International Journal of Assessment Tools in Education
2019, Vol. 6, No. 1, 9–24

DOI: 10.21449/ijate.447780

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

An Effective Way to Provide Item Validity: Examining Student Response Processes

Omer Kutlu [1], **Hatice Cigdem Yavuz** [1,*]

[1] Department of Educational Measurement and Evaluation, Ankara University, Turkey

ARTICLE HISTORY

Received: 01 August 2018
Revised: 11 December 2018
Accepted: 19 December 2018

KEYWORDS

Cognitive interviews,
Item validity,
TIMSS,
Response processes

Abstract: Studies based on response processes of individuals can provide information that supports the assessment and increases the validity of the items in the scale or tests. The purpose of this study is to present the extent to which the student response processes are effective in identifying and developing the characteristics of the items in an achievement test and in collecting validity evidence. For this purpose, 28 Turkish fourth-grade students were chosen, half were high-achieving students and the remaining half were low-achieving students. The items for the study were chosen from the Trends in International Mathematics and Science Study TIMSS 2007 and 2011 by taking into consideration several item characteristics. Before cognitive interviews, an interview guide was also prepared. In the study, it was determined that cognitive interviews, especially those conducted with the high-achieving students, can serve to develop item validity. In the cognitive interviews with the low-achieving students, information was gathered concerning how students who did not have specific knowledge measured with an item were able to respond to that item.

1. INTRODUCTION

One of the most important characteristics sought in tests used in education and psychology is validity. The ways of increasing validity by obtaining evidence of this characteristic are among the important issues that concern psychometricians. Validity is defined as "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), 1999, p. 9). For this reason, according to the Standards in Education and Psychology, validity in this context does not refer to the actual test but to the validity of the interpretations and evaluations made in consideration

CONTACT: Hatice Cigdem Yavuz hcyavuz@ankara.edu.tr Department of Educational Measurement and Evaluation, Ankara University, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

of the intended uses of the test. In this context, psychometricians search for evidence from different sources in relation to the validity of test scores (McDonald, 1999).

There are different opinions on the definition and classification of the term validity (see Sireci, 2007). Some researchers conceptualise validity within a general framework (AERA, APA, & NCME, 2014; Kane, 2013; Sireci & Foulkner-Bond, 2014; Sireci, 2007) and some suggest that validity cannot be interpreted generally (e.g. Borsboom, Mellenbergh, & van Heerden, 2004; Lissitz, & Samuelsen, 2007). Since tests used in psychology and education are developed depending upon specific purposes, people or conditions, it is not possible to develop a perfect test, which would serve all the required characteristics (Cronbach, 1984). From this point of view, in 1999 and 2014, the Standards in Education and Psychology (AERA, APA, & NCME, 1999, 2014) approached validity as a whole in the form of the types of validity without separation according to the types, such as content validity, criterion-related validity, and structural validity. The current study is based on this approach.

According to the Standards in Education and Psychology (AERA, APA, & NCME, 2014)), the ways of collecting the evidence of validity are divided into five sources of validity evidence; test content, response processes, internal structure, relations to other variables, and testing consequences. With these sources of validity evidence, the ways to obtain validity based on response processes has become an attention-grabbing subject in the field literature (Desimone & Le Floch, 2004; Padilla & Benítez, 2014; Ryan, Gannon-Slater & Culbertson, 2012). Evidence based on response processes is defined as "concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA, APA, & NCME, 2014, p.12). In this way, thorough information is gathered regarding the cognitive processes shown by the respondents and response processes, and it is possible to determine to what extent these processes are in accord with the purposes of the test (Padilla & Benítez, 2014). In addition, together with the response processes it is possible to reveal how items are interpreted by individuals (DeWalt et al., 2007). Thus, studies based on response processes can provide information that supports the evaluation and increases the validity of the items in the scales or tests.

There are wide-ranging ways to obtain evidence based on the response processes, such as think aloud, focus groups and interviews (Padilla & Benítez, 2014). Among these, cognitive interview is a method composed of thinking aloud and verbal probing techniques (Willis, 2005). The role of the cognitive interview is important and useful in understanding the response processes of individuals (DeSimone & Le Floch, 2004; Ryan, Gannon-Slater, & Culbertson, 2012). With cognitive interviews, it is easier to discover which strategies individuals use and what they really think when responding to an item (Hopfenbeck & Maul, 2011). According to Desimone & Le Floch (2004), cognitive interviews can reveal mistakes in the item, different interpretations regarding the item, and the effect of the social desirability on the response to the item. Thus, measures to increase the validity of the items can be undertaken with the obtained data. Through cognitive interviews, it can be determined whether the items in measurement tools need to be reorganised (Conrad & Blair, 2004). In this context, cognitive interviews are used in pilot applications of scale development research and produce effective results (e. g. Johnstone, Figueroa, Attali, Stone, & Laitusis, 2013; Peterson, Peterson, & Powell, 2017; Snow & Katz, 2009; Wildy & Clarke, 2009).

In the literature, cognitive interviews are being investigated in various fields, such as health, education and social sciences. In their study, DeWalt et al. (2007) researched whether items in a scale related to a psychological structure were clear and understandable, and how individuals interpreted the items through the cognitive interview. With a similar purpose, Nicolaidis, Chienello and Gerrity (2011) showed through cognitive interviews that a 10-item scale is clear and understandable in terms of the focus group. In another study (Ding, Reay, Lee, & Bao,

2009), cognitive interviews were used in order to identify validity problems that could not be identified by experts and to present students' different perspectives towards the items. Ercikan, Arim and Law (2010) used the response processes of the students in order to examine the differential item functioning (DIF), which results from linguistic differences. With a similar purpose, Benitez and Padilla (2013) investigated DIF in student questionnaires used in the Program for International Student Assessment (PISA) and sought to reveal the possible sources of DIF by carrying out cognitive interviews with students. According to the findings of the study, words in some items were variously interpreted by students in different groups. Wildy and Clarke (2009) conducted cognitive interviews to test the preliminary test of the scale they used in their work and to test whether the meaning the scale writers attributed to an item was understood in the same way by the respondents. With a similar purpose, Ryan et al. (2012), identified measurement errors that had not emerged in other analysis methods using cognitive interviews to assess the validity of a scale. Ouimet, Bunnage, Carini, Kuh and Kennedy (2004) re-analysed the "College Student Report" tool developed for university students, within the framework of cognitive interview, focus group interview and expert opinions. The findings of the study show that these descriptive methods are important in improving the clarity, validity, appearance and reliability of the tool, as well as in revealing the strengths and weaknesses of the item.

Apart from the studies carried out on scales and questionnaires, cognitive interviews are also used on tests in the field literature. For example, Johnstone et al. (2013) used cognitive interviews to determine how students with disabilities interpreted the items in large-scale tests and to receive their feedback regarding the test items. In this framework, the differences were revealed between the responses of the disabled and non-disabled students in relation to the test items. In another study, cognitive interviews were carried out on the iSkills ™ test developed by the Educational Testing Service (ETS) to measure the digital literacy skills of students (Snow & Katz, 2009). According to the findings of the study, evidence of validity was obtained for iSkills ™ with regard to determining students' digital literacy. In their study, Noble, Rosebery, Suarez, Warren and O'Connor (2014) analysed the response processes of English Language Learner (ELL) students and non-ELL students through items gathered from a high-stakes science test. The findings of the study show that even though the ELL students have knowledge of the item, the linguistic features of the items led them to the wrong answer.

In the field of literature, it is seen that cognitive interviews are used on test items quite narrowly, and they are generally applied to improve the validity of the questionnaire or scales or to obtain evidence of validity. Thus, this study attempts to fill this gap in the field literature, and to show that cognitive interviews can also be applied to younger age groups and this application can be informative in terms of validity. Within this framework, the purpose of this study is to present the extent to which the student response processes are effective in identifying and developing the characteristics of the items in an achievement test and in collecting validity evidence. The research questions developed for this purpose are: What are students' response processes concerning (i) the necessity for the figure or table in the item root, (ii) the clarity of the text or figure given in the item root, (iii) the level of difficulty of the item, (iv) the level of knowledge given in the item, and (v) the reason for selecting the relevant choice in the item?

2. METHOD

2.1. Participants

The participants of this study were 24 Turkish fourth-grade elementary school students (10 girls, 14 boys) aged between 9 and 10. Cognitive interview studies are conducted with typically small sample sizes (Willis, 2015) thus the sample size in this study was enough to obtain well detailed student response processes. Moreover, the ability of items to measure the desired

feature without involving the group characteristics is very crucial. While gathering information about whether the items possess this feature, taking opinions from individuals with different characteristics will lead to more realistic and accurate information being obtained. For this reason, the schools to be involved in the research were carefully selected. Purposive sampling was employed in this study with the selection of two schools. After receiving permission from the authorities, one class was randomly selected from each school. In addition, participants were informed that participation was voluntary.

In this study, students were divided into two groups of high achieving and low achieving. One reason for this division is that students in these two different achievement levels would have different perspectives towards the items, which would frame their responses to the questions directed to them. The level of achievement of the students that the classroom teachers verbally indicated was taken into account. To determine the students' achievement level, the teachers were asked to consider students' academic achievement performance, participation in class, performing assignments and undertaking homework adequately, the level of interest and curiosity they have during the lesson, and briefly the students' performance in the classroom.

2.2. Data Collection Tools

The items for the study were chosen from the TIMSS 2007 and 2011 (TIMSS 2007 Assessment, 2009; TIMSS 2011 Assessment, 2013) and are presented in Appendix 1. In this study, students answered the Turkish version of these items which were translated and adapt in the framework of TIMSS asessments. Item statistics, cognitive domains and the scope of the items were taken into account in the selection of the items and attention was paid to the items being heterogeneous in terms of the related features. The TIMSS items were chosen on the basis of the item parameters being estimated according to item response theory and that the information regarding the items can be easily accessed. Table 1 gives the characteristics regarding the items and the number of participants that gave correct answers.

Table 1. The characteristics regarding the items

Item	Item type	Item discrimination	Item difficulty	Guessing parameter	Context	Cognitive Domain	The number of participants that gave correct answers	
							High-achieving	Low-achieving
1	MC*	.76	-1.64	.22	Biology	Knowing	12	10
2	MC*	1.12	-1.14	.26	Biology	Applying	11	7
3	MC*	.71	.14	.22	Chemistry	Knowing	9	8
4	MC*	.84	.39	.18	Chemistry	Reasoning	12	6
5	OE**	.53	1.07	-	Physics	Reasoning	12	3
6	MC*	.75	-1.63	.22	Biology	Applying	11	7
7	OE**	.95	1.13	-	Chemistry	Applying	8	3
8	OE**	1.00	.28	-	Biology	Knowing	6	1

* MC: Multiple choice, ** OE: Open ended

Table 1 shows that a total of six multiple-choice and three open-ended items were selected for use in the study. This distribution was chosen for the students to be able to respond to items during class time.

In this study, an interview form was another data collection tool. The interview form was composed of questions about the length, language, and the use of visual materials of the items. Techniques specific to the method of cognitive interviewing (Willis, 2015; Bowen, Bowen & Woolley, 2004) were employed in the preparation of the items in the interview form. After the interview form was prepared, the opinions of two experts in the field of educational measurement were taken. The experts were asked to express their views on the language and

expression characteristics of the questions, suitability for the age level, and the appropriateness of the extent of the questions. For pretesting, the interview form was applied to five students of the same age range from a class but were not participating in the study. The interview form was modified within the framework of the findings obtained from pretesting and the expert opinions.

2.3. Procedure

In this study, first, the items were applied to the students, and then cognitive interviews were administered. Prior to the applications, information was given to the classroom teachers and students. It was explained to the students that their participation in the study would be kept confidential and their performance in response to the science items would not be shared with third parties apart from the researchers. It was also stated that participating in the test and interviews was voluntary.

After the required explanations in class were given by at least one researcher, the selected items were applied to the students. It took students approximately 30-40 minutes to respond to the items. After the test was applied, to avoid creating a negative impression towards the selection of the students, it was specified that interviews were to be conducted with the randomly selected students, and they were randomly summoned one by one from their classes from a list. During interviews, same protocols were followed with each group, and students were not told about the correctness of their choices on items as part of the interview.

Cognitive interviews were conducted on one-on-one basis with students in a different classroom in the school. Cognitive interviews are a combination of many techniques and provide a comprehensive understanding of how individuals comprehend and respond to items (Tourangeau, Rips & Rasinski, 2000). In this method, it is sought to determine the processes and thoughts that individuals experienced while replying to the items (Willis, 2015). Moreover, as pointed out by Willis, inferences are made regarding why the respondents in the cognitive interviews responded to a related item and how they responded in that way. Bowen et al. (2006) state that cognitive interviews can be carried out in the following way: First, the respondent is asked to read the item, and afterwards it is determined what meaning the item has for him and what he is being asked in the item. Why the student chose that answer is asked according to the student's response. Probing questions to be addressed in cognitive interviews can also reveal what the student actually thinks about an item when responding to that item. In this study, these steps were followed in each interview.

2.4. Data Analysis

Before the analysis all interviews were transcribed, verbatim, yielding a total of 51 pages of written transcriptions. The data obtained from the interviews was analysed using descriptive analysis (Strauss and Corbin, 1990). In the context of this approach, first, the views of the students were placed in the predetermined themes and response categories. Determination of themes and response categories were established according to the answers of questions in the interview form. The placing the views into categories were conducted by the authors. This process was undertaken separately for each item. In the analysis of the data, the quotations that best explain the answers of the low and high achieving students were collated in accordance with each theme. In the study, the data belonging to four students selected impartially from the 24 students were also re-coded by an expert in educational measurement. With her re-coded data, the rater-reliability was determined as 96%.

3. FINDINGS

3.1. Responses to the necessity of a Figure/Table

The responses of students on the necessity of a figure/table to answer the items are given in order in Table 2.

Table 2. The responses of the students on the necessity of a figure/table

Item	2		4		5		6		7	
Student group	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A
Needed	6	10	10	8	7	6	6	4	11	7
Not needed	6	2	2	4	5	6	6	8	1	5

H-A: High achieving, L-A: Low achieving

According to Table 2, most of the students stated that there was no need for a figure or table to answer item 6; yet, it was needed for item 7. The students who thought that a figure was needed to answer item 2 expressed this as, "I could see the characteristics of the walrus by looking at the figure". An examination of the items shows that being able to respond to items 4 and 7depends on using the given figure. However, Table 2 reveals that there were also students who stated that a figure/table was not necessary to respond to these items. The reason why a group of students did not consider a table necessary for responding item 4 is that the experiment column is not given in the table. In relation to this topic, one student commented, "I was confused by the table because the experiments were not stated one by one." According to the students' responses, the absence of a separate column that showed four experiments was considered to be "puzzling".

The students found the figure given in item 5 necessary since it "makes it easier to answer the item". Students presented their opinions as, "You understand their distance better in the visual", and "In the figure, A pulls from a farther distance". Generally, the figure in item 6 was considered as complicated by both high- and low-achieving students with comments such as "It is not clear that it is a frog", "the figure is confusing", and "the picture could be less blurred". For item 7, the higher-achieving students expressed their thoughts as, "the figure is necessary, I would not be able to understand if it were not for the figure" and "the figure is necessary, I did not understand the question when they were given one after the other", and the low-achieving students stated that "Items were long for me" and "Items were confusing".

3.2. Responses on text/figure clarity

Table 3 presents the responses of students on the comprehensibility of the text/ figure in the items.

Table 3. The responses of the students on the comprehensibility of the text/ figure

Item	1		2		3		4		5		6		7		8	
Student group	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A
Clear	12	7	9	7	12	8	9	6	11	6	11	7	8	4	7	4
Not clear	-	5	3	5	-	4	3	6	1	6	1	5	4	8	5	8

H-A: High achieving, L-A: Low achieving

According to Table 3, the text in items 1, 3, 5 and 6 was considered as comprehensible by the high-achieving students. The choices in the item were the reason why items 1 and 6 were not understood by low-achieving students. For item 2, students stated that they did not understand the word "walrus". The responses of low achieving students to item 3 were: "The item is not completely expressed well" and "I could not understand the text very much, I did not understand the top part". For item 5, generally the comments of the low-achieving students regarding the text in the item were: "I did not understand the first sentence", "Puzzling", and "Text is too long". The low-achieving students stated the reason why they could not understand item 7 was as follows: "I could not understand the cups", "x, y cup names", and "I could not understand the text very much". For the same item, the high-achieving students gave these responses: "I

was confused by the expression" and "I did not understand how he/she placed the container without taking out the glass in the figure".

The text of item 8 was not found to be understandable by a large number of both the high- and low-achieving students. The low-achieving students expressed their views as, "I perceived it as we can see plants and seed from a distance", "I did not understand the word 'from far away'", and "The text is complicated, providing a figure would be better". The views of the high-achieving students were given as "puzzling", "I understood it when I read it twice", and "The words 'far away' made me confused".

Generally, as can be understood from these responses, when making decisions about data to determine whether items are clear and understandable, the group which responded correctly to the item most of the time should be taken into consideration. Taking the students' responses and Table 3 into account, for the less successful students, the comprehensibility of the text becomes more difficult as the length of the text increases.

3.3. Students' responses on item difficulty

Table 4 presents the responses of students about difficulty of the item.

Table 4. The responses of the students about difficulty of the item

Item	1		2		3		4		5		6		7		8	
Student group	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A
Easy	8	8	12	8	8	5	6	6	10	8	11	8	5	9	5	6
Medium	4	3	-	2	4	4	3	3	-	2	-	1	5	-	3	5
Most difficult	-	1	-	2	-	3	3	3	2	2	1	3	2	3	4	1

H-A: High achieving, L-A: Low achieving

According to Table 4, the high-achieving students found item 2 the easiest whereas items 7 and 8 were found to be the most difficult. For the low-achieving students, item 7 was the easiest and item 3 was the most difficult. The high-achieving students explained why they generally found item 1 easy or of medium difficulty as, "This information is in cartoons", " I learned this in the documentary", "Because I've read in books", "Among the choices, it made the most sense " and "I could not decide between B and C " whereas the low-achieving students stated, "I had a bird and therefore I know it", "I am good with animals", and "I know about birds". The high-achieving students gave the reasons for finding item 2 easy as, "I learned it from a documentary" and "I learned from documentaries and books". Those among the low-achieving students who found item 2 easy explained, "I learned it from the TV" and "Choices are nonsense", and the remaining students expressed their view as, "I have never heard of a walrus", "I do not know what palette is", and "Walrus is a different animal".

For item 3, the higher-achieving students who found the item easy expressed their view as, "We saw it in class" and "It melts instantly because it is hot", and those who found it to be of medium difficulty stated, "I was confused because first it was cold then warm and hot." Those of the low-achieving students who found this item easy commented, "I knew it, it dissolves in cold when the heat is high", and those who found it to be of medium difficulty or difficult stated, "warm and hot water puzzled me". Those who found item 5 easy generally expressed their view as, "I solved it with the help of other options" and "the other options are meaningless". Those who found the item to be of medium difficulty or difficult stated, "it is because of the table" and "the table confused me".

Both high- and low-achieving students generally found item 5 easy. The high-achieving students gave reasons, such as "the answer is written directly" and "the answer is definitely

clear", and the low-achieving students gave responses; for example, "the figure made it easier" and "the picture made it easier".

Item 6 was also generally found to be easy by the high- and low-achieving students. The higher-achieving students responded as, "there were things that could not be possible in other options" and "I saw it in the movies that the frogs leave" whereas the responses of the low-achieving students were "I love animals", "Options are easy", and “the answer is apparent”.

Both high- and low-achieving students generally found item 7 difficult or of medium difficulty giving responses, such as "I did not understand the complicated text much, it was a little long", "I do not understand how they put the x in the box", "I was confused because of the figures", and "I had difficulty because of the text". The high- and low-achieving students who found this item easy commented, "Figures made it easier", "I learned it in science lessons", and "the options made it easier".

Item 8 was also found to be difficult or moderately difficult by the majority of the high and low achieving students. The students expressed their views as: "The sentence is complicated, it is not clear", "it would be better if they added a figure", "I did not understand the seed in the text", "I did not understand the words ‘far away’ in the text". The students who considered the item to be easy gave responses such as: "I imagined, estimated in my mind", "I like the plant kingdom", "I see it in villages", "I see it in documentaries" and "I pictured it in the soil".

3.4. Responses regarding the level of knowledge given in the item

In Table 5, the responses on the level of knowledge given in the item are given according to responses of the high- and low-achieving students.

Table 5. The responses of the students on the level of knowledge given in the item

Item	1		2		3		4		5		6		7		8	
Student group	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A
Enough	12	9	11	9	11	7	12	8	11	8	11	8	10	8	6	3
Not enough	-	3	1	3	1	5	-	4	1	4	1	4	2	4	6	9

H-A: High achieving, L-A: Low achieving

According to Table 5, the high-achieving students indicated that items 1 to 7 contained sufficient information to find the answer. The low-achieving students generally expressed their views as bird features could be given in item 1, features about walrus could be given in item 2, figures could be given in item 3 and 7, a table was missing in item 4, and the item root was inadequate in item 6. To solve item 8, both the high- and low-achieving students stated that there was not enough information. The opinions of the high-achieving students were: "I think there is a need for some sentences and some figures", "a little information can be added", "there is no information, direction", "it could have been expressed in a different way", and “a figure can be added”, and the low-achieving students shared these views with their comments of "’how far away?", "I was puzzled by the words ‘far away’", "there could be a figure", "’ ‘far away’ is not something that can be expressed", and "it is a little hard; it could have been simplified."

3.4. Responses on the students’ selection of the relevant item option

The responses of the high- and low-achieving students on the reasons for selecting the relevant option are given in Table 6. According to this table, the reason for students’ selection of an option they considered to be the correct answer was generally related to the other options. Thus, the students found the correct answer by eliminating the other options that were given before questioning the information in the item. For item 3, most of the high-achieving students found

the right answer using their knowledge. In addition, when finding the right answer, both the high- and low- achieving students were assisted by the other choices with most assistance provided by item 6.

Table 6. The responses of the students on the reasons for selecting the relevant option

Item	1		2		3		4		6	
Student group	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A	H-A	L-A
Knowledge	6	5	3	3	10	6	8	1	2	1
Other options	6	7	9	7	2	6	4	8	9	7

H-A: High achieving, L-A: Low achieving

Generally, the high-achieving students who selected the correct answer by benefiting from the options in all items expressed their opinions as, “it was clear to me because of the other choices”, "I eliminated the other options", "the other options cannot be correct because they are unreasonable", "A was the most meaningful option", and “it was quite clear in this option” while the high-achieving students who selected the correct answer by based on their knowledge commented, "I learned it from books and documentaries," "it does not melt in cold water; I know it from tea", and "because the temperature and the mixture are always the same". The opinions of the low-achieving students were similar to these views. Among these students, those who marked the item based on the knowledge they possessed explained, "the layer of fat will protect the animal" and "heat melts sugar" while those who first eliminated the other options gave opinions, such as "the options give a lot of clues", "by eliminating the options", and "it makes more sense than the other options".

4. DISCUSSION and CONCLUSION

The aim of this study was to show that the responses of the fourth-grade students to some items chosen from the TIMSS application could be used as validity evidence and to increase item validity. In the study, opinions were obtained from the high- and low-achieving students, and it was determined that cognitive interviews, especially those conducted with the high-achieving students, can serve to develop item validity. In the cognitive interviews with the low-achieving students, information was gathered concerning how students who did not have specific knowledge measured with an item were able to respond to that item.

Generally, this study shows that students, like experts, can have a role in providing evidence for item validity and increase validity. The findings of this study were found to be similar to those of other studies in the field literature (Ercikan et al., 2010; Nicolaidis, Chienello & Gerrity, 2011; Noble et. al, 2014). Validity evidence obtained through this research show that students are important as consulting experts in the processes of test development and adaptation; thus, this is an effective method to find solutions to validity issues. Even though cognitive interviews are time-consuming and costly (Desimone & Le Floch, 2004), it appears that this type of studies would be useful in improving the validity of the items and the tests. Although referring to experts’ opinions is an important and widespread method, it is clear that student’s response processes also provide validity evidence since experts cannot possibly have any knowledge about the student's response processes (Benitez & Padilla, 2013).

In situations where students’ high-level cognitive skills such as problem-solving and critical thinking are to be measured, verbal, numerical materials or materials; e.g. tables and figures are often used in the root of the developed item; therefore, attention should be paid to the selection of these materials (Beddow, Elliot & Kettler, 2013). According to the findings regarding the necessity of the materials in the items selected for this study, the students stated that some materials are necessary for some items and unnecessary for others. Considering that materials

that do not contribute to the solution of an item may negatively affect the item validity (Linn & Grounland, 1995, Nitko & Brookhart, 2007), it is clear that the visuals of such items need to be modified. For example, the material used in item 6 in the current study did not contribute to the student(s)' response to that item. In this sense, it is suggested that visuals that are not essential for students to respond to an item should not be included as item materials.

There may be a different approach to the inclusion of material in items which aim to measure psychological structures to ensure that such items are developed without bias (Haladyna, 1997; Osterlind, 2002). Since international measurement applications such as TIMSS and PISA are applied in numerous different cultures, material can sometimes be used in the root of the item even if it is not necessary, and the opinions that students expressed for item 2 in the current study support this. Some students stated that they benefited from the picture since they had not seen a walrus before, and a picture was needed in order to solve the item. In this sense, considering the cultural and socioeconomic characteristics of the groups to which the items will be applied is also significant regarding the validity of the item. The interviews with the students revealed that the material in item 5 provided a clue for the low-achieving students to find the correct answer. However, the materials in the root of the item should not give the students a clue to the correct answer (Haladyna, 1997). It is important to recognise that reliability and validity in tests depend on the selection of the items (Linn, 1989); therefore, as an example, the material in item 5 should be changed or removed.

In tests that measure psychological characteristics, the comprehensibility of the items is important in terms of language and expression. In the processes of writing items, the language used in the items and generally in the whole test must be clearly written and comprehensible, obeying spelling and punctuation rules (Haladyna, 1996; Osterlind, 2002). According to the student opinions, the items selected for this study from TIMSS appeared to have problems in their Turkish language. In addition, in order for the material in the base of the item to be understandable, the class and age level to which the item will be applied should also be considered (Linn & Gronland, 1995). The current study revealed that in the cognitive interviews, item 7 was more complicated than it should be for fourth-grade students. The x and y letters used in item 7 were confusing for the students. Furthermore, the word 'walrus' in item 2 also prevented the students from comprehending the item.

In tests, such as TIMSS, which have been translated from the language in which they are originally developed into a different language, the process of translation is important in the sense that gives the meaning which the item desires to measure, and represents the characteristics in the item. From this point of view, regarding the findings related to the items in the current study, it can be seen that in particular, the meaning of item 8 was lost. Regardless of whether the students were lower or higher achieving, it is necessary to create a situation in which all students understand the items and only the ones who have the knowledge and skills related to the item can answer it correctly (Linn & Grounland, 1995). The findings obtained from the current study show that there are problems in this respect; thus, it is suggested that changes are made to the items based on the opinions of the students.

Significant findings about the difficulty level of the items were obtained in the study. According to the cognitive interviews, the findings indicated that the items based on remembering a related specific knowledge were easy for the students to solve; items such as comprehension, problem solving, and critical thinking were more difficult because they measure information in a more complex way. In addition, the students attributed the ease of solving some of the items to the distractors not being related to the correct answer. Furthermore, extra-curricular resources such as documentaries and books also contributed to the response of the items.

It is observed in the current study that students generally had difficulty in open-ended items. Similarly, in a study conducted by Johnstone et al. (2013), the students achieved more correct

answers in multiple-choice tests. Another reason for the Turkish students having difficulties with open-ended items is that they are not familiar with this type of items in their instructional programme. To resolve this problem, the students' responses can be consulted to determine the level of difficulty of the items to be included in the tests. In large-scale international tests, such as TIMSS and PISA, differences in socioeconomic and sociocultural characteristics of the participating countries can also affect the difficulty levels of the items. The reason for the distractors of some of the items not being connected to the correct option may be due to the differences between countries and the results of such examinations being a particular concern to the educational policy makers. In such measurement applications, the percentage of items to which there is a correct response is a particular concern for educational policy makers since it leads to various inferences being made about the countries. Hopfenbeck and Maul (2011) emphasise the need to take care over comparisons of countries with the information gained from these measures.

When the findings in the current study are analysed in relation to the sufficiency of the information required for the solution of the items, the high-achieving students generally found the given information sufficient with the exception of item 8. It is understood from the opinions of the low-achieving students that the information given in the item was not adequate. Taking the other responses given to item 8 into account, it was problematic in many respects. Therefore, it would be wrong and biased to compare Turkish students with other students who responded to this item because this item does not operate in the same way for the Turkish-speaking students. Mistakes originating from translation can be observed in tests such as TIMSS and PISA (see Goldstein, 2008). The diversity of participating countries taking these tests and that the tests do not follow the various curricula in the schools makes the item writing process difficult. In this framework, the tests should be based on common learning topics, and preliminary research on the test content should be undertaken by obtaining opinions from the participating countries. This process will be significant in developing the scientific accuracy of the items and the general validity of the test.

Considering the students' reasons for selecting an option for the items addressed to them in the study, it is seen that writing multiple-choice questions is as important as writing the item base. In this context, various precautions should be taken in writing the options in multiple-choice items (Haladyna, 1997, Nitko & Brookhart, 2007). Moreover, the response process of a multiple-choice item depends on the characteristics of the student as much as the characteristics of the - item itself. While some students read the item and look for the expression in the choices they think is the correct response, some students try to obtain the answer by comparing the choices with each other after reading the item (Pehlivan Tunç & Kutlu, 2014). In this sense, some students can develop test-wiseness behaviour when responding to the items. This situation decreases the validity of the test, and in order to avoid such strategies, measures should be taken in the development and review of the items (Towns, 2014). It was also determined in this study that distractors should be rational and consistent with the context especially in multiple-choice tests (Osterlind, 2002).

When the findings obtained from this study are considered as a whole, taking into account student responses in item writing and undertaking the corresponding improvement of the items will provide significant contributions to the validity of the tests. There are only a few studies on this subject; therefore, more studies being conducted will help gain different perspectives in test development processes. In future studies, it would be appropriate to use different item types, benefit from a larger number of items and refer to the opinions of students who are studying at different levels. In addition, researchers can conduct intercultural cognitive interview studies to support the development studies of scales and tests, such as TIMSS, which are applied in different cultures. This study has some limitations and premises. In this sense, the results of this

study need to be evaluated within this framework. A significant limitation is that the students were fourth graders, and they had not participated in this type of interviews before. For this reason, some data loss was experienced in some questions. Moreover, this study was conducted with only 24 students. Another limitation is that the study used a total of eight multiple-choice and three open-ended items. The assumptions that students responded to the questions without being affected by social desirability and that students responded to questions solely based on their own knowledge were accepted as the premises of the study.

Acknowledgement

This study was presented at the 10th International Test Commission (ITC) Conference in 1 - 4 July 2016 in Vancouver, Canada.

ORCID

Ömer KUTLU https://orcid.org/0000-0003-4364-5629
Hatice Çiğdem YAVUZ https://orcid.org/0000-0003-2585-3686

5. REFERENCES

AERA, APA, & NCME (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2013). Test accessibility: Item reviews and lessons learned from four state assessments. *Education Research International*, *2013*, 1-12. doi:10.1155/2013/952704

Benitez, I., & Padilla, J. L. (2013). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research, 8*(1) 52-6. doi: 10.1177/1558689813488245

Borsboom, D., Mellenbergh, G. J., & Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-71. doi: 10.1037/0033-295X.111.4.1061

Bowen, N. K., Bowen, G. L., & Woolley, M. E. (2004). Constructing and validating assessment tools for school-based practitioners: The Elementary School Success Profile. In A. R. Roberts & K. Y. Yeager (Eds.) *Evidence-based practice manual: Research and outcome measures in health and human services* (pp. 509-517). New York: Oxford University Press.

Conrad, F., & Blair, J. (2004). Aspects of data quality in cognitive interviews: The case of verbal reports. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, et al. (Eds.) *Questionnaire development, evaluation and testing methods* (pp. 67-88). New York: Wiley.

Cronbach, L. J. (1984). *Essentials of psychological testing.* NY: Harper.

Desimone, L., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*(1), 1-22. doi:10.3102/01623737026001001

DeWalt, D. A., Rothrock, N., Yount, S., et al. (2007) Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(1), 12-21. doi: 10.1097/01.mlr.0000254567.79743.e2

Ding, L., Reay, N. W., Lee, A., & Bao, L. (2009). Are we asking the right questions? Validating clicker question sequences by student interviews. *American Journal of Physics, 77*(7), 643-650. doi:10.1119/1.3116093

Ercikan, K., Arim, R., & Law, D. (2010). Application on think aloud protocols for examining and confi rming sources of differential item functioning identified by experts review. *Educational Measurement: Issues and Practices, 29*, 24-35. doi:10.1111/j.1745-3992.2010.00173.x

Goldstein, H. (2008). *How may we use international comparative studies to inform education policy.* Retrieved from http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/how-useful-are-international-comparative-studies-in-education.pdf

Haladyna, T. M. (1996). *Developing and validating multiple-choice test items*. NJ: Lawrence Erlbaum associates, publishers.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. USA: Allyn & Bacon.

Hopfenbeck, T. N., & Maul, A. (2011) Examining evidence for the validity of PISA Learning Strategy Scales based on student response processes. *International Journal of Testing, 11*(2), 95-121. doi: 10.1080/15305058.2010.529977

Johnstone, C., Figueroa, C., Yigal, A., Stone, E., & Laitusis, C. (2013). *Results of a cognitive interview study of immediate feedback and revision opportunities for students with disabilities in large scale assessments (Synthesis Report 92).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73. doi: 10.1111/jedm.12000

Linn, R. L. (1989). *Educational measurement*. NJ: American Council on Education and Macmillan Publishing Company.

Linn, R. L., & Gronlund, N. E. (1995). Measurement and assessment in teaching (7th ed.). Englewood Cliffs, New Jersey; Prentice Hall.

Lissitz, W. R., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. Educational Researcher, 36(8), 437-448.

Nicolaidis, C., Chienello, T., & Gerrity, M. (2011). The development and initial psychometric assessment of the centrality of Pain Scale. *Pain Medicine, 12*, 612-617.

Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education, 27*(4), 248–260.

Osterlind, S. J. (2002). *Constructing test items: Multiple-choice, constructed-response, performance and other formats.* New York: Kluwer Academic Publishers.

Ouimet, J. A., Bunnage, J. C., Carini, R. M., Kuh, G. D., & Kennedy, J. (2004). Using focus groups, expert advice, and cognitive interviews to establish the validity of a college student survey. *Research in Higher Education, 45*(3), 233-250.

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26,* 136-144. doi: 10.7334/psicothema2013.259

Pehlivan Tunç, E. B., & Kutlu, Ö. (2014). Investigation of Answering Behaviour in Turkish Test. *Journal of Measurement and Evaluation in Education and Psychology, 5*(1), 61-71.

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: Validity evidence based on content and response processes, *Measurement and Evaluation in Counseling and Development, 50*(4), 217-223, doi: 10.1080/07481756.2017.1339564

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small- and medium-scale evaluations. *American Journal of Evaluation, 33*(3), 414-30. doi:10.1177/1098214012441499

Sireci, S., & Faulkner-Bold, M. (2014). Validity evidence based on test content. Psicothema, 26, 1, 100-107. doi: 10.7334/psicothema2013.256

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481. doi: 10.3102/0013189X07311609

Snow, E. & Katz, I. (2009). Using cognitive interviews to validate an interpretive argument for the ETS ISKILLS assessment. *Communications in Information Literacy*, *3*(2), 99-127.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage Publications, Inc.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response.* NY: Cambridge University Press.

Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education, 91*(9), 1426-1431. doi: 10.1021/ed500076x

TIMSS 2007 Assessment. Copyright © 2009 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Wildy, H., & Clarke, S. (2009). Using cognitive interviews to pilot an international survey of principal preparation: A Western Australian perspective. *Educational Assessment, Evaluation and Accountability*, *21*(2), 105-117. doi: 10.1007/s11092-009-9073-3

Willis, G. (2015). *Analysis of the cognitive interview in questionnaire design (understanding qualitative research).* NY: Oxford University Press.

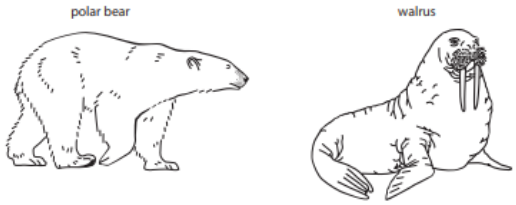
Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Appendix 1 – Selected Items

Item 1

Most birds sit on their eggs until they hatch. Which of these is the most important reason why birds sit on their eggs?
a) to keep the eggs inside the nest
b) to keep the eggs warm
c) to protect the egss from wind
d) to protect the eggs from the rain

Item 2



Polar bears and walruses look very different, but both can survive in extremely low temperature. Polar bears have a thick coat of fur that helps it keep itself warm. On the other hand, walruses have no fur.
What do walruses have to keep them warm?
a) Fat layers
b) Tusks
c) Whiskers
d) Flippers

Item 3

Sue measured how much sugar would dissolve in a cup of cold water, a cup of warm water, and a cup of hot water. What did she most likely observe?
a) The cold water dissolved the most sugar.
b) The warm water dissolved the most sugar.
c) The hot water dissolved the most sugar.
d) The cold water, warm water and hot water all dissolved the same sugar.

Item 4

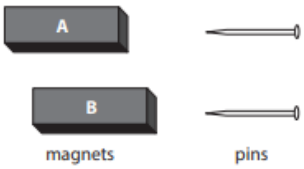
Maria designed an experiment using salt and water. Th e results of her experiment are shown in the table.

Amount of Salt Dissolved	Water Volume	Water Temperature	Was Mixture Stirred?
15 grams	50 ml	25°C	Yes
30 grams	100 ml	25°C	Yes
45 grams	150 ml	25°C	Yes
60 grams	200 ml	25°C	Yes

What was Maria studying in her experiment?
a) How much salt will dissolve in diff erent volumes of water.
b) How much salt will dissolve at diff erent temperatures.
c) If stirring increases how fast salt will dissolve.
d) If stirring decreases how fast salt will dissolve

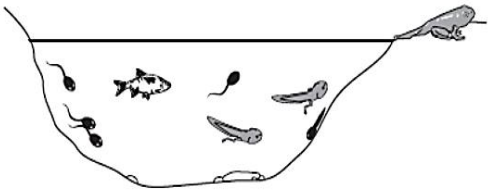
Item 5

Betty has two magnets (A and B) and two metal pins that are the same.
She slides Magnet A along a table until a pin is attracted to the magnet.
She slides Magnet B along a table until a pin is attracted to the magnet.



She finds that Magnet A attracts the pin from 15cm and Magnet B attracts the pin from 10cm.
Steven says that both magnets are equally strong. Do you agree? Explain your answer.

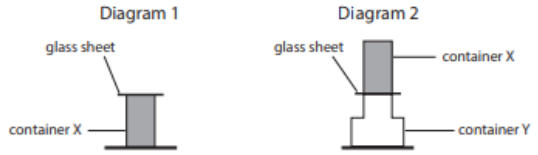
Item 6



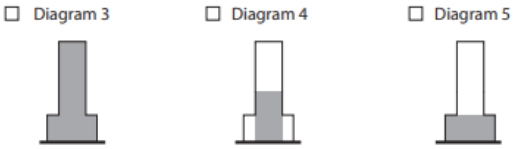
Melissa found some tadpoles and fish in a pond as shown above. How did the tadpoles get there?
a) They hatched from eggs laid by fish in the pond.
b) They formed from mud at the bottom of the pond.
c) They were made from materials dissolved in pond water.
d) They developed from eggs laid by frogs in the pond

Item 7

Diagram 1 shows a container X that is filled with a material that could be a solid, liquid, or gas. Th e container has been sealed with a glass sheet. Container X is placed upside down on an empty container Y, as shown in Diagram 2.



Th e glass sheet is removed.
A. Which of the diagrams below shows what you would see if the material in container X is a gas?
(Check one box.)



B. Explain your answer.

Item 8

Seeds from a plant can end up a long way away from the plant.
Describe one way that this can happen.