesi. *The Journal of Academic Social Science Studies, 26*, 473–496. http://dx.doi.org/10.9761/JASSS2364.

Başol, G., Balgalmış, E., Karlı, M.G., & Öz, F.B. (2016). TEOG sınavı matematik sorularının MEB kazanımlarına, TIMSS seviyelerine ve yenilenen Bloom Taksonomisine göre incelenmesi. *Uluslararası İnsan Bilimleri Dergisi*, *13*(3), 5945–5965. http://dx.doi.org/10.14687/jhs.v13i3.4326

Bell, J.F. (1985). Generalizability theory: Software problem. *Journal of Educational Statistics*, *10*(1), 19–29. http://dx.doi.org/10.3102/10769986010001019.

Biggs, J.B. (1979). Indıvıdual differences in study processes and the quality of learning outcomes. *Higher Education*, *8*(4), 381–394. http://dx.doi.org/10.1007/BF01680526.

Biggs, J.B. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364. http://dx.doi.org/10.1007/BF00138871.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic.

Biggs, J.B., & Tang, C. (2011). *Teaching for quality learning at university*. Maidenhead, UK: Open University.

Brabrand, C., & Dahl, B. (2009). Using the SOLO-taxonomy to analyze competence progression of university science curricula. *Higher Education, 58*(4), 531–549. http://dx.doi.org/10.1007/s10734-009-9210-4.

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

Brennan, R.L. (2003). *Coefficients and indices in generalizability theory*. CASMA Center for Advanced Studies in Measurements and Assessments. Research Report No.1

Briggs, D., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*(2), 131–155. http://dx.doi.org/10.1111/j.1745-3984.2007.00031.x

Cansız Aktaş, M. (2014). Nitel veri toplama araçları. In M. Metin (Ed.), *Kuramdan uygulamaya eğitimde bilimsel araştırma yöntemleri* (pp. 337–371). Ankara: Pegem Akademi.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Routledge.

Chan, C.C., Tsui, M.S., Mandy, Y.C., & Hong, J.H. (2002). Applying the structure of the observed learning outcomes (SOLO) taxonomy on student's learning outcomes: An empirical study. *Assessment and Evaluation in Higher Education*, *27*(6), 511–527. http://dx.doi.org/10.1080/0260293022000020282.

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: Wiley.

Çalışkan, H. (2011). Öğretmenlerin hazırladığı sosyal bilgiler dersi sınav sorularının değerlendirilmesi. *Eğitim ve Bilim*, *36*(160), 120–132.

Çetin, B., Boran, A., & Yazıcı, N. (2014). Fizik eğitiminde başarının ölçülmesinde SOLO taksonomisine göre hazırlanan rubriklerin incelenmesi. Bayburt Eğitim Fakültesi Dergisi, *9*(2), 32–71.

Çetin, B., & İlhan, M. (2016). SOLO taksonomisi. In E. Bingölbali, S. Arslan, & İ.Ö. Zembat (Ed.), *Matematik eğitiminde teoriler* (pp. 861–879). Ankara: Pegem Akademi.

Demirel, Ö. (2012). *Kuramdan uygulamaya eğitimde program geliştirme*. Ankara: Pegema.

Dindar, H., & Demir, M. (2006). Beşinci sınıf öğretmenlerinin fen bilgisi dersi sınav sorularının Bloom taksonomisine göre değerlendirilmesi. *Gazi Eğitim Fakültesi Dergisi*, *26*, 87–96.

Doğan, N. (2002). *Klasik test teorisi ve örtük özellikler kuramının örneklemeler bağlamında karşılaştırılması*. Unpublished doctoral dissertations, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Erkuş, A. (2011). *Davranış bilimleri için bilimsel araştırma süreci*. Ankara: Seçkin Yayıncılık.

Fensham, P., & Bellocchi, A. (2013). Higher order thinking in chemistry curriculum and ıts assessment. *Thinking Skills and Creativity*, *10*, 250–264. http://dx.doi.org/10.1016/j.tsc.2013.06.003

Gezer, M., & İlhan, M. (2014). 8. Sınıf vatandaşlık ve demokrasi eğitimi dersi kazanımları ile değerlendirme sorularının SOLO taksonomisine göre incelenmesi. *Doğu Coğrafya Dergisi*, *19*(32), 193–207. http://dx.doi.org/10.17295/dcd.88376.

Gezer, M., & İlhan, M. (2015). Sosyal bilgiler dersi öğretim programı kazanımları ile ders kitabı değerlendirme sorularının SOLO taksonomisine göre incelenmesi. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi*, *29*, 1–25.

Göçer, A., & Kurt, A. (2016). Türkçe dersi öğretim programı 6, 7 ve 8. sınıf sözlü iletişim kazanımlarının SOLO taksonomisine göre incelenmesi. *Bitlis Eren Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *5*, 215–228.

Gökler, Z.S., Aypay, A. & Arı, A. (2012). İlköğretim İngilizce dersi hedefleri kazanımları SBS soruları ve yazılı sınav sorularının yeni Bloom taksonomisine göre değerlendirilmesi. *Eskişehir Osmangazi Üniversitesi Eğitimde Politika Analizi Dergisi*, *1*(2), 114–133.

Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Genellenebilirlik Kuramı*. Ankara: Pegem Akademi.

Güler N., & Taşdelen Teker, G. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenirliğin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1), 12–24. http://dx.doi.org/10.21031/epod.63041.

Hathcoat J.D., & Penn, J.D. (2012). Generalizability of student writing across multiple tasks: A challenge for authentic assessment. *Research and practice in assessment*, *7*, 16-28.

Hattie, J. A., & Purdie, N. (1994). *Using the SOLO taxonomy to classify test items*. Unpublished manuscript, University of Western Australia, Graduate School of Education, Perth, Aus.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies*, *5*(1), 1–17.

İlhan, M., & Çetin, B. (2016). The identification of the views of raters on standard rubrics and rubrics based on the SOLO taxonomy. *Eğitimde Kuram ve Uygulama*, *12*(1), 1–16.

Jones, B.L., Collis, K.F., & Watson, J.M. (1993). Towards a theoretical basıs for students' alternative frameworks in science and for scıence teaching. *Research in Science Education*, *23*, 126–135. http://dx.doi.org/10.1007/BF02357053.

Karasar, N. (2009). *Bilimsel Araştırma Yöntemi.* Ankara: Nobel.

Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218. http://dx.doi.org/10.1207/s15430421tip4104_2.

Kumar, R. (2008). *Research methodology.* New Delhi: Balaji Offset.

Kumazawa, T. (2009). Revision of a criterion-referenced vocabulary test using generalizability theory. *Japan Association for Language Teaching (JALT), 31*(1), 81–100.

Leung, C.F. (2000). Assessment for learning: Using SOLO taxonomy to measure design performance of design & technology students. *International Journal of Technology and Design Education*, *10*(2), 149–161. http://dx.doi.org/10.1023/A:1008937007674.

Lucas, U., & Mladenovic, R. (2009). The identification of variation in students' understandings of disciplinary concepts: The application of the SOLO taxonomy within introductory accounting. *Higher Education*, *58*(2), 257–283. http://dx.doi.org/10.1007/s10734-009-9218-9.

Mabe, M. (2013). *Progress monitoring for prerequisite social skills: A generalizability study for measure development*. Unpublieshed master thesis, University of Rhode Island, Rhode Island, ABD.

Milli Eğitim Bakanlığı. (2015). *İlköğretim sosyal bilgiler 7. sınıf ders kitabı*. Ankara: Devlet Kitapları Müdürlüğü Basımevi.

Murray, R.P. (1984, July). *Application of generalizability theory in the development of quality of care measurement*. Paper presented at the Third International Conference on System Science in Health Care, Munih. http://dx.doi.org/10.1007/978-3-642-69939-9_198.

Mushquash, C., & O'Connor, B.P. (2006). SPSS and SAS Programs for generalizability theory analyses. *Behavior Research Methods*, *38*(3), 542–547. http://dx.doi.org/10.3758/BF03192810.

Näsström, G. (2008). *Measurement of alignment between standards and assessment*. Retrieved from http://umu.diva-portal.org/smash/get/diva2:142244/FULLTEXT01.

Özdemir, A.S., & Göktepe-Yıldız, S. (2015). The analysis of elementary mathematics preservice teachers' spatial orientation skills with SOLO model. *Eurasian Journal of Educational Research*, *61*, 217–236. http://dx.doi.org/10.14689/ejer.2015.61.12.

Pegg, J., & Tall, D. (2004, December). *Fundamental cycles in learning algebra: An analysis*. Paper presented to the 12th ICMI Study Conference on the Future of the Teaching and Learning of Algebra. Melbourne, Australia.

Peter, F., & Alberto, B. (2013). Higher order thinking in chemistry curriculum and its asessment. *Thinking Skills and Creativity*, *10*, 250–264. http://dx.doi.org/10.1016/j.tsc.2013.06.003.

Renkl, A. (2003). The dependability of test scores: Generalizability theory and hierarchical models. In D. Leclercq, and J.E. Bruno (Eds.), *Item banking: Interactive testing and self-assessment* (pp. 167–176). Berlin: Springer.

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

Sönmez, V. (2001). *Program geliştirmede öğretmen elkitabı.* Ankara: PegemA.

Stora, B., Hagtvet, K.A., & Heyerdahl, S. (2013). Observations of families in structured interactions: Parenting therapists provide reliable ratings of mothers' parenting. *Personality and Social Psychology, 23*(4), 448–463. http://dx.doi.org/10.1080/10503307.2012.733830.

Sun, A., Valiga, M.J., & Gao, X. (1997). Using generalizability theory to assess the reliability of student ratings of academic advising. *The Journal of Experimental Education*, *65*(4), 367–379. http://dx.doi.org/10.1080/00220973.1997.10806611.

Tarman, B., & Kuran, B. (2015). Examination of the cognitive level of questions in social studies textbooks and the views of teachers based on Bloom taxonomy. *Kuram ve Uygulamada Eğitim Bilimleri*, *15*(1), 213–222. http://dx.doi.org/10.12738/estp.2015.1.2625.

Taylor, M.A., & Pastor, D.A. (2013). *An application of generalizability theory to evaluate the technical quality of an alternate assessment*. *Applied Measurement in Education*, *26*(4), 279–297, http://dx.doi.org/10.1080/08957347.2013.824450.

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme.* Ankara: Yargı.

Thompson, J.G. (2012). *First year teacher's survival guide: Ready to use strategies, tools & activities for meeting the challenges of each school day*. San Francisco, CA: Josse-Bass.

Turgut, M.F., & Baykul Y. (2012). *Eğitimde ölçme ve değerlendirme.* Ankara: PegemA.

Webb, N.M., Rowley, G.L., & Shavelson, R.J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development*, 21, 81-90.

Webb, N.M., Shavelson, R.J, & Haertel, E.H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao and S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics*, *26*, 81–124. Amsterdam: Elsevier B. V. http://dx.doi.org/10.1016/S0169-7161(06)26004-8.

Yanmaz, E. (Ed.). (2015). *İlköğretim fen ve teknoloji dersi 8. sınıf ders kitabı*. Ankara: Yıldırım.

Yıldırım, A., & Simşek H. (2011). *Sosyal bilimlerde nitel araştırma yöntemleri.* Ankara: Seçkin.

Zorluoğlu, S.L., Kızılaslan, A., & Sözbilir, M. (2016). Ortaöğretim kimya dersi öğretim programı kazanımlarının yapılandırılmış Bloom taksonomisine göre analizi ve değerlendirilmesi. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi (EFMED)*, *10*(1), 260–279. http://dx.doi.org/10.17522/nefefmed.22297.