Analysis of Mathematics Teachers' Class Management Behaviors in Turkey, Bulgaria and Greece using Anchoring Vignettes

Assist.Prof.Dr . Meltem Yurtçu
İnönü University - Turkey
ORCID: 0000-0003-3303-5093
meltem.yurtcu@inonu.edu.tr

Prof.Dr. Nuri Doğan
Hacettepe University - Turkey
ORCID: 0000-0001-6274-2016
nurid@hacettepe.edu.tr

Abstract

Countries participate in common tests to compare their positions with other countries and monitor their progress. While conducting these tests, comparing the achievement or attitudes between these countries or individuals without considering the differences between them will result in biased results. The anchoring vignette method was used in PISA 2012 as an alternative to eliminate this bias. In this study, the responses of Turkish, Bulgarian, and Greek students' to the survey questions in PISA 2012 about their mathematics teachers' class management and adjusted response categories based on their responses to anchoring vignettes were analyzed, and whether countries differ in class management or not was investigated. The data collected from 4,895 people were analyzed (1,665 from Bulgaria, 1,662 from Greece, and 1,568 from Turkey). All responses were not used; only three anchoring vignettes in the CLSMAN questionnaire and the 3rd item of the original questionnaire were used. The analyzes were performed in the R program with the ancors package. As a result of the research, it was found in both calculations that Greek teachers are quite flexible in starting class on time, and there is a significant difference between the response categories of other countries. According to the self-assessment responses, there is no difference between Turkish and Bulgarian teachers' class management; however, there is a significant difference between them according to the adjusted response. Different methods can be used in the process of evaluating anchoring vignettes using a non-parametric approach. Weighting is different in these methods. A single method can be used regarding the feature to be investigated or the researcher's algorithm preference; or as in this study, all the four methods can be used, and the comparisons can be interpreted.

Keywords: Mathematics education, Mathematics teacher, Class Management, Anchoring vignettes

E-International Journal of Educational Research

Vol: 3, No: 1, pp. 283-299

Research Article

Received: 2021-11-29
Accepted: 2022-02-24

Suggested Citation

Yurtçu, M. & Doğan, N. (2022). Analysis of mathematics teachers' class management behaviors In Turkey, Bulgaria and Greece using anchoring vignettes, *E-International Journal of Educational Research, 13*(1), 253-299. DOI: https://doi.org/10.19160/e-ijer.1029751

INTRODUCTION

Large-scale tests help countries see their position nationally and internationally (Avvisati, et al., 2019; Hirve et al., 2013; Van Soest et al., 2011). Factors that will affect the achievement of individuals can be examined through these tests, which aim to highlight affective characteristics and cognitive features. Generally, the scales and questionnaires having standard response categories are used for this purpose. Self-reporting questionnaires and scales are the assessment tools used to estimate the participant's personality, attitudes, values, beliefs. (Hirve et al., 2013; Hu, 2018; Primi et al., 2016; van Wilgenburg, 2010; Weiss and Roberts, 2018). However, individuals may respond to the questionnaire or scales consisted of standard categories at different threshold levels because many affective characteristics, such as feelings and perceptions, differ; this is usually ignored (Avvisati, et al., 2019; Bradbury-Jones, et al., 2014; Cope et al., 2016; Gottemoller, 2011; He, et al., 2017a; He et al., 2017b; Hinz, 2017; Hirve et al., 2013; Hu, 2018; Hu, et al., 2019; Van Soest, 2011; Visser et al., 2005). Even individuals with the same ability or achievement level may evaluate the items differently depending on their past and give different responses at different times. Therefore, making comparisons between variables such as achievement and attitude without considering the differences between participants/countries may cause biased interpretations (Azzolina et al., 2017; Cope et al., 2016; Grol-Prokopczyk, et al., 2011; Hinz, 2017; Hinz et al., 2020; Murray & Chen, 1992; Salomon et al., 2004; Van Soest, 2011), and the correlation between the variables of interest may give a different result than expected (Hinz et al., 2016; Hinz et al., 2020; Mottus et al., 2012). Consequently, different researchers cannot obtain consistent results for the variables (Hu et al., 2019), and the problem of validity arises in cross-country studies (Avvisati et al., 2019; Cope et al., 2016; Hu, 2018; Mottus et al., 2012; Primi et al.2016; Rice et al., 2008). There are various approaches that help to overcome this problem. One of them is Differential Item Functioning (DIF). It eliminates the differences between the responses of individuals with the same ability using their responses to anchoring vignettes (Hu, 2018; King et al., 2004; Raju, 1990; van Soest et al., 2011; Wand et al., 2011; Weiss & Roberts, 2018; Zumbo, 2007). The use of the levels expressing different cases as anchoring vignettes is an attempt to minimize the differences between groups, and DIF is used as a solution by rescaling the responses (Chevalier & Fielding, 2011; Gottemoller, 2011; Grol-Prokopczyk et al., 2011; King & Wand, 2007; Kapteyn et al., 2011; Weiss & Roberts, 2018; Xu & Xie, 2016). The aim is to verify the universality of the theories and psychological foundations in different cultures through intercultural studies (Solano, 2014; 30). In this respect, PISA and TIMSS are the leading tests. There are questionnaires for both students and stakeholders in these tests, revealing their cognitive and affective characteristics. Taking the differences between cultures under control is important in evaluating countries' real positions compared with other countries. For the first time, a new arrangement was introduced in PISA 2012 questionnaires to ensure cross-cultural comparability objectively (OECD, 2014). In this format, short stories called *anchoring vignettes*, in which hypothetical individuals/situations are defined were given to the participants, and they were asked to evaluate these individuals/situations. The objective is to reveal the differences between participants or countries after the process (He et al., 2017b; Hinz et al., 2020; Hirve et al., 2013; Hu et al., 2019; Korfage et al., 2007; Primi et al., 2016; Van Soest et al., 2011; Vonkova et al., 2018; Vonkova & Hrabak, 2015; Weiss & Roberts, 2018). The responses of participants/groups to the questionnaire or scales are called self-assessments; the responses of the original question are then rescaled according to the responses given to the anchoring vignettes. Thus, it is aimed to verify the self-assessment responses by considering the differences between the groups (Cope et al., 2016; Kristensen & Johansson, 2006; Primi et al., 2016; Rice et al., 2008; Salomon et al., 2004; van Wilgenburg, 2010; Vonkova & Hrabak, 2015; Weiss & Roberts, 2018).

Anchoring vignettes

Anchoring vignettes is a method used to determine the difference between groups, where short stories of hypothetical individuals, consisting of one or more vignettes representing different levels of a situation, are given to the participants (Chevalier & Fielding, 2011; Gottemoller, 2011; Grol-Prokopczyk et al., 2011; He et al., 2017a; Hopkins & King, 2010; Kapteyn et al., 2011; Kristensen & Johansson, 2006; van Wilgenburg, 2010). The objective is to identify and revise the differences in the cut-off scores of the response categories through these anchoring vignettes (He et al., 2017a). In this way, the cases are standardized as much as possible considering the norms and expectations of the participants (Hu et al., 2019; Salomon et al., 2004).

Anchoring vignettes are based on two assumptions. One of them is response consistency. This assumption means that the threshold used for the self-assessment is also used for the anchoring vignettes (Hirve et al., 2013; 2016; Gottemoller, 2011; Kristensen & Johansson, 2006; King et al., 2004; Salomon et al., 2004; Weiss & Roberts, 2018; van Soest et al., 2011). In violation of this assumption, participants give different responses to self-assessment items (Weiss & Roberts, 2018). The other assumption is the vignette equivalence. According to this assumption, all levels of the variable represented in the vignettes should be understood in the same way by all participants (Gottemoller, 2011; Grol-Prokopczyk et al., 2011; Hirve et al., 2013; Kristensen & Johansson, 2006; King et al., 2004). In this way, response categories are standardized (Gottemoller, 2011; Grol-Prokopczyk et al., 2011; He et al., 2017a; Hinz, 2017; Kapteyn, et al., 2011; Korfage, et al. 2007; van Wilgenburg, 2010; Xu & Xie, 2016).

PISA 2012 results were recalculated using 12 index anchoring vignettes. These indexes are also included in the data set with the ANC prefix. The names of these indexes and their variable codes are as follows: *ANCATSCHL*, *ANCATTLNACT*, *ANCBELONG*, *ANCCLSMAN*, *ANCCOGACT*, *ANCINSTMOT*, *ANCINTMAT*, *ANCMATWKETH*, *ANCMTSUP*, *ANCSCMAT*, *ANCSTUDREL* ve *ANCSUBNORM* (OECD, 2014).

The responses of two variables' anchoring vignette sets are available in the data. These variables are "Class Management/CLSMAN" and "Teacher Support/MTSUP." Regular and systematic class management contributes to the improvement of each student, and teachers play the most important role in the effective management of the class (Toprakcı, 2012; Weinstein et al., 2004).

Students in PISA 2012 data evaluated their mathematics teachers. The difference between the Class Management response categories by country is evaluated through the responses given to these questions. However, the differences arising in such an assessment might be due to the variable itself or the countries' perceptions of that variable; this is ignored in the analysis. Anchoring vignettes are necessary to reflect countries' perceptions to the responses of questionnaire items.

In this study, the assessment of the mathematics teachers about Class Management by the students of three countries that are geographically close to each other (Turkey, Bulgaria, and Greece), which participated in the 2012 PISA test, are considered. The responses of the third question of the CLSMAN questionnaire, *Teachers' starting class on time*, and the responses to the anchoring vignettes prepared for this questionnaire were analyzed by country, and new categories have been obtained. The response categories of the survey questions and the response categories rescaled according to the anchoring vignettes were compared. In this framework, the research question was set as: "Is there a real difference between the Class Management levels of Turkish, Bulgarian, and Greek mathematics teachers in PISA 2012 data?"

METHOD

The study that uses anchoring vignettes is based on two basic assumptions. To not violate the equality of vignettes assumption, the sample was selected as Turkey and two neighbor countries, which could be regarded as having similar perceptions, included in the B form of PISA 2012. The data belonging to Turkey, Greece, an OECD country, and Bulgaria, which participated in PISA 2012, were used. Before starting the analysis, missing values and participants who did not participate in this part of the survey were excluded from the data set. The data collected from 4,895 people were analyzed (1,665 from Bulgaria, 1,662 from Greece, and 1,568 from Turkey). All responses were not used; only three anchoring vignettes in the CLSMAN questionnaire and the 3rd item of the original questionnaire were used.

PISA 2012 includes the CLSMAN questionnaire consisting of four items (code ST85Q) and hypothetical anchoring vignettes (code ST84Q) to evaluate mathematics teachers' class management. Among the four items in the CLSMAN questionnaire, the self-assessment is mostly based on item ST85Q03, which has four response categories and better represents hypothetical questions. Items

ST84Q01-ST84Q03 were used as hypothetical vignettes. The information about the self-assessment question and anchoring vignettes are given in Table 1.

Table 1. *Questions for self-assessment and anchoring vignettes*

Question type/level	Content of the question	Question Code
Low Level (Q3)	Students in Mr./Mrs <name>'s class often interrupt their class. As a result, he/she usually comes to class five minutes late.	ST84Q03
Medium Level (Q1)	Students in Mr./Mrs <name>'s class often interrupt their class. As a result, he/she usually comes to class five minutes early.	ST84Q01
High Level (Q2)	Students in Mr./Mrs. <name>'s class are calm and organized. He/she always comes to class on time.	ST84Q02
Self-assessment question	My teacher starts class on time.	ST85Q03

The order of the anchoring vignettes at different levels can be determined by the researcher or by analyzing the options (Wand et al., 2011). In the study, the order of the vignettes related to Class Management was used as Q2 (High- ST84Q02)>Q1 (Medium- ST84Q01)>Q3 (Low-ST84Q03), as in PISA 2012 format. Each item takes a value between 1-4, representing "completely agree," "agree," "disagree," and "completely disagree," respectively. The order of the students' responses to anchoring vignettes is expected to be based on the responses consistent with this order, which is the Class Management level. For the cases that violate this order, making adjustments, rescaling the responses according to the countries' perceptions, and examining the response categories allow more objective comparisons between countries.

The analysis used in evaluating anchoring vignettes is based on two approaches: parametric and non-parametric approaches (Wand et al., 2011). In this study, the assumptions of the parametric approach were not met; therefore, the non-parametric approach, which does not require the confirmation of a statistical assumption (Hu et al, 2019), was used in the evaluation of anchoring vignettes. In this approach, participants ' self-assessment responses are recalculated by taking anchoring vignettes that are given at different levels. Self-assessment responses are placed in a category according to the number of anchoring vignettes (total number of categories = 2n+1) and take a new value. This new value is called the adjusted response category or rescaled eigenvalue responses and is denoted by the letter C.

Since there are three anchoring vignettes/hypothetical cases, eigenvalue responses can take 2* 3 + 1 = 7 different values. Accordingly, category values (C) obtained from rescaling vary between 1 and 7. The orders obtained using 3 anchoring vignettes are given in Table 2 (Van Soest & Vonkova, 2014; Wand, et al., 2011; Weiss & Roberts, 2018).

Table 2. *Possible oreders and corresponding categories (C)*

Possible order	C
s<v1<v2<v3	1
s=v1<v2<v3	2
v1<s<v2<v3	3
v1<s=v2<v3	4
v1<v2<s<v3	5
v1<v2<s=v3	6
v1<v2<v3<s	7

NOTE: Self-assessment (s) and anchoring vignettes (v1, v2, v3) symbols represent the following.
s = self-assessment response, v1= low = ST84Q03, v2= medium= ST84Q01, v3= high = ST84Q02

According to the responses given to the hypothetically offered anchoring vignettes, the self-assessment response of a participant takes a value, as shown in Table 2. However, sometimes participants may give responses contradictory to these orders. For example, if the participant prefers a high category for the hypothetical case representing a teacher with low-level Class Management and the same participant chooses a low category for the hypothetical case representing a teacher with high-level Class Management, this order is violated. In non-parametric approaches, the adjusted category value (C) can be obtained according to 4 different methods depending on the processing of violations. These methods are: omitting interval values (Omit), uniform allocation within intervals (Uniform), censored ordered probit model (Cpolr), and Mint (minimum entropy) (King & Wand, 2007; Wand et al, 2011). Omit ignores the values calculated as intervals when recalculating categories using the responses of anchoring vignettes. In this respect, it constitutes a reference for the other methods to give ideas about the weighting made in intermediate values. In Uniform, intermediate values recalculated according to the responses given to the anchoring vignettes are equally distributed among the intermediate values that the variable can take. With this aspect, the uniform method can be interpreted as equal distribution among the categories falling within the same range and taking the same threshold levels in different categories. Clopr is built on the ordered probit model. In this model, the intervals are assigned to the categories according to the thresholds. In Minimum Entropy, the smallest category of the recalculated values is enlarged by adding the probabilities of the other values in this interval to this category (Wand et al., 2011). These methods allow recalculation of the categories according to different weights and provide similar results in the absence of intermediate values.

Ancors package in the R program was used to obtain the categories rescaled according to anchoring vignettes using the non-parametric approach (Wand et al, 2016). Students' self-assessment of Teachers' starting class on time (Teacher's class management) and the response categories adjusted according to their responses to the items in the anchoring vignettes were examined together. The Kruskal Wallis test, which is suitable for categorical variables, was used to test the difference between mathematics teachers' classroom management by country. Mann-Whitney U test was performed for the groups with a significant difference.

FINDINGS

The frequency and percentages of the responses to the "Teachers' starting class on time" self-assessment question in the Class Management questionnaire are given in Table 3 for three countries.

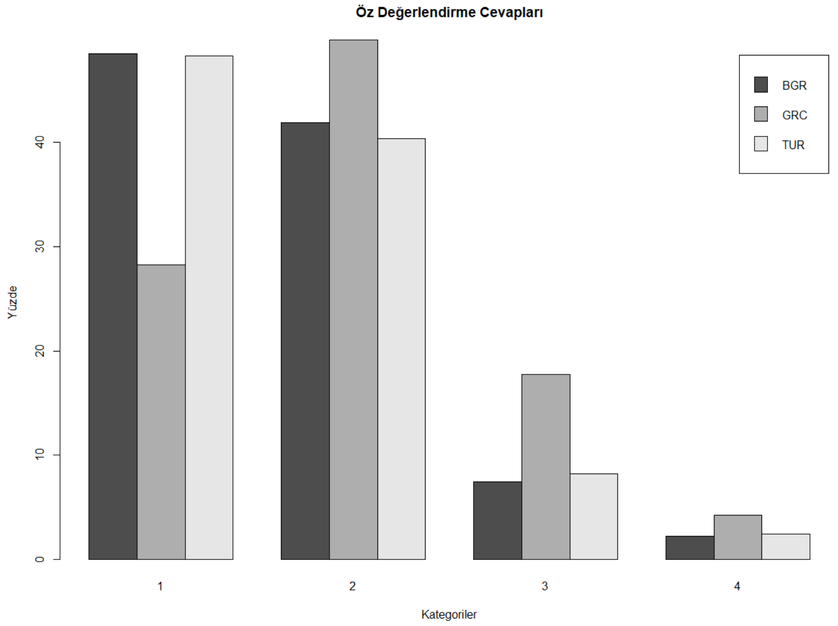
Table 3. *Descriptive statistics of three countries and the number of participants in each category*

	Category Level	Bulgaria	Greece	Turkey	Overall
Self-assessment responses (For question 3)	**1 (Completely agree)**	805	469	760	2,034
	~ %	48	28	48	41.6
	2 (Agree)	696	825	640	2,161
	~ %	42	50	41	44.14
	3 (Disagree)	126	296	129	551
	~ %	8	18	8	11.22
	4 (Completely disagree)	38	72	39	149
	~ %	2	4	3	3.04
Total		1,665	1,662	1,568	4,895
Descriptive statistics for self-assessments	**Mean**	1.66	2.00	1.65	1.77
	Standard deviation	0.83	0.88	0.76	0.84
	Median	2	2	2	2

According to Table 3, the responses given by Bulgarian and Turkish students are similar. "Teachers' starting class on time," which reflects Class Management, is evaluated the same by the students of both countries. Approximately half of the students in both countries stated that they completely agree with this statement (48%). The most preferred category in overall is "agree" (44.14%). When the percentages of the first two categories that support "Teachers' starting class on time" are merged, it is understood that 90% of teachers in Bulgaria, 89% in Turkey, and 78% in Greece come to class on time.

Descriptive statistics are used to evaluate the response given to the relevant item by country. These statistics provide information about the distribution within groups. The overall mean is 1.77, and the median is 2. In this case, it can be said that the mean is clustered in the "agree" category, which is the second one. Regarding the self-assessment of the three countries separately, Greece has the highest mean, whereas Bulgaria and Turkey have a similar mean. This shows that Class Management behavior in Greece is more flexible in mathematics teachers' starting class on time than in other countries. The standard deviations show that Turkish students' responses are closer to each other than in other countries.

The categories of self-assessment responses are given in Figure 1.



* BGR: Bulgaria; GRC: Greece; TUR: Turkey

Figure1. Responses to categories by country

Figure 1 shows that there is a difference between self-assessment responses of the countries. Regarding the percentages, "Teachers' starting class on time" behaviors are quite similar in Turkey and Bulgaria; teachers in Bulgaria are slightly more sensitive about starting the class on time. Although the overall is clustered in the 2nd category, Bulgaria and Turkey have been placed to the 1st category by more students. The normality of the data was tested to reveal whether there is a difference between self-assessment responses by countries. The results of the normality test are given in Table 4.

Table 4. *Normality test of self-assessment responses by country*

	Kolmogorov-Smirnov	
Country	**Statistics**	**p**
Bulgaria	.269	.000
Greece	.278	.000
Turkey	.289	.000

None of the groups showed normal distribution. Therefore, the Kruskal Wallis test, which compares the differences between groups according to medians, was used since the 3rd item used for Class Management has a 4-level structure. Kruskal Wallis-H Test results are given in Table 5.

Table 5. *Kruskal Wallis-H test results for the difference between countries' self-assessment responses*

Chi-Square Value	**df**	**p**
228.65	2	.00

Kruskal-Wallis-H chi-square value showed that the difference between the countries' responses was statistically significant (x^2 = 228.65; p <0.05). Paired comparisons were made using the Mann-Whitney U test to determine the source of this difference. The results of the test are given in Table 6.

Table 6. *Comparison of differences between countries*

Compared Countries	Z	p
BGR – GRC	-13.226560	.000
BGR - TUR	-0.1310416	.896
GRC - TUR	12.8958219	.000

Regarding the differences, there is no difference between the assessments of Bulgaria and Turkey; however, the assessment of Greece differs significantly from both Turkey and Bulgaria. Therefore, it was concluded that teachers' starting class time is a bit late for mathematics teachers in Greece, and they are poorer in Class Management than other countries. Accordingly, when the class management of the countries is evaluated according to their responses to the 3rd item, the order of the countries by best class management is Bulgaria, Turkey, and Greece.

In PISA 2012, three anchoring vignettes that help to standardize the assessors' perceptions were given to the students, and they were asked to evaluate each vignette in addition to self-assessment. Anchoring vignettes have the same category levels as self-assessment categories. The responses are given in Table 7 according to the countries.

Table 7. *Responses to anchoring vignettes by country*

	Category Level	Bulgaria	Greece	Turkey	Overall
Q2	1	864 (%51.9)	891 (%53.6)	985 (%62.8)	2,740 (%56)
	2	660 (%39.6)	625 (%37.6)	467 (%29.8)	1,752 (%35.8)
	3	116 (%7)	113 (%6.8)	95 (%6.1)	324 (%6.6)
	4	25 (%1.5)	33 (%2)	21 (%1.3)	79 (%1.6)
Q1	1	271 (%16.3)	124 (% 7.5)	227 (%14.5)	622 (%12.7)
	2	572 (%34.3)	392 (%23.6)	408 (%26)	1,372 (%28)
	3	666 (%40)	865 (%52)	600 (%38.3)	2,131 (%43.6)
	4	156 (%9.4)	281 (%16.9)	333 (%21.2)	770 (%15.7)
Q3	1	183 (%11)	109 (%6.6)	126 (%8)	418 (%8.6)
	2	334 (%20.1)	206 (%12.4)	244 (%15.6)	784 (%16)
	3	671 (%40.3)	749 (%45)	549 (%35)	1,969 (%40.2)
	4	477 (%28.6)	598 (%36)	649 (%41.4)	1,724 (%35.2)
N		1,665	1,662	1,568	4,895

Information about how mathematics teachers' classroom management is perceived or students' expectations in different countries can be obtained by analyzing the responses given to anchoring vignettes. The vignettes are ranked as Q2> Q1> Q3 for each country, which is an indication that countries perceived anchoring vignettes in the same order, and their priorities and expectations were the same. At the same time, when categories 1 and 2 are merged, which shows the teacher's starting the class on time, the agreement with vignettes are as follows: Q2 (91.5% for Bulgaria; 91.2% for Greece; 92.6% for Turkey), Q1 (50.6% for Bulgaria; 31.0% for Greece; 40.5% for Turkey) and Q3 (31.0% for Bulgaria; 19.0% for Greece; 23.6% for Turkey).

The analysis of the responses given to anchoring vignettes to see the perception of the countries showed that countries mostly prefer the "completely agree" category in anchoring vignettes 2, "teacher's starting the class on time." Differences are observed with the responses they gave in the self-assessment. Regarding self-assessment responses, Bulgaria has the highest agreement percentage of "Teachers' starting class on time," however, according to anchoring vignettes 2, the highest agreement percentage is observed in Turkey (62%). In the self-assessment results, Greece is the country that chooses the "completely agree" category at the lowest rate; on the other hand, Bulgaria is the county with the lowest agreement to the highest hypothetical vignette. This fact can be interpreted that the expectations for Bulgaria are higher.

Q3, the lowest-level hypothetical vignettes, evaluates the teachers' coming to the class 5 minutes late. Responses to anchoring vignettes 3 support the results of anchoring vignettes 2. The country with the highest percentage of participants who say "completely disagree" to this article is Turkey, and Bulgaria is the country with the lowest percentage. Q1 is the anchoring vignette reflecting the thoughts about teachers' coming to class 5 minutes early. The response categories of this medium vignette are 2 and 3. The category with the highest percentage is "disagree" (43.6%). The disagreement percentage of countries to this anchoring vignette is taken as the sum of categories 3 and 4, and the highest disagreement is observed among Greek students (69%). Therefore, less teacher comes to class early in Greece compared to other countries.

Regarding their responses in hypothetical vignettes, countries have a different order in "teachers' starting class on time" when the perceptions are standardized. Therefore, the responses in self-assessment were rescaled with a non-parametric approach, and adjusted response categories were obtained.

Adjusted categories consisting of 7 levels were obtained using the responses given to the three anchoring vignettes and self-assessment questions. Histograms of the adjusted self-assessments are given in Figure 2.

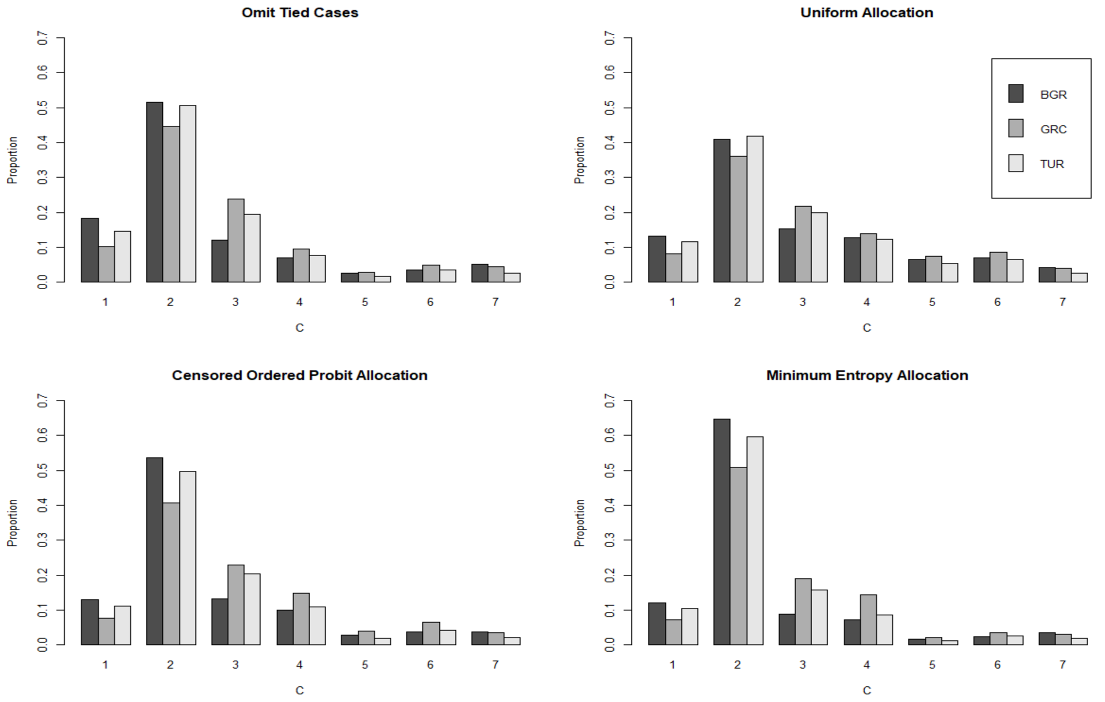


Figure 2. *Distribution of adjusted responses obtained after rescaling*

Adjusted response categories were obtained through four different methods. The percentages of these seven categories for each method are given in Table 8.

Table 8. *Agreement percentages of countries in the adjusted response categories*

Category	**Bulgaria**				**Greece**				**Turkey**			
	Omit	**Unifor**	**Cpol**	**Mint**	**Omit**	**Unifor**	**Cpolr**	**Mint**	**Omit**	**Unif**	**Cpol**	**Mint**
1	304	221	215	200	170	136	128	121	227	182	174	165
~ %	0.183	0.133	0.129	0.120	0.102	0.082	0.077	0.073	0.145	0.116	0.111	0.105
2	857	283	891	1076	741	602	676	844	792	657	779	935
~ %	0.515	0.41	0.535	0.646	0.446	0.362	0.407	0.508	0.505	0.419	0.497	0.596
3	201	255	219	147	395	361	382	316	306	312	318	246
~ %	0.121	0.153	0.131	0.088	0.238	0.218	0.229	0.190	0.195	0.199	0.203	0.157
4	115	211	167	120	155	229	246	239	121	193	169	135
~ %	0.069	0.127	0.10	0.072	0.094	0.138	0.148	0.143	0.077	0.123	0.108	0.086
5	43	110	48	28	48	123	65	33	25	83	30	17
~ %	0.026	0.066	0.029	0.017	0.029	0.074	0.039	0.020	0.016	0.053	0.019	0.011
6	58	115	63	38	80	143	108	57	56	102	66	41
~ %	0.035	0.069	0.038	0.023	0.048	0.086	0.065	0.034	0.036	0.065	0.042	0.026
7	87	10	62	57	73	68	57	52	41	39	31	28
~ %	0.052	0.042	0.037	0.034	0.044	0.041	0.034	0.031	0.026	0.025	0.020	0.018
Mean	2.55	2.80	2.63	2.42	2.82	3.18	2.99	2.76	2.53	2.87	2.63	2.45
Std Dv.	1.54	1.45	1.48	1.29	1.47	1.59	1.47	1.33	1.31	1.48	1.29	1.17
Medium	2	2	2	2	2	3	3	2	2	2	2	2

As shown in Figure 2 and Table 7, the adjusted responses to the questions that mathematics teachers are evaluated on Class Management were grouped under seven categories. These categories were analyzed using four different methods; the countries are clustered in the first three categories. The C value is clustered in categories 1 and 2 for Bulgaria and Turkey; it is clustered in categories 2 and 3 for Greece. Therefore, it can be said that Greece data in Class Management is at a lower level compared to the other two countries.

The category with the highest percentage is the 2nd category for all countries in all four methods

Regarding the descriptive statistics of the countries' adjusted assessments from different methods, the following results were revealed: the mean of Greece is higher than other countries in all four methods; the mean of Turkey is higher than Bulgaria in Uniform and Mint methods; Turkey and Bulgaria have similar means according to the Cpolr method. Adjusted Greece results are similar to self-assessment, but Bulgaria and Turkey's results are different from self-assessment results. Regarding the medians, the data of Bulgaria and Turkey are clustered in the category with a C value of 2. On the other hand, Greece data are clustered in the category with a C value of 3 according to Uniform and Cpolr methods and in the category with a C value of 2 according to other methods. Since the categories of anchoring vignettes, which have the same categories as the self-assessment, are scored in reverse order, lower categories represent higher agreement. Therefore, in a study with three anchoring vignettes, a category with a C value of 2 corresponds to category 6; and a category with a C value of 3 corresponds to category 5. According to the classification of Van Soest and Vonkova (2014) and Wand, King and Lau (2011) given in Table 2, the order of the case that corresponds to the C value of 6 is v1 <v2 <s = v3; and the order corresponding to the C value of 5 is v1 <v2 <s <v3. Using the codes of the vignettes, these rankings can be expressed as Q3 <Q1 <y = Q2 for C = 6 and Q3 <Q1 <y <Q2 for C = 5. Hence, according to adjusted assessments, Bulgarian and Turkish students' perceptions about their teachers' class management behavior are similar to the teacher's behavior in the 2nd vignettes (Q2).

Regarding the perception of Greek students about their mathematics teachers' Class Management, it is similar to the teacher's behavior in Q2 according to the Omit and Mint methods; it is lower than the behavior of the Q2 teacher and higher than the behavior of the Q1 teacher according to the Uniform and Cpolr methods.

First of all, normality assumptions were tested to compare the clusters of the adjusted self-assessment responses.

Table 9. *Normality test for adjusted self-assessment responses*

	Omit		Uniform		Cpolr		Mint	
Country	**Statistics**	**p**	**Statistics**	**p**	**Statistics**	**p**	**Statistics**	**p**
Bulgaristan	.338	.000	.274	.000	.333	.000	.394	.000
Greece	.259	.000	.216	.000	.235	.000	.296	.000
Turkey	.305	.000	.257	.000	.295	.000	.352	.000

The distribution of the countries is not normal in all four different methods. Kruskal Wallis-H test was conducted to test the difference between self-assessment responses and adjusted responses of the countries, and the results are given in Table 10.

Table 10. *Results of the Kruskal Wallis-H Test conducted to determine the difference between countries' self-assessment response categories*

Method	**Chi-Square Value**	**df**	**p**
Omit	78.556	2	.00
Uniform	79.606	2	.00
Cpolr	105.49	2	.00
MinEnt	119.9	2	.00

According to Table 10, the Kruskal-Wallis-H chi-square value is statistically significant in all four different methods. This shows that the difference between the orders in adjusted response categories is significant. The source of the difference was analyzed by dual comparison, and the results are given in Table 11.

Table 11. *Comparison of the difference between countries*

	Omit		Uniform		Cplor		MintEnt	
Country	**Z**	**p**	**Z**	**p**	**Z**	**p**	**Z**	**p**
BGR - GRC	-8.54960	.000	-8.77223	.000	-9.73426	.000	-10.62442	.000
BGR - TUR	-2.17429	.029	-2.89991	.003	-2.92293	.005	-1.93930	.003
GRC - TUR	6.24714	.000	5.74110	.000	7.648800	.000	7.542319	.000

Regarding the comparisons of adjusted response categories, the responses of all three countries differ from each other. Therefore, the country-wise comparison of the self-assessment responses and adjusted responses shows different results. According to the self-assessment responses, the only differences are between Greece-Bulgaria and Greece-Turkey. In contrast, there is a difference between the responses of all three countries according to the adjusted self-assessment response categories. Regarding the order of the countries' adjusted response categories, Greece's response categories get closer to the other countries than the self-assessment. The order difference between Bulgaria and Turkey becomes significant. The gap between Greece and other countries is closing according to the adjusted response categories; however, the gap between Turkey and Bulgaria is widened.

DISCUSSION AND SUGGESTIONS

The process of measuring the feelings, thoughts, and attitudes of the individuals is very important. A miscalculation in these processes will result in misjudgment and all related steps to proceed incorrectly. All participants should be examined under the same conditions to reveal these characteristics properly. Since it is very difficult to draw their boundaries for a human being, a social entity, one should be careful while expressing these features in measurement tools such as scale or questionnaire. As a result of individual perceptions or differences, various variables may affect the objective of the measurement, and a bias, consequently a validity problem may arise. Anchoring vignettes can be used in large-scale tests as a method to solve this problem. With these vignettes, response categories form a standard perception (Gottemoller, 2011; He et al., 2017b; Hinz, 2017; Korfage et al.2007; van Wilgenburg, 2010) and offer a solution to bias (Chevalier & Fielding, 2011; Gottemoller, 2011; Grol-Prokopczyk et al., 2011; Kapteyn et al., 2011; King & Wand, 2007; Xu & Xie, 2016; Weiss & Roberts, 2018). However, attention should be paid to the assumptions of anchor vignettes in providing an effective solution to this problem. The satisfaction of these assumptions is possible with the similarity of perceptions. Therefore, the data of Bulgaria, Greece, and Turkey, which are geographically close to each other, were used in the study. The self-assessment responses of these three countries to the 3rd item of the CLSMAN scale and the responses given to the anchoring vignettes in the PISA 2012 Class Management questionnaire were used. According to the information obtained from the self-assessment responses, the country that pays the most attention to mathematics teachers' starting class on time is Bulgaria. Turkey follows this behavior with a slight difference, and unlike these countries, the teachers in Greece show flexible behaviors about starting class in time. According to the self-assessment answers, the only difference is observed between Greece and other countries; no significant difference is found between Turkey and Bulgaria.

Regarding the countries' responses to the anchoring vignettes, the highest agreement percentage to the Q2 statement, representing the highest Class Management behavior, is from Turkey. This fact can be interpreted as that the perceptions of Turkish students fit with what is directly explained in the posed question. However, it can be concluded that Bulgarian and Greek students did not answer what was asked, and there are higher expectations for mathematics teachers in these countries. It can be interpreted as those Greek and Bulgarian students considered higher criteria in answering the questions and consequently answered in lower categories.

The answers given to anchoring vignettes were processed with self-assessment answers and rescaled using the non-parametric approach. The countries' response categories were re-analyzed using four different methods, and the differences between countries were compared. According to the adjusted response categories, Bulgarian and Turkish mathematics teachers' behaviors mostly overlapped with the Q2 expression ("Students in Mr./Mrs. <name>'s class are calm and organized. He/she always comes to class on time"), which represents the highest level. The behavior of Greek teachers in Class Management is between the expression representing the highest level (Q2) and the expression representing the medium level (Q1). After setting the standards, Greece's adjusted self-assessment responses approached other countries' responses as Greece expects high standards. However, this analysis has not changed the conclusion that teachers in Greece are flexible in starting class on time. Although there is no difference between Turkish and Bulgarian teachers' classroom management according to their self-assessment responses, a significant difference is found between them according to the analysis of adjusted response categories.

Hence, perceptions and expectations are very important when making comparisons in international tests such as PISA, in which countries participate to see their position. The scores obtained only from Likert-type scales can lead to misinterpretations. It is possible to obtain more accurate and valid results with anchoring vignettes, which are used to eliminate this mistake. This result is supported by many studies in the literature, in which response categories differed by using anchoring vignettes (Mottus et al., 2017; Korfage et al., 2007; van Wilgenburg, 2010; Kristensen & Johansson, 2006; He et al., 2017b). When comparing groups/countries, the answers given to the questionnaire or scale items should not be used directly without checking the differences that may arise due to different factors such as history or geography. Attention should be paid to the translations/adaptations of common tests used for the same purpose in different countries. Different methods can be used in the process of evaluating anchoring vignettes using a non-parametric approach. Weighting is different in these methods. A single method can be used regarding the feature to be investigated or the researcher's algorithm preference; or as in this study, all the four methods can be used, and the comparisons can be interpreted.

REFERENCES

Avvisati, F., Le Donne, Noemie & Paccagnella, M. (2019). A meeting report: cross-cultural comparability of questionnaire measures in large-scale international surveys. *Avvisati et al. Measurement Instruments for the Social Sciences, 1*:8. Doi: 10.1186/s42409-019-0010-z

Azzolina, D., Minto, C., Boschetto, S., Martinato, M., Bauce, B. Illiceto, S. & Gregori, D. (2017). Anchoring vignettes in EQ-5D-5L questionnaire: validation of a new instrument. *The open Nursing Journal, 11,* 144-156. Doi: 10.2174/1874434601711010144

Bradbury-Jones, C., Taylor, J. & Herber, O.R. (2014). Vignette development and administration: a framework for protecting research participants. *International Journal of Social Reseach Methodology, 17;* 427-440. Doi: 10.1080/13645579.2012.750833

Chevalier, A. & Fielding, A. (2011). An introduction to anchoring vignettes. *Journal of the Royal Statistical Society. Series A (Statistics in Society),* 174(3), 569-574

Cope, A.B., Ramirez, C., Devellis, R. F., Agans, R., Schoenbach, V.J. & Adimora, A. A. (2016). Measuring concurrency attitudes: development and validation of a vignette-based scale. *Plos ONE, 11(10).* Doi: 10.1371/journal.pone.0163947

He, J., Buchholz, J. & Klieme, E. (2017a). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. Journal of Cross-Cultural Psychology, Vol. 48(3) 319– 334. Doı: 10.1177/0022022116687395

He, J., Van de Vijver, F. J. R., Fetvadjiev, V. H., Dominguez-Espinosa, A., Adams, B. G., Alonso-Arbiol, I., Aydinli-Karakulak, A., Buzea, C., Dimitrova, R., Fortin Morales, A., Hapunda, G., Ma, S., Sargautyte, R., Schachner, R. K., Sim, S., Suryani, A., Zeinoun, P., & Zhang, R. (2017b). On enhancing the cross-cultural comparability of Likert-scale personality and value measures: A comparison of common procedures. *European Journal of Personality*.

Hinz, A., Hauser, W., Glaesmer, H. & Brahler, E. (2016). The relationship between perceived oun health state and health assesment ofanchoring vignettes. *International Journal of Clinical and Health Psychology, 16,* 128-136.

Hinz, A. (2017). Using anchoring vignettes in evaluation of beast canser survivors' quality of life. Breast Care, 12: 34-38. Doi: 10.1159/000455002.

Hinz,A. Karoff, J., Kittel, J., Brähler, E., Zenger, M., Schmalbach, B., & Kocalevent, R-D. (2020). Associations between self-rated health and the assessments of anchoring vignettes in cardiovascular patients. International Journal of Clinical and Health Psychology, 20, 100---107. Doi: 10.1016/j.ijchp.2020.04.001

Hirve, S., Gomez-Olive, X., Oti, S., Debpuur, C., Juvekar, S., Tollman, S., Blomstedt, Y., Wall, S. & Ng, N. (2013). Use of anchoring vignettes to evaluate health reporting behavior amongst adults aged 50 years and above in Africa and Asia-testing assumptions. Global Health Action, 6(1), 21064, Doi: 10.3402/gha.v6i0.21064

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly, 74*, 201-222. Doi:10.1093/poq/nfq011.

Hu, M., Lee, S. & Xu, H. (2019). Using anchoring vignettes to correct for differential response scale usage in 3mc surveys. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, First Edition. Edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke A.L. Stoop, and Brita Dorer, John Wiley & Sons, Inc.

Hu, M. (2018). *Anchoring vignettes for health comparisons: the validity of a multidimensional ırt model approach and design ımprovements using visual vignettes.* University of Michigan,

Gottemoller, P. G. (2011). *White Americans' Affect Toward African Americans: Predictive Power on Political Behavior and Measurement Problems*. Doctor of Philosophy, Southern Illinois University, Political Science, Illinois

Grol-Prokopczyk, H., Freese, J. &Hauser, R. M. (2011). Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior. 52(2) 246– 261.* Doi: 10.1177/0022146510396713

Kapteyn, A., Smith, J. P., van Soest, A. & Vonkova, H. (2011). Anchoring vignettes and response consistency. RAND. Labor and Population

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and crosscultural comparability of measurement in survey research. *American Political Science, 98*, 191-207.

King, G. & Wand, J. (2007) Comparing incomparable survey responses: evaluating and selecting anchoring vignettes. *Polit. Anal., 15, 66*

Korfage, I.J., de Koning, H. J. & Essink-Bot, M-L. (2007). Response shifting due to diagnosis and primary treatment of localized prostate cancer: a then test and a vignette study. *Qual Life Res, 16:* 1627-1634. Doi: 10.1007/s11136-007-9265-6.

Kristensen, N. & Johansson, E. (2006). New evidence on cross-country differences in job satisfaction using anchoring vignettes. ISBN 87-7882-095-2 (online). Working Paper 06-1. Retrieved from: https://core.ac.uk/download/pdf/7107252.pdf

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., Amoussou-Yéyé, D., Bäckström, M., Barkauskiene, R., Barry, O., Bhowon, U., Björklund, F., Bochaver, A., Bochaver, K., de Bruin, G., Cabrera, H. F., Chen, S. X., Church, A. T., Cissé, D. D., ... Johnson, W. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin, 38*, 1423-1436. doi:10.1177/0146167212451275

Murray, C. J. L. & Chen, L. C. (1992). Understanding morbidity change. *Popul Dev Rev,18:*481- 503.

OECD (Organisation for Economic Cooperation and Development) (2014). *PISA 2012 technical report.* [Çevrim-içi: http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf, Erişim tarihi: 1 Eylül 2015]

Primi, R., Zanon, C., Santos, D., De Fruyt F., & John, O. P. (2016). Anchoring vignettes, *Europan Journal of Psychological Assesment. 32*(1), 30-51. Doi: 10.1027/1015-5759/a000336

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement,* 14(2), 197– 207.

Rice, N., Robone, S. & Smith, P. C. (2008). *International compason of public sector performance: the use of anchoring vignettes to adjust self-reported data.* HEDG Working Paper 08/28

Toprakcı, E. (2012) Rethinking Classroom Management: A new perspective, a new horizon. e-*International Journal of Educational Research*, 3(3), 84-110.

Salomon, J. A, Tandon, A. & Murray, C.J. (2004). Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. *BMJ, 328*, 258

Solano, A. C. (2014). *Latin-American studies on well-being.* In A. Castro Solano (Ed.), *Cross-cultural advancements in positive psychology: Vol. 10. Positive psychology in Latin America* (p. 21–36). Springer Science + Business Media. Doi: 10.1007/978-94-017-9035-2_2

Van Soest, A., Delaney, L, Harmon, C., Kapteyn, A. & Smith, J. P. (2011). Validating the use of anchoring vignettes for corection of response scale differences in subjective questions. *J R Stat Soc Ser A Stat Soc. 174(3),* 575-595. Doi:10.1111/j.1467-985X.2011.00694.x.

Visser, M. R., Oort, F. J., & Sprangers, M. A. (2005). Methods to detect respose shift in quality of life data: A convergent validity study. *Quality of Life Research, 14(3*), 629-639.

van Soest A, Vonkova H (2014) How sensitive are retirement decisions to financial incentives? A stated preference analysis. *Journal of Applied Econometrics 29*(2):246–264

van Wilgenburg, K. (2010). *The Validity Of Anchoring Vignettes, Testng Response Consistency With An Experiment.* Erasmus University Rotterdam, Institute of health policy and management. Erasmus Unievrsity Roterdam Instiute of Health Policy & Management (Master Thesis).

Vonkova, H., Zamarro, G., & Hitt, C. (2018). Cross-country hetero-geneity in students' reporting behavior: The use of the anchoringvignette method. *Journal of Educational Measurement*, *55*,3-31. Doi: 10.1111/jedm.12161

Vonkova, H. & Hrabak, J. (2015). The (in) comparability of ICT knowledge and skill self-assessments among upper secondary school students: The use of the anchoring vignette method. *Computers & Education 85,* 191-202. Doi: 10.1016/j.compedu.2015.03.003

Wand, J., King, G., & Lau, O. (2011). Anchors: Software for anchoring vignette data. *Journal of Statistical Software. 42*(3).

Wand, J., King, G. & Lau, O. (2016). Package 'anchors': Statistical analysis of surveys with anchoring vignettes. Retrieved from: https://cran.r-project.org/web/packages/anchors/anchors.pdf

Weinstein, C.S., Tomlinson-Clarke, S., & Curran, M. (2004). Toward a conception of culturally responsive classroom management. *Journal of Teacher Education, 55*(1), 25-38. doi: 10.1177/0022487103259812

Weis, S. & Roberts, R. D. (2018). Using anchoring vignettes to adjust self-reported personality: a comparison between countries. *Frontiers in Psychology,* 9, 325. Doi:: 10.3389/fpsyg.2018.00325

Xu, H. & Xie, Y. (2016). Assessing the effectiveness of anchoring vignettes in bias reduction for socioeconomic disparities in self-rated health among Chinese adults. *Sociological Methodology, 46*(1) 84–120, Doi:: 10.1177/0081175015599808

Zumbo, B. D. (2007) Three Generations of DIF analyses: considering where ıt has been, where ıt ıs now, and where ıt ıs going, *Language Assessment Quarterly, 4*(2), 223- 233.

Türkiye, Bulgaristan ve Yunanistan Matematik Öğretmenlerinin Sınıf Yönetimi Davranışlarının Ortak Senaryolarla Incelenmesi

Dr. Öğr. Üyesi Meltem Yurtçu
İnönü Üniversitesi - Türkiye
ORCID: 0000-0003-3303-5093
meltem.yurtcu@inonu.edu.tr

Prof.Dr. Nuri Doğan
Hacettepe Üniversitesi - Türkiye
ORCID: 0000-0001-6274-2016
nurid@hacettepe.edu.tr

Özet

Ülkeler kendi konumlarını diğer ülkeler ile kıyaslamak ve kendi ilerlemelerini gözlemlemek üzere ortak sınavlara katılım sağlamaktadır. Bu sınavların yapılması sürecinde ise bireyler/ülkeler arası farklılıkların göz önüne alınmadan bu ülke veya bireyler arasında başarı veya tutumların karşılaştırılması yanlı sonuçlar elde edilmesine sebebiyet verecektir. Bu yanlılığı ortadan kaldırmaya yönelik bir alternatif olarak PISA 2012 de ortak senaryolar yöntemi kullanılmıştır. Bu araştırma kapsamında 2012 PISA uygulamasına katılmış coğrafi olarak birbirine yakın üç ülkenin (Türkiye, Bulgaristan ve Yunanistan (Bulgaristan'dan 1.665, Yunanistan'dan 1.662 ve Türkiye'den 1.568).) öğrencilerinin matematik öğretmenlerini sınıf yönetimi konusunda değerlendirmeleri dikkate alınmıştır. CLSMAN anketinin 3. Sorusu olan öğretmenin zamanında sınıfta olma sorusuna verilen cevaplar ile bu anket için hazırlanmış ortak senaryolara verilen cevaplar ülkelere göre incelenmiş ve yeni kategoriler elde edilmiştir. Analizler R programında ancors paketi ile gerçekleştirilmiştir. Araştırma sonucunda Yunanistan'daki öğretmenlerin zamanlama konusunda her iki hesaplamada da derse katılımlarında esnek davrandığı ve diğer ülkelerle cevap kategorileri arasında anlamlı düzeyde farklılık olduğu sonucu ortaya çıkmıştır. Öz değerlendirme yanıtlarına göre Türk ve Bulgar öğretmenlerin sınıf yönetimleri arasındaki fark ortaya çıkmaz iken analiz sonucunda düzeltilmiş cevap kategorilerine göre aralarında anlamlı düzeyde fark olduğu bulunmuştur. Parametrik olmayan bir yaklaşım kullanılarak ortak senaryoların değerlendirme sürecinde farklı yöntemler farklı ağırlıklandırmaları dikkate alarak kullanılabilir. İncelenecek özelliğe veya araştırmacının algoritma tercihine göre tek bir yöntem kullanılabilir; veya bu çalışmada olduğu gibi, dört yöntemin tümü kullanılabilir ve karşılaştırmalar yorumlanabilir.

Anahtar Kelimeler: *Matematik eğitimi, Matematik öğretmeni, Sınıf Yönetimi, Ortak senaryolar*

E-Uluslararası Eğitim Araştırmaları Dergisi

Cilt: 3, No: 1, ss. 283-299

Araştırma Makalesi

Gönderim: 2021-11-29
Kabul: 2022-02-24

Önerilen Atıf

Yurtçu, M. ve Doğan, N. (2022). Türkiye, Bulgaristan ve Yunanistan matematik öğretmenlerinin sınıf yönetimi davranışlarının ortak senaryolarla incelenmesi, *E-International Journal of Educational Research, 13*(1), 253-299. DOI: https://doi.org/10.19160/e-ijer.1029751

GENİŞLETİLMİŞ ÖZET

Problem: Öz bildirim sağlayan anketler ve ölçekler, bir katılımcının kişiliği, tutumları, değerleri, inançları vb. hakkında önemli bilgileri tahmin etmek için kullanılan bir değerlendirme olarak ele alınmaktadır (Hirve, vd, 2013; Hu, 2018; Korfage, de Koning ve Essink-Bot, 2007; Primi, vd. 2016; van Wilgenburg, 2010; Weiss ve Roberts, 2018). Ancak araştırmaya katılan bireylerin duyguları, algıları gibi birçok duyuşsal özellikler farklılık gösterdiği için standart kategorilerin kullanıldığı anket veya ölçeklere bireylerin farklı eşik seviyelerinde tepkiler verecekleri göz ardı edilmektedir (Avvisati, vd., 2019; Bradbury-Jones, vd., 2014; Cope vd., 2016; Gottemoller, 2011; He, vd., 2017a; He vd., 2017b; Hinz, 2017; Hirve vd., 2013; Hu, 2018; Hu, vd., 2019; Van Soest, 2011; Visser vd., 2005). Hatta aynı yetenek veya başarı düzeyine sahip bireyler farklı geçmiş yaşantılarına bağlı olarak farklı kategorilerde değerlendirme yapabilmekte, farklı zamanlarda farklı cevaplar verebilmektedir. Dolayısı ile bireyler/ülkeler arası farklılıkların göz önüne alınmadan başarıları, tutumlar gibi değişkenler arasında karşılaştırmalar yapmak yorumlamaların yanlı olmasına (Azzolina vd, 2017; Cope vd., 2016; Grol-Prokopczyk, vd., 2011; Hinz, 2017; Hinz vd., 2020; Murray ve Chen, 1992; Salomon vd, 2004; Van Soest, 2011) ve ilgilenilen değişkenler arasında hesaplanan korelasyonun beklenenden farklı bir sonuç vermesine (Hinz vd, 2016; Hinz vd., 2020; Mottus vd., 2012) neden olmaktadır. Bu durumun sonucunda ise farklı araştırmacılar ilgilenilen değişkenler için tutarlı sonuçlar elde edememekte (Hu vd., 2019) ve ülkeler arası çalışmalarda ilgilenilen değişken için geçerlik sorunu ortaya çıkmaktadır (Avvisati, vd., 2019; Cope vd., 2016, Hu, 2018; Mottus vd., 2012; Primi vd., 2016; Rice vd., 2008). Bu geçerlik probleminin üstesinden gelmek üzere ortak senaryolara cevap veren aynı yeteneğe sahip bireylerin tepkilerindeki farklılıkları ortadan kaldırmaya yardımcı yaklaşımlardan bir tanesi de Değişen madde fonksiyonudur (DMF-Differenatial Item Functioning-DIF-) (Hu, 2018; King vd., 2004; Raju, 1990; van Soest vd., 2011; Wand vd., 2011; Weiss ve Roberts, 2018; Zumbo, 2007). Farklı durumları ifade eden düzeylerin ortak senaryolar olarak kullanılması, gruplar arası farklılıkları en aza indirgemeye çalışmakta ve tekrardan ölçekleme ile DMF bir çözüm yolu olarak kullanılmaktadır (Chevalier ve Fielding, 2011; Gottemoller, 2011; Grol-Prokopczyk vd., 2011; King ve Wand, 2007; Kapteyn vd., 2011; Weiss ve Roberts, 2018; Xu ve Xie, 2016). Özellikle kültürler arası çalışmalar ile farklı kültürlerdeki teori ve psikolojik temellerin evrenselliğini doğrulamak amaçlanmaktadır (Solano, 2014). Yapılan çalışmaların başında ise PISA, TIMSS gibi uygulamalar gelmektedir. Bu uygulamalarda bilişsel ve duyuşsal özellikleri ortaya çıkarmak adına hem öğrencilere hem de paydaşlara yönelik anketler yer almaktadır. Bu uygulamalarda kültürler arası farkların kontrol altına alınması ülkelerin diğer ülkelere göre gerçek konumlarını karşılaştırmada önemlidir. PISA 2012 çalışmasında ilk defa kültürler arası karşılaştırılabilirliği nesnel olarak sağlamak adına anketlere yeni bir düzenleme getirilmiştir (OECD, 2014). Bu format ile ortak senaryolar (anchoring vignette) olarak adlandırılan kısa hikayeler bireylere yönlendirilmektedir. Senaryolarda hipotetik bireyler/durumlar tanımlanarak katılımcıların bu bireyleri/durumları değerlendirmeleri istenmektedir. İşlem sonrasında ise katılımcılar veya ülkeler arası farklılıkların ortaya çıkarılması amaçlanmaktadır (He vd., 2017b; Hinz vd., 2020; Hirve vd., 2013; Hu vd., 2019; Korfage vd., 2007; Primi vd., 2016; Van Soest vd., 2011; Vonkova vd., 2018; Vonkova ve Hrabak, 2015; Weiss ve Roberts, 2018). Bireylerin/grupların anket veya ölçeklere verdikleri cevaplar öz değerlendirmeler olarak adlandırılmakta ve bu cevaplar ortak senaryolara verilen cevaplara göre tekrardan ölçeklenmektedir. Böylece gruplar arasındaki farklılıklar dikkate alınarak öz değerlendirme cevaplarını doğrulamak amaçlanmaktadır (Cope vd., 2016; Kristensen ve Johansson, 2006; Primi vd., 2016; Rice vd., 2008; Salomon vd., 2004; van Wilgenburg, 2010; Vonkova ve Hrabak, 2015; Weiss ve Roberts, 2018).

Ortak senaryolar bir duruma ait farklı düzeyi temsil edecek hipotetik bireylerin kısa hikayesi olarak bir veya daha fazla senaryodan oluşacak şekilde katılımcılara yöneltilen gruplar arası farklılığı belirlemek için kullanılan bir yöntemdir (Chevalier ve Fielding, 2011; Gottemoller, 2011; Grol-Prokopczyk vd., 2011; He vd., 2017a; Hopkins ve King, 2010; Kapteyn vd., 2011; Kristensen ve Johansson, 2006; van Wilgenburg, 2010).

PISA 2012 de 12 tane indeks ortak senaryoların kullanılması ile tekrardan hesaplanmıştır. Bu indekslerden "Sınıf Yönetimi-Class Managment/CLSMAN" ve "Öğretmen desteğidir-Teacher Support/MTSUP" için senaryolardaki bireylere ait tepkiler veri setinde yer almaktadır. Ülkelere göre cevap kategorileri arasında farklılığın olup olmadığı bu verilerde anket sorularına verilen cevaplar üzerinden değerlendirilmektedir. Ancak böyle bir değerlendirmede ortaya çıkan farklılıkların gerçekte incelenen özellikten mi yoksa ülkelerin o konuya yönelik algılarından mı kaynaklandığı göz ardı edilmektedir.

Ülkelerin algılarının anket maddelerine verdiği cevaplara yansıtmak için ortak senaryo maddelerinden yararlanmak gerekmektedir.

Bu araştırma kapsamında 2012 PISA uygulamasına katılmış coğrafi olarak birbirine yakın üç ülkenin (Türkiye, Bulgaristan ve Yunanistan) öğrencilerinin matematik öğretmenlerini sınıf yönetimi konusunda değerlendirmeleri dikkate alınmıştır. CLSMAN anketinin 3. Sorusu olan öğretmenin zamanında sınıfta olma sorusuna verilen cevaplar ile bu anket için hazırlanmış ortak senaryolara verilen cevaplar ülkelere göre incelenmiş ve yeni kategoriler elde edilmiştir. Ülkelerin hem anket sorularına verdiği cevap kategorileri hem de ortak senaryo sorularına göre yeniden ölçeklenmiş anket sorularına ilişkin cevap kategorileri karşılaştırılmıştır. Bu çerçevede araştırma sorusu;

"PISA 2012 de Türk, Bulgar ve Yunan matematik öğretmenlerinin sınıf yönetimi düzeyleri arasında geçekte bir farklılık var mıdır?" şeklinde kurulmuştur.

Yöntem: Ortak senaryoların kullanıldığı bu çalışmada temel iki varsayım göz önüne alınmıştır. Çalışmada senaryoların eşitliği varsayımını ihlal etmemek için örneklem olarak PISA 2012 da B formunda yer alan ve yaklaşık olarak aynı coğrafyada benzer algıya sahip olarak ele alınabilecek Türkiye ile birlikte komşu iki ülke verisi kullanılmıştır. Çalışma grubu olarak Türkiye ile birlikte OECD ülkesi olan Yunanistan ve PISA 2012'de katılımcı olarak yer alan Bulgaristan'a ait veriler ele alınmıştır. Araştırmada analize başlamadan önce kayıp değerler ve bu anket uygulamasına katılmayan bireyler veri setinden çıkarılmıştır. Çalışmaya Bulgaristan için 1665, Yunanistan için 1662 ve Türkiye için 1568 kişiye ait veriler dahil edilmiştir. Çalışma grubu olarak ülkelerin tüm cevapları değil CLSMAN anketi için verilmiş olan üç farklı düzeydeki ortak senaryoya ilişkin cevaplar ve bu anketin 3. maddeye ait cevaplar kullanılmıştır.

PISA 2012'de matematik öğretmenlerinin sınıf yönetimini değerlendirmek üzere ST85Q kodlu dört maddeden oluşmuş CLSMAN anketi ve ST84Q kodlu hipotetik ortak senaryolar yer almaktadır. CLSMAN anketinde yer alan bu dört maddeden hipotetik soruları daha çok temsil eden 4 cevap kategorili ST85Q03 maddesi temel alınmıştır. Bu soruya verilen cevaplar öz değerlendirme yanıtı olarak değerlendirilmiştir. Araştırmada PISA 2012 formatındaki gibi senaryoların sınıf yönetimi düzeyinin sırası olarak Q2 (Yüksek-ST84Q02) > Q1 (Orta- ST84Q01) > Q3 (Düşük-ST84Q03) kullanılmıştır. Öğrencilerin ortak senaryolara vermiş oldukları olası cevapların sırasının sınıf yönetimi düzeyi olan düşük-orta-yüksek olarak verilen bu sıra ile tutarlı cevaplara dayanması beklenmektedir.

Bu çalışmada ortak senaryoların değerlendirilmesinde parametrik yaklaşıma ait varsayımların sağlanmamasından dolayı istatistiksel bir varsayımın teyidini gerektirmeyen (Hu, Lee ve Xu, 2019) parametrik olmayan yaklaşım kullanılmıştır.

Parametrik olmayan yaklaşımlarda ihlal durumlarını işleme almaya bağlı olarak düzeltilmiş kategori değeri (C değeri) 4 farklı yaklaşıma göre elde edilebilir. Bu yaklaşımlar; Omit, Uniform, Cpolr ve Mint kısaltmaları ile anılmaktadır. Kısaltmaların açılımları ise sırası ile aralık değerlerini göremezden gelme (omitting interval values), aralıkları kendi içinde eşit dağıtma (uniform allocation within intervals), sansürlenmiş sıralı probit model (censored ordered probit model) ve minimum entropi (minimum entropy) şeklindedir (King ve Wand, 2007; King, Wand ve Lau, 2011). Bu yaklaşımlar farklı ağırlıklandırmalara göre kategorilerin tekrardan hesaplanmasını sağlamakta ve ara değerlerin olmaması durumunda benzer sonuçlar vermektedir.

Araştırmada ortak senaryoları kullanılarak tekrardan ölçeklenen kategorileri parametrik olmayan yaklaşıma göre elde etmek için R programında ancors paketinden (Wand, King ve Lau, 2016) yararlanılmıştır. Bu çalışma ile öğrencilerin başlangıçta öğretmenlerin sınıf yönetimi için zamanında sınıfta olma maddesine ait öz değerlendirme ve ortak senaryolarda yer alan maddelere vermiş oldukları cevaplara göre düzeltilmiş cevap kategorileri birlikte incelenmiştir. Ülkelere göre matematik öğretmenlerinin sınıf yönetimleri arasında gerçekte farklılık olup olmadığını incelemek üzere kategorik değişkenler için uygun olan Kruskall Wallis testi kullanılmıştır. Kruskall Wallis testine göre aralarında manidar bir farklılık oluşan gruplar Mann Whitney U testi ile incelenmiştir.

Sonuçlar: Bireylere ilişkin duygu, düşünce ve tutum gibi yapıların ölçülmesi süreci oldukça önemlidir. Bu süreçlerin yanlış ölçülmesi yanlış değerlendirilmesine ve buna bağlı bütün adımların yanlış ilerlemesine sebep olacaktır. Bu yapıların doğru bir şekilde ortaya çıkarılması için bütün katılımcıların aynı şartlar altında incelenmesi gerekmektedir. Sosyal bir varlık olan insana ait bu özelliklerin sınırlarının çizilmesi oldukça güç olmakla birlikte bu özelliklerin ölçek veya anket gibi ölçme araçları ile ifade edilmesinde dikkatli davranılması gerekmektedir. Bireysel algılara veya farklılıklara bağlı olarak ölçülmek istenen amaca çeşitli değişkenlerin karışması ile yanlılık durumu ve dolayısı ile geçerlik sorunu ortaya çıkmaktadır. Büyük ölçekli sınavlarda ise bu soruna cevap olabilecek bir yöntem olarak ortak senaryolar kullanılabilmektedir. Bu senaryolar sayesinde cevap kategorileri standart bir algı oluşturmakta (Gottemoller, 2011; He vd., 2017b; Hinz, 2017; Korfage vd., 2007; van Wilgenburg, 2010) ve yanlılığa bir çözüm sunmaktadır (Chevalier ve Fielding, 2011; Gottemoller, 2011; Grol-Prokopczyk vd., 2011; Kapteyn vd., 2011; King ve Wand, 2007; Xu ve Xie, 2016; Weiss ve Roberts, 2018). Ancak ortak senaryoların etkili olarak bu duruma çözüm sunabilmesi için varsayımlarına dikkat edilmesi gerekmektedir. Bu varsayımların sağlanması ise algıların benzerliği ile mümkündür. Bu yüzden çalışmada coğrafi olarak birbirlerine yakın Bulgaristan, Yunanistan ve Türkiye verileri kullanılmıştır. Bu üç ülkenin CLSMAN ölçeğinin 3. Sorusuna verdikleri öz değerlendirme cevapları ile sınıf yönetimi anketi için PISA 2012 de yer alan ortak senaryolara verilmiş cevaplar ele alınmıştır. Öz değerlendirme cevaplarından elde edilen bilgilere göre matematik öğretmenlerin derse zamanında başlamaya en çok dikkat eden ülke Bulgaristan olarak bulunmuştur. Çok az bir fark ile bu davranışı Türkiye takip etmekte, Yunanistan'daki öğretmenlerin ise biraz daha bu ülkelerden farklı olarak derse başlama konusunda esnek davranışlar gösterdiği sonucu ortaya çıkmıştır. Öz değerlendirme cevaplarına göre sadece Yunanistan ile diğer ülkeler arasında farklılık var iken, Türkiye ve Bulgaristan verilerine ait öz değerlendirme cevap kategorileri arasında anlamlı düzeyde farklılık bulunmamıştır. İşlemi sürecinde ele alınan ortak senaryolara ülkelerin cevapları incelendiğinde en yüksek sınıf yönetimi davranışını temsil eden Q2 ifadesine katılım yüzdesinin Türkiye'den olduğu gözlemlenmiştir. Bu durum yöneltilen sorularda direkt anlatılmak istenilen ile Türk öğrencilerin algılarının daha çok uyuştuğu şeklinde yorumlanabilir. Ancak Bulgaristan ve Yunanistan'daki öğrenciler ise tam olarak sorulmak istenilene cevap veremedikleri sonucu ortaya çıkmakta ve bu ülkelerdeki matematik öğretmenlerine yönelik daha yüksek düzeyde beklentilerinin olduğu ifade edilebilir. Bir diğer deyişle Yunan ve Bulgar öğrencilerin sorularındaki kriterlerin daha yüksek olarak düşünüp ona göre düşük kategorilerde cevap verdikleri şeklinde yorumlanabilir.

Ortak senaryoların hepsine verilen cevaplar öz değerlendirme cevapları ile ele alınarak parametrik olmayan yaklaşımlara göre tekrardan ölçeklenmiştir. Dört farklı yaklaşıma göre tekrardan ülkelerin cevap kategorileri incelenmiş ve ülkeler arasındaki farklar karşılaştırılmıştır. Düzeltilmiş cevap kategorilerine göre Bulgaristan ve Türkiye'de matematik öğretmenlerin davranışlarının çoğunlukla ortak senaryolardan en yüksek düzeyi temsil eden Q2= "Bayan < isim>'in sınıfındaki öğrenciler sakin ve düzenlidir. O her zaman derse zamanında gelir." ifadesi ile örtüştüğü sonucu ortaya çıkmıştır. Yunanistan'daki öğretmenlerinin sınıf yönetimi konusundaki davranışlarının ise en yüksek düzeyi temsil eden ifade-Q2- ile orta düzeyi temsil eden ifade-Q1- arasında bir tutum sergiledikleri gözlemlenmektedir. Standartların belirlenmesi ile Yunanistan'ın daha yüksek standartlarda beklentisi olduğundan düzeltilmiş öz değerlendirme cevapları diğer ülkelerin cevaplarına yaklaşmıştır. Ancak bu yaklaşma yine de diğer ülkelere göre Yunanistan'daki öğretmenlerin zamanlama konusunda derse katılımlarında esnek davrandığı sonucunu değiştirmemiştir. Ayrıca öz değerlendirme yanıtlarına göre Türk ve Bulgar öğretmenlerin sınıf yönetimleri arasındaki fark ortaya çıkmaz iken analiz sonucunda düzeltilmiş cevap kategorilerine göre aralarında anlamlı düzeyde fark olduğu bulunmuştur.

Öneriler; Ülkelerin kendi konumunu görmek için katıldıkları PISA gibi uluslararası sınavlarda karşılaştırma yaparken algılar ve beklentiler oldukça önemlidir. Sadece likert türü ölçeklerden alınan puanlar yanlış yorumlamalara yol açabilmektedir. Özellikle bu yanlışlığı ortadan kaldırmak için kullanılan ortak senaryolar ile daha doğru ve geçerli sonuçlar elde etmek mümkündür. Ortak senaryoların uygulanması ile cevap kategorilerinin farklılaştığı birçok çalışma literatürde bu sonucu desteklemektedir (Mottus, vd. 2017; Korfage, vd. 2007; van Wilgenburg, 2010; Kristensen ve Johansson, 2006; He vd., 2017b).). Gruplar/ülkeler arası karşılaştırmalar yapılırken geçmiş yaşantılarından veya bulundukları coğrafi koşullar gibi farklı etmenlerden kaynaklı olarak yaşanabilecek farklılaşmalar kontrol altına alınmadan anket veya ölçek maddelerine verilen cevaplar direkt kullanılmamalı, farklı ülkelerde aynı amaç için kullanılan ortak sınavlarda çevirilere/uyarlamalara dikkat edilmelidir. Parametrik olmayan yaklaşımları kullanarak ortak senaryoların değerlendirilmesi sürecinde farklı yaklaşımlar kullanılabilmektedir. Yaklaşımlar farklı farklı ağırlıklandırmalara göre hesaplanmaktadır. İncelenen özelliğe veya araştırmacının uygulamak istediği algoritma tercih edilerek tek bir yaklaşım kullanılabileceği gibi bu araştırmada olduğu şekli ile dört yaklaşım birlikte ele alınarak karşılaştırmalar yorumlanabilir.