International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 218–235
https://doi.org/10.21449/ijate.1089472
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Identifying the presence of context and item-writing flaws in practice items: The case of Turkish mathematics textbooks

Munevver Ilgun Dibek[1,*], **Zerrin Toker**[2]

[1]TED University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]TED University, Faculty of Education, Department of Mathematics and Science Education, Ankara, Türkiye

ARTICLE HISTORY

Received: Mar. 17, 2022
Revised: Oct. 14, 2022
Accepted: Nov. 04, 2022

Keywords:
Context,
Item writing rules,
Mathematics textbooks.

Abstract: This study seeks to ascertain the degree to which context-based items are offered in Turkish mathematics textbooks as well as the quality of the items in terms of item writing guidelines, whether or not they are given as traditional or context-based. A qualitative research approach is used in this study. The eighth-grade mathematics textbook used in public schools and a textbook used in certain private school chains constitute its sample. The practice items (i.e, exercises without solutions given) included in the textbooks were analyzed by performing document analysis. The results revealed that both textbooks contain several flawed items in terms of item writing rules, as well as having mainly non-contextual items.

1. INTRODUCTION

Ensuring that students gain the requisite knowledge and skills to satisfy the demands and expectations of contemporary society is one of the goals of education in schools. To what extent people are citizens who have gained the knowledge and skills required for both personal and social life depends on their level of mathematics literacy (Geiger *et al*., 2015). It is crucial to foster mathematical literacy in the mathematics classroom to attain the ultimate goal of education (Bolstad, 2020). The term "mathematics literacy," which is one of the competencies assessed in the Programme for International Student Assessment (PISA), is defined as follows: (i) understanding and defining the role of mathematics in real life; (ii) making decisions based on mathematics in constructive, associative, and reflective ways in life; and (iii) making it a lifestyle (OECD, 2009).

A strategy to strengthen students' mathematics literacy is to use situations from life outside of school, considering the mathematical needs of current living. According to Kaiser and Willander (2005), students should be given questions that incorporate real-world scenarios where mathematical models can be employed to increase their mathematical literacy; thus, they can formulate the issue, create a model, and mathematically assess their findings. Goos *et al*. (2012) suggested a model (see Figure 1) to describe the complicated nature of mathematics

*Corresponding Author: Munevver Ilgun Dibek ✉ munevver.ilgun@tedu.edu.tr TED University, Faculty of Education, Department of Educational Sciences, Türkiye

e-ISSN: 2148-7456 /

literacy in general and numeracy in particular. They claimed that mathematics literacy is a broad interpretation of numeracy.

Figure 1. *A model for mathematics literacy (Goos et al., 2012, p. 149).*

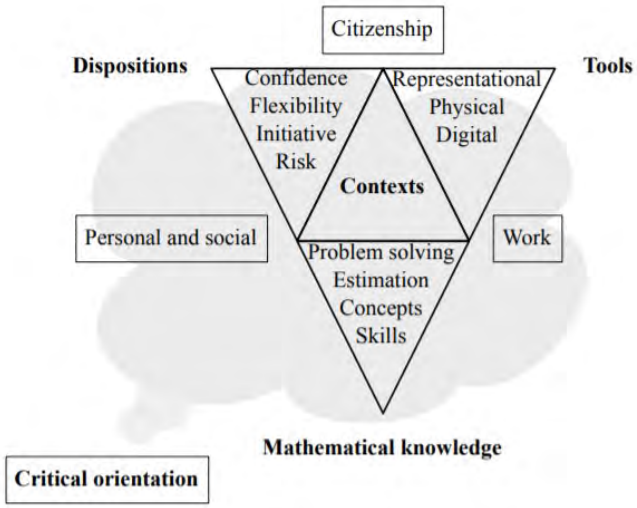


Figure 1 illustrates how literacy in many contexts is necessary for mathematical literacy. It is important to incorporate context into mathematics instruction and use context-based questions to increase students' mathematical literacy. Despite the importance of context in the development of students' mathematical literacy skills, students in diverse countries face difficulties when it comes to correctly answering these questions (Schwarzkopf, 2007; Verschaffel *et al*., 2000). A country where a majority of students had trouble responding to questions with context is Turkey. For example, the same problem is observed when the results regarding the mathematics literacy tests administered in different PISA cycles are examined. Although the average mathematics literacy scores of Turkish students increased slightly in each implementation year, the increase was not sufficient to exceed the OECD average. Specifically, in PISA 2018, Turkish students' mean score regarding mathematics literacy is 454 although the OECD average is 489. Similarly, although this increase brought Turkey to the forefront in some PISA cycles, it did not yield in great changes in its ranking in general (MoNE, 2019).

When given context-based questions or activities, students frequently struggle to discriminate between relevant and irrelevant information in the question as well as comprehend the nature of the problem and define the requisite steps to solve them (OECD, 2019a). Context-based questions are difficult for students to answer because they are solely used in evaluation procedures and do not have a place in the teaching process (Başaran, 2005; Fidan, 2018). In addition to the context, the role of the quality of items regarding compliance with item writing rules in teaching and learning is considered important. The high number of multiple-choice items in the books, where the writing process of both the item stem and the plausible distractors is difficult (Shin *et al*., 2019), necessitates revision of some items in Turkish textbooks according to the item writing principles (Kul *et al*., 2018; Simsek, 2016).

In the light of these issues, this study deals with revealing the situation regarding the extent to which students encounter such questions in the textbooks and the quality of the questions in the textbooks as one of the possible reasons for the difficulties experienced by Turkish students while solving context-based tasks and traditional items. Therefore, we analyzed practice items in textbooks in terms of context and quality regarding compliance with item writing rules. The approach adopted in the study is to examine what Turkish eighth grade mathematics textbooks offered to Turkish students regarding solving context-based items. Despite being conducted in a Turkish context, the study has the potential to contribute to the international literature by

offering details on the connections between students' learning and a number of aspects of the textbook's practice items.

2. THEORETICAL BACKGROUND AND RESEARCH QUESTION

2.1. The Role of Context in Teaching and Learning

Context can be considered as real-world settings, imaginary situations, or the formal world of mathematics (Van den Heuvel-Panhuizen, 2005). The realization of learning depends on the use of context-based questions. Learning can occur effectively when students can relate to an idea and its applications to their own culture, family, friends, or their daily lives (Yam, 2005). At some point during the learning process, every student wonders, "Why do we need to learn this?" However, very few students are able to provide appropriate responses to the questions that arise when they attempt to make sense of what they are being required to perform (Krouse, 2016). Therefore, using contexts for the development of mathematical thinking contributes to an understanding of mathematical concepts and prevents or eliminates misconceptions by improving the students' ability to use mathematics in various contexts of daily life. The use of daily life contexts as a didactic tool to support learning provides a meaningful basis for the concepts in the mathematics curriculum.

2.2. The Role of Quality of Items Regarding Compliance with Item Writing Rules in Teaching and Learning

Regardless of the psychological construct that is being tested, the method of creating the test items that will be used to measure it is crucial because the test items make up the structure of mental properties. Test items must therefore be described succinctly and clearly (Osterlind, 2002). Moreover, the high level of validity and reliability of the results obtained from the test depends on the quality of the items that make up the test. Specifically, the item must reflect the structure or content to be measured for the results from the test to provide valid interpretation (Peeters *et al*., 2013). If a strong link is not established between the test item and the psychological construct to be measured or its content, the item will lose its purpose and will not be different from a thought that circulates freely on a test page (Osterlind, 2002). Besides, the difficulty level of the test item increases due to item-writing flaws. In other words, construct-irrelevant variance is introduced to the results obtained from the test item; therefore, the reliability of the results to be obtained decreases (Downing, 2005). Hence, the interpretability of test results is closely associated with the quality of the item.

Certain technical features should be considered to ensure the high quality of the test item. For example, the use of the correct item format, level of complexity of the words used, use of a sufficient number of answer options, and absence of negative words are a few of these features. Every word is valuable in a test item. The test-taker should be able to understand the meaning of the item's stem and recognize the incorrect choices/distractors from the correct one (Osterlind, 2002). The way the items are built is crucial for the students, the researchers who will use the assessment results, and the program evaluators since the test items serve as the fundamental building blocks of the assessment tools. In this context, examining the existence of item-writing flaws in the textbooks to be used in the teaching and learning process will provide valuable information.

2.3. Context of the Study and the Case of Turkey

The teaching process has not regularly used context-based questions until now because textbooks do not contain enough context-based questions, and instructors could lack expertise in this subject and feel unqualified (Kayhan Altay *et al*., 2020). For instance, in a study conducted by Fidan (2018) teachers said that assessment questions from textbooks and context-based questions are incompatible. Similar results were found in the study of Kayhan Altay *et al*. (2020) who focus on the context and daily life in sixth-grade mathematics textbook. In

studies focusing on a particular section of the textbooks, it is stressed that items presented directly, rather than through a mathematization context, come to the fore (e.g, Kar & Işık, 2015). In parallel with teachers' opinions, since statewide exams play a significant role in students' life, students state that they desire tests that reflect what is expected of them in their textbooks and lectures (Başaran, 2005). In Turkey, when compared with previous administrations, the recent statewide exam called the High School Entrance Exam (LGS in Turkish) includes many context-based questions (Güler & Ülger, 2018). However, recent research indicates that teachers have complained that the exam is incompatible with educational materials like textbooks (e.g., Korkmaz *et al*., 2020).

Students need to be familiar with tasks involving contexts within the teaching and learning process for them to be successful in answering such questions. Textbooks, that play an important role in the planning of teaching, are expected to include such tasks (Korkmaz *et al*., 2020). Given the strong relationship between textbooks and educational processes, it is crucial to understand how much opportunity for activities, items, and other contents—including context—are provided by textbooks, which support educational processes. To the best of our knowledge, while numerous studies (e.g., Hadar, 2017; Törnroos, 2005; Wijaya *et al*., 2015) have examined mathematics textbooks in relation to learning opportunities and students' mathematics achievement, no study has looked at the items in mathematics textbooks in two dimensions, such as context and the quality of item regarding compliance with item writing rules. Examining to what extent and how such questions are included in the textbooks currently in use will contribute to understanding the difficulties experienced by students.

2.4. Research Purpose

This study attempts to investigate the practice items in Turkish eighth grade mathematics textbooks, which are utilized as main course materials by teachers, in terms of context and their compliance with the item writing principles. Within this context, this study seeks answers to the following research questions:

(1) To what extent do Turkish eighth grade mathematics textbooks offer context-based practice items?

(2) What are the item-writing flaws of practice items in Turkish eighth grade mathematics textbooks?

3. METHOD

3.1. Research Design

The present study aims to examine several aspects of the practice items included in Turkish eighth grade mathematics textbooks, such as context and quality regarding item writing rules. In this regard, a document analysis is used in this study. It is a systematic procedure for reviewing or evaluating documents including text and images that the researcher did not interfere with (Bowen, 2009).

3.2. Sample

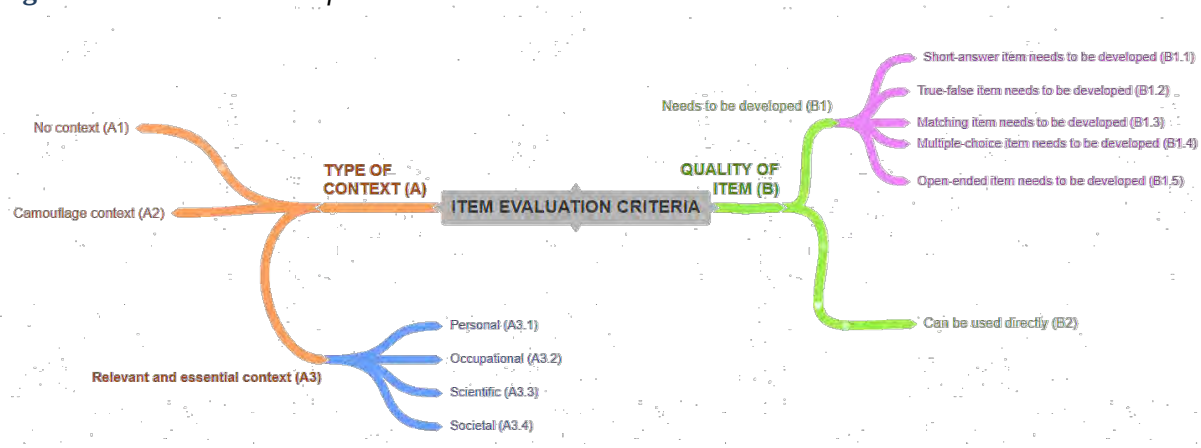
The eighth-grade mathematics textbooks, used by public schools and one of the private school chains, constitute the sample of the study. These two textbooks (hereafter referred to as Book 1 and Book 2) have been selected using the purposive sampling method that enables researchers to select their sample according to predefined criteria (Fraenkel *et al*., 2012). The reason for choosing Book 1, approved by MoNE, is that this book is used as the principal course resource in all schools, while Book 2 was chosen to increase the representativeness of the eighth grade mathematics textbooks used in private schools. More precisely, a private school, where the number of students in the 8th grade level is higher than other private schools, uses Book 2. Moreover, some of the other private schools use Book 2 as a supplementary material in the

mathematics course at the 8th grade level. Eight grade level students were chosen because, according to different PISA cycles that are pioneering applications where mostly context-based items are used, grade 8 can be considered a relevant grade level to prepare students to be able to solve context-based tasks (Wijaya *et al*., 2015). Also, 8th grade Turkish students attend centralized exams, indicating the tendency to include context-based items.

3.3. Data Collection and Analysis

Data collection and analysis were performed by using a two-dimensional framework given in Figure 2.

Figure 2. *Two-dimensional framework.*



Context Analysis Form (CAF) was utilized to provide an answer to the first research question regarding the context of the practice items. In addition, Checklists for Evaluating Item Quality (CEIQs) were employed to address the second research question about item-writing flaws in practice items. By using these tools, data collection and analysis were performed simultaneously.

3.3.1. *Context analysis form (CAF)*

One of the tools utilized to collect data for the study was the CAF, which was used to assess the context-related aspects of the items from the textbooks that were the subject of the current investigation. The Wijaya *et al*. (2015) classification, which is more appropriate for real-world circumstances and 21st century abilities, is the basis for the subcategories of the CAF. They were coded as no context (A1), camouflage context (A2), and relevant and essential context (A3). The framework of PISA (OECD, 2019b) was taken as the basis for the items that were determined to be context-based. Accordingly, personal, occupational, societal and scientific contexts are coded as A3.1, A3.2, A3.3, and A3.4, respectively. Explanations related to each code and category of CAF are provided in Appendix 1.

3.3.2. *Checklists for evaluating item quality (CEIQs)*

CEIQs were employed as additional data collecting tools to get information regarding the second dimension depicted in the framework shown in Figure 2. More specifically, different Checklists for Evaluating Item Quality (CEIQs) (see Appendix 2) were created by considering the recommendations made by Miller *et al*. (2013) to determine the quality related conformity with item writing rules of the items. These checklists were used to determine whether the item violates any item writing guidelines and, if so, what specific violations it may contain. At this point, it has been decided whether the item will be used directly, based on the rules in the relevant checklist, depending on the type of item (open ended, multiple choice, true-false, etc). The item that does not have the item-writing flaws indicated in the relevant checklist is defined as “the item that can be used directly” by giving the B2 code.

3.3.3. *Training of the coders*

Two researchers, one with expertise in measurement and evaluation and the other in mathematics education, carried out the process of assessing the practice questions in the chosen books in terms of several dimensions and assigning codes to them in this study. These coders have knowledge of the dimensions addressed by the current study. Specifically, throughout their doctoral studies, these researchers completed a number of graduate courses about test items and item structures. They currently teach undergraduate-level courses focusing on these topics and various taxonomies for the classification of learning outcomes. A third coder—a pre-service teacher—was brought in when the two researchers could not agree on a particular item. It was ensured that the chosen preservice teacher had sufficient understanding of the various item types and how to write an appropriate item in accordance with item-writing guidelines. The pre-service teacher received in-depth instruction from the two researchers on the aspects of analysis and coding of sample items prior to the use of coding. Each category in the data gathering tools was examined individually during this training, and what was meant to be communicated was discussed. Consequently, it was ensured that each coder assigned identical meanings to each code.

3.3.4. *Coding procedure*

Each item received a location code in the book and a dimension code that covered the two dimensions of the analysis demonstrated in Figure 2 throughout the coding process. For example, the dimension of "context" received the letter "A," whereas the dimension of "item quality" received the letter "B." Additionally, sub-codes were given to the items to precisely specify the subcategory they belong to. Figure 2, that demonstrates the framework used in the present study's coding process, contains more details. Moreover, in cases where a poor item in the textbook exists, possible item-writing flaws that an item possesses have been previously listed in CEIQs, and codes have been assigned to each item-writing flaw to indicate the kind of item-writing flaw an item possesses. Since each rule included in each CEIQ is stated as a question, the negative answer to each question considered that the particular item has an item-writing flaw, and was coded as B1. Further code was assigned to specify which item-writing flaw an item had. On the other hand, items with the positive answer to these questions are coded as B2. More than one code was given to the item when the item included more than one problem regarding item quality. The number of items in each category was counted at the end. A sample of codes given to an item is shown in Figure 3.

Figure 3. *The sample of codes given to an item.*

This question asks how many laps both motorcycle racers completed when they first came together at the starting point after completing a lap in 9 minutes and 12 minutes respectively. The location code assigned to the item shown in Figure 3 was "B2, U1, P38, I13," and the dimension code for this item was "A2, B1.4.2." First of all, if we interpret the location code, this is a question from the book used in certain private schools that we consider as Book 2 (B2). Moreover, this is the thirteenth item (I13) on page 38 (P38) in unit 1 (U1). To continue with the dimension code, the camouflage context (A2) is used in this item. In addition, the item has a problem in terms of being a qualified item. The negative answer to rule B1.4.2 (see Appendix 2) indicated that the stem of the item presents an unnecessary element. More specifically, the use of the image in this item is not necessary to solve the problem.

In addition to the authors of the current study, one measurement and evaluation specialist and one mathematics educator were consulted regarding the dimensions and definitions in the data collection tools developed or adapted in order to establish the validity of the results obtained from various data collection tools, such as CAF and CEIQs. In response to their suggestions, the data collection tools were changed. Additionally, the coders conducted pilot coding using all of the data collection tools before the researchers coded every task in the two textbooks to ensure that they comprehended each criterion in the same manner. This strengthened the validity of the conclusions drawn from the measurement results. Within this context, as in similar studies (e.g., Wiijaya *et al*., 2015), 15% of the items included in each textbook selected within the scope of the research were coded independently by all the coders. Items to be coded by all the coders were randomly selected. Interrater reliability was calculated for each dimension of the analysis to determine the reliability of the results obtained from this coding procedure. For this, "the agreement percentage formula" developed by Miles and Huberman (1994, p. 64) was used. Accordingly, the formula is as follows:

$$\text{Agreement percentage} = \frac{(\text{the number of agreement})}{(\text{the number of agreement} + \text{the number of disagreement})} \times 100$$

The scorer agreement coefficients of context dimension and item quality during the pilot coding procedure were found to be .95 and .90, respectively. The raters' coding is consistent because the coefficient is larger than .90 (Miles & Huberman, 1994). The items that the coders were not in agreement about were returned and the coding for them was repeated until agreement was achieved. The frequency and percentage values for the number of items gathered under each dimension were then presented following the final item coding.

4. FINDINGS

4.1. Context Dimension

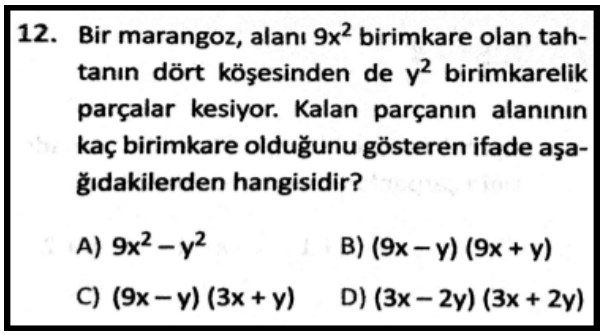
Table 1 displays the findings of the analysis of the context dimension of the items found in Books 1 and 2.

Table 1. *Results of the analyses of items in terms of context dimension.*

Context		Book 1											Book 2	
		Multiple-Choice		Open-Ended		Short Answer		Matching		T-F Items			Multiple-Choice	
		f	%	*f*	%	*f*	%	*f*	%	*f*	%	Total	*f*	%
A1		131	74.70	9	81.80	16	76.20	5	83.30	22	78.60	183	464	80.00
A2		35	16.30	-	-	2	9.50	1	16.70	4	14.30	42	55	9.50
A3	A3.1	4	2.20	-	-	1	4.80	-	-	2	7.10	7	21	3.60
	A3.2	6	4.50	1	9.10	1	4.80	-	-	-	-	8	6	1.00
	A3.3	-	-	-	-	-	-	-	-	-	-	-	13	2.20
	A3.4	2	2.20	1	9.10	1	4.80	-	-	-	-	4	21	3.60
Total		178	100	11	100	21	100	6	100	28	100	244	580	100

Table 1 demonstrates that when the context of the items in the two books is compared, the majority of the items in Book 1 (f = 183) do not contain any contextual components, while some of them use camouflage context (f = 42), and a small number of items include occupational, personal, and scientific context elements. A similar pattern is noticed for Book 2. A few items (f = 55) in Book 2 have camouflage context, while the majority of items (f = 464) do not use context. Similar to Book 1, Book 2 has a small number of items (f = 61) with relevant and essential contexts. It was also discovered that there are more items with personal and scientific context than any occupational or societal context. Figures 4 and 5 show examples of contexts related to camouflage and personal context, respectively.

Figure 4. *Example for camouflage context (coded as A2).*

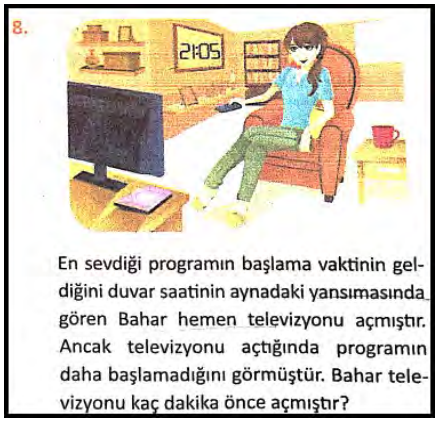

12. Bir marangoz, alanı $9x^2$ birimkare olan tahtanın dört köşesinden de y^2 birimkarelik parçalar kesiyor. Kalan parçanın alanının kaç birimkare olduğunu gösteren ifade aşağıdakilerden hangisidir?

A) $9x^2 - y^2$ B) $(9x - y)(9x + y)$

C) $(9x - y)(3x + y)$ D) $(3x - 2y)(3x + 2y)$

In the item displayed in Figure 4, it is stated that a carpenter cuts sections of y^2 square units from each of the board's four corners, with an area of 9x2. The students are then asked to calculate the area of the piece that is left over. Due to the fact that it does not just refer to mathematical objects, symbols, or structures, this item is classified as having "camouflage context." On the other hand, the context is not necessary and the operations needed to solve the problems are already obvious; the answer is simply obtained by adding the numbers provided in the item.

Figure 5. *Example of personal context falling under the category of relevant and essential context (coded as A3.1).*



In this item, it is said that a person who noticed her favorite television program through the reflection of a wall clock in a mirror realized that the program had not yet started when she turned on the television. The students were asked to calculate how many minutes earlier she might have turned on the television. Given that context is necessary to comprehend the issue and find a solution, this item is placed under the category of "relevant and essential context," and categorized under "personal" because the context relates to personal life.

4.2. Quality Dimension regarding Compliance with Rules of Item Writing

Table 2 displays the findings of the study of items from Books 1 and 2 in relation to the item quality dimension.

Table 2. *Analysis of items in terms of the quality dimension.*

Book		Book 1											Book 2	
Item Type		Multiple-choice		Open-ended		Short Answer Completion		Matching		T-F Items		Total	Multiple-choice	
		f	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	*f*	%
Quality	B1	31	17.40	3	39.40	4	19.00	6	100	0	-	44	190	32.80
	B2	147	82.60	8	72.70	17	81.00	0	-	28	100	200	390	67.20

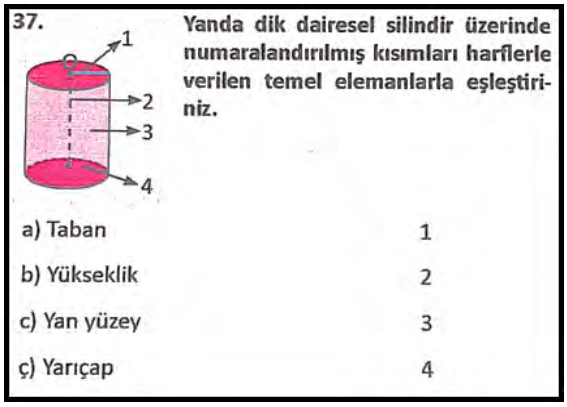
When the items selected within the scope of the research were examined in terms of their compliance with the principles of item writing and their quality, it was found that the number of items written by considering these principles (coded as B2) was higher than the number of items in which these principles were neglected (coded as B1). However, it should be noted that there are still a sizable number of items created that do not consider these principles. The various types of items, such as short answers, matching, and open-ended questions, are provided in Table 3 to highlight common item-writing flaws in Book 1, as Book 2 only contains multiple-choice questions.

Table 3. *Frequently-made mistakes in writing qualified items included in Book 1.*

Item Type	Criteria List	Total Number of Items Reviewed	*f*	*(%)*
Short-answer	Do the answers blank place at the end of the items?	21	2	9.50
	Do the items contain any clues?		2	9.50
Matching	Do the responses rank alphabetically or numerically?	6	2	33.30
	Do the directions specify the number of times each response may be used?		4	66.70
Open-ended	Does the material to be interpreted contain some novelty to require interpretation?	11	1	9.10
	Does each question specify the expected response?		2	18.20

Table 3 shows that out of the 21 short-answer questions, 4 contained some specific item-writing flaws. The short answer questions in Book 1 specifically had item-writing flaws in that the blanks were not at the end of the items and some clues would reveal the solution within the item. All matching type items found in Book 1 have errors according to item-writing rules. Common item-writing flaws observed in the matching type items are that there is no information on how many times the expressions/numbers can be used in the response column, and that these expressions/numbers are not in alphabetical or numerical order. Among 11 open-ended items, the neglected item writing principle is that the expected answer from the student should be made explicit. An example of an item-writing flaw in a matching question is provided in Figure 6.

Figure 6. *Example of a mistake in a writing matching item (coded as B1.3.4 and B 1.3.6).*



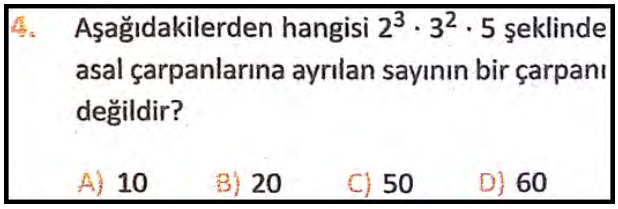
The students were required to match the numbers of a right circular cylinder with the cylinder's fundamental components in this item. This item's quality was judged to contain two item-writing flaws. The first one is that each response is not allowed to be used more than once in the item's instructions. Another is that the total number of premises in the premise column and the total number of responses in the response column are the same. However, more statements in the response column are needed. Otherwise, even though they are unfamiliar with the concept needed to match the final premise and response, the students are still able to accomplish it. Table 4 lists the common item-writing flaws that were made when creating multiple-choice items for the two books analyzed within the scope of the research.

Table 4. *Frequently-made mistakes in the creation of multiple-choice items.*

Criteria List for Multiple-Choice Item	Book 1		Book 2	
	f	%	*f*	%
Is each item stem meaningful?	6	3.40	7	1.20
Do the item stems contain irrelevant material?	1	.60	9	1.60
If used, has negative wording been given special emphasis (for example, capitalization)?	-	-	75	12.90
Is there grammatical consistency between the alternatives and the item stem?	-	-	1	.20
Are the alternative answers brief and free of unnecessary words?	-	-	5	.90
Are the length and form of the alternatives similar?	4	2.20	15	2.60
Are the distractors plausible to low achievers?	10	5.60	21	3.60
Do the items contain any verbal clues to the answer?	-	-	2	.30
Do the verbal alternatives rank alphabetically?	1	.60	-	-
Do the numerical alternatives rank numerically?	9	5.10	55	9.50

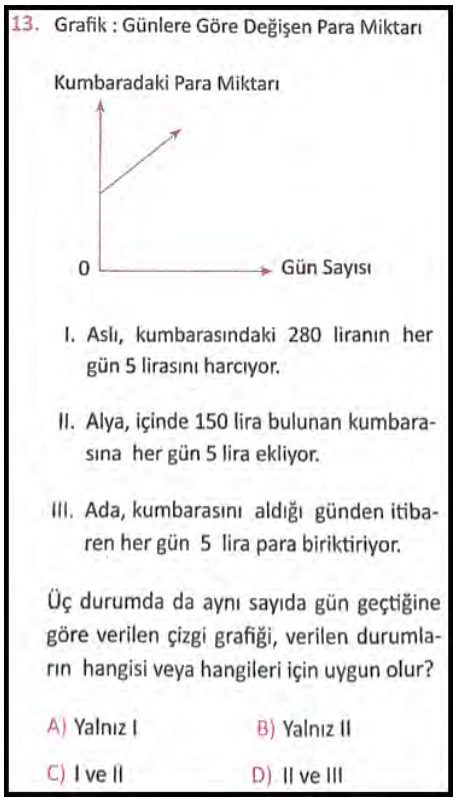
Table 4 demonstrates that failing to make the distractors plausible enough was the most frequently observed item-writing flaw in Book 1. Another typical one was that the response options that were numerical were not presented in a sequential order. On the other hand, failing to emphasize the negative statements at the stem of the multiple-choice questions was the item-writing flaw that was observed frequently in Book 2. The distractors were not written in a numerical order, another common item-writing flaw in the items in this book, similar to the case in Book 1's items. Figures 7 and 8 illustrate examples of item-writing flaws observed when creating multiple-choice questions.

Figure 7. *Example of a mistake in writing multiple-choice item (coded as B1.4.3).*



"Which of the numbers provided in the response options is the prime factorized number?" is the question in this item. This item violates the rules for item writing because the negative phrase "it is not" (in Turkish, "değildir") was not highlighted or stressed. Another common item-writing flaw in the item writing approach is demonstrated in Figure 8 for another item.

Figure 8. *Example of an Item-Writing Flaw in the Multiple-Choice Item (coded as B1.4.7).*



In the item displayed in Figure 8, it is asked for which of the cases listed in I, II and III can the graph showing the amount of money in the penny bank and the number of days be created. In Case I, it is stated that Aslı spends 5 liras every day of her 280 liras in the penny bank. In Case II, it is indicated that Alya adds 5 liras every day to her penny bank, which currently holds 150 liras. In the last case, Case III, it is stated that Ada has saved 5 liras every day since the day she received her penny bank. The distractors of this item are not plausible enough, because a student who knows that this graph cannot be drawn for Case I can directly rule out options A and C.

5. DISCUSSION and CONCLUSION

The present study demonstrates that Turkish eighth-grade mathematics textbooks rarely include context-based items. Most of the items in these textbooks is non-contextual and does not require mathematization. In other words, this study shows that the items in the eighth-grade mathematics textbooks, commonly-used in Turkey, are insufficient in making connections to real-world situations in terms of personal, scientific, occupational, and social aspects. The results of two studies—one by Kayhan Altay *et al*. (2020), that investigated the contexts used

for real-life connections in mathematics textbooks for sixth graders and found that more than half of the tasks presented in the textbook are not related to real life, and another by Kar & Işık (2015), that examined Turkish mathematics textbooks in a more specific area, concentrating on addition and subtraction operations with integers—support this conclusion. This situation with Turkish textbooks is also observed in the textbooks of a few other countries that fall behind OECD average like Turkey in PISA, where context plays an important role in the measurement of literacy. For example, Indonesia shows similar patterns in terms of mathematics literacy performance in PISA 2018 (OECD, 2019a) and mathematics textbooks. It appears that the results of the current study are consistent with those of Wijaya *et al*. (2015), who looked at the learning opportunities provided by Indonesian textbooks for completing context-based mathematical activities. One reason for this situation might be that sufficient information about curriculum change must be given for the existing curriculum framework to be implemented, (Rea-Dickson & Germanie, 2001).

The results of this study show that multiple-choice items make up the majority of the material covered in Turkish textbooks. This result is in line with the results of the study conducted by Kul *et al*., (2018) which analyzed the item types in Turkish and Canadian textbooks and discovered that multiple-choice items made up a higher percentage of the items in Turkish textbooks. Multiple-choice items are more prevalent than other item types in the eighth-grade mathematics textbooks used in Turkey, which may be explained by the fact that these types of items also appear in the middle to high school transition exam. The 8th grade level, the level covered by the mathematics textbook under investigation in this study, is the stage between secondary school and high school. Students take a centralized test at this transitional level. They are exposed to questions that are similar to the item types in this exam during the learning and teaching process to succeed in this high stakes exam. In other words, this exam system, where significant decisions are made depending on the results, also impacts the teaching process (Kahraman, 2014). Consequently, the course textbooks now contain more multiple-choice questions.

When the frequencies of the item-writing flaws in multiple choice were compared for both books, it was concluded that Book 1 had fewer item-writing flaws than Book 2. Since Book 1 was approved by MoNE, both field experts and assessment and evaluation experts took part in the item writing process in Book 1. Therefore, the item writing process could have been conducted more meticulously, and the relevant item redactions could have been made. Accordingly, there may have been a decrease in the number of item-writing flaws related to multiple-choice items. Additionally, in terms of the type of the most commonly observed item-writing flaws in constructing the multiple-choice items in eighth-grade textbooks addressed in the present study, this study shows that negative statements in the stem of the item are not emphasized, and plausible distractors are not developed. The learner might not pay attention if the negative term at the stem of the multiple choice item is not highlighted. Even though the student is aware of the right answer, they may still respond incorrectly since they failed to notice the negative word. However, the primary goal of multiple-choice questions is to discover whether students have acquired the idea being measured, not to gauge how attentive they are (Chiavaroli, 2017). Additionally, asking students to identify the incorrect options is not a preferred method in teaching. Just because someone is aware of the incorrect options does not imply that they are also aware of the solution (Burton *et al*., 1991).

This study's conclusion is consistent with the results of the study conducted by Simsek (2016), who compared the items created by teachers and trainers and found that almost 60% of them need improvement. The two most frequently observed item-writing flaws were the use of implausible distractors and the use of negative items without emphasizing the negative features of the items. The use of distractors like this (i.e., using implausible distractors) causes the

question with more response options to function as an item with fewer response options, increasing the possibility of getting the right answer just by chance (Royal & Stockdale, 2017); even if the students do not know the answer to the question, it causes them to eliminate distractors without prompting them to think and directly turn to the right answer. This reduces the item's ability to discriminate (Rush *et al*., 2016). In other words, the item will no longer be sufficient to distinguish between students who met the required learning goals and those who did not. Since creating plausible distractors and producing a high-quality multiple-choice test item stem are challenging tasks that require time and expertise, it may be understandable to use a lot of multiple-choice items with problematic distractors (Shin *et al*., 2019). For instance, in a test with 100 multiple-choice questions, each with five response alternatives, 400 distractors should be prepared along with 100 item stems and 100 right answers (Gierl *et al*., 2017). So even if it is not ideal, it is fairly obvious that writing illogical distractors is an often made blunder.

The findings of this study offer important insights into how context-based textbooks are currently written, as well as an understanding of the qualities of good context-based items to educational politicians who direct item writers and textbooks writers. Consequently, this study might be able to provide information that can be used in textbook preparation. More specifically, it is suggested that, in light of the findings of the present study, mathematics textbooks should include more context-based materials and students should be required to employ mathematization for these questions. To put it another way, more relevant and essential contexts should be used in the eighth grade mathematics textbooks. Additionally, whether in traditional form or a context-based form, the items' quality in terms of conformity with item writing rules should take precedence. The item cannot assess the material in a valid and reliable manner if the item writing principles are ignored.

The study, even if its primary focus is on the analysis of the practice problems in the Turkish eighth grade mathematics textbooks, also has the potential to provide a framework for increasing practitioners' knowledge of selecting qualified items. Teachers can choose from the pre-existing items, make necessary modifications, or use the forms as a checklist to create new items using the present study's forms. Along with the quantitative and qualitative results of the study, the implementation process can therefore aid future practices.

When conclusions are drawn from the findings of the present study, the following limitations need to be considered, since they also point to future possible research trajectories. First, it is incorrect to just attribute the low achievement of Turkish students, particularly in large scale assessments like PISA, to the inadequateness of the textbooks used by that age group in Turkish schools. As previously mentioned, different teachers differentiate their instructions even while using the same items. The fact that the items in the books meet the criteria for the dimensions considered in the context of this study does not, therefore, ensure the quality of the instruction. Future research should look into how much teachers use these textbooks and particularly items in those books in their lessons. Second, this study is limited only to mathematics textbooks. As context-based items are also featured in other subjects on national central exams and in large-scale assessments, textbooks of other courses, such as Turkish, Science, and Social Studies, could also be analyzed in the framework of the criteria stated in this study. Third, because the age group for the test, eighth graders, is the only one included in this study, additional research may be conducted with students of other grade or age levels. Lastly, although this study reflects the situation regarding 8th grade mathematics textbooks in Turkey, its results may also be useful for the 8th grade students in other countries below the OECD average in terms of mathematics performance in large scale assessments. International comparative studies might be carried out by identifying and selecting the textbooks of such countries to generalize.

Acknowledgments

This research was supported by the TEDU IRF Project (T-20-B2010-90025).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Authors conducted the whole research process including conceptualization, research design, literature review, data collection and processing, interpretation, writing and critical review together.

Orcid

Munevver Ilgun Dibek https://orcid.org/0000-0002-7098-0118
Zerrin Toker https://orcid.org/0000-0001-9660-0403

REFERENCES

Başaran, S. (2005). *Diğer ülkelerde lise bitirme sınavları ve Türk eğitim sistemi için lise bitirme sınavı önerisi [High school leaving exams in other countries and high school leaving exam recommendation for the Turkish education system]*. MEB Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

Bolstad, O.H. (2020). Secondary teachers' operationalisation of mathematical literacy. European *Journal of Science and Mathematics Education, 8*(3), 115-135. https://doi.org/10.30935/scimath/9551

Bowen, G.A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27-40. https://doi.org/10.3316/QRJ0902027

Burton, S.J., Sudweeks, R.R., Merrill, P.F., & Wood, B. (1991). How to prepare better multiple-choice test items: Guidelines for university faculty. Bringham Young University Testing Services and the Department of Instructional Science. http://testing.byu.edu/info/handbooks/betteritems.pdf

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research & Evaluation, 22*(3), 1-14. https://doi.org/10.7275/ca7y-mm27

Downing, S.M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences, 10,* 133-143. https://doi.org/10.1007/s10459-004-4019-5

Fidan, M. (2018). Ortaokul öğrencilerinin Türkçe ders kitaplarının tasarımına yönelik görüşlerinin analizi [Analysis of middle school students' views on the design of Turkish textbooks.]. *Bayterek International Journal of Academic Research,1*(2), 178–189.

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw Hill.

Geiger, V., Goos, M., & Forgasz, H. (2015). A rich interpretation of numeracy for the 21st century: A survey of the state of the field. *ZDM Mathematics Education, 47*(4), 531–548. https://doi.org/10.1007/s11858-015-0708-1

Gierl, M.J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Goos, M., Geiger, V., & Dole, S. (2012). Auditing the numeracy demands of the middle years curriculum. *PNA, 6*(4), 147-158. https://doi.org/10.30827/pna.v6i4.6138

Güler, H.K. & Ülger, B. (2018). PISA, TIMSS ve TEOG sınavlarının temele aldığı öğrenme kuramları [Learning theories based on PISA, TIMSS and TEOG exams]. In S. Çepni (Ed.), *PISA ve TIMSS mantığını ve sorularını anlama* (ss.111-153). Pegem A Yayıncılık.

Hadar, L.L.(2017). Opportunities to learn: Mathematics textbooks and students'achievements, *Studies in Educational Evaluation, 55,* 153-166. http://dx.doi.org/10.1016/j.stueduc.2017.10.002

Kahraman, İ. (2014). Merkezi ortak sınav uygulamasının etkilerine ilişkin öğretmen görüşleri [The effect of common implementation that related to teachers' opinion]. *Tunceli Üniversitesi Sosyal Bilimler Dergisi, 2*(4), 53-74.

Kaiser, G., & Willander, T. (2005). Development of mathematical literacy: results of an empirical study. *Teaching Mathematics and Its Applications, 24*(2-3), 48-60. https://doi.org/10.1093/teamat/hri016

Kar, T. & Işık, C. (2015). Comparison of Turkish and American seventh grade mathematics textbooks in terms of addition and subtraction operations with integers. *Education and Science, 40*(177), 75-92. https://doi.org/10.15390/EB.2015.2897

Kayhan Altay, M., Kurt Erhan, G. & Batı, E. (2020). Contexts used for real life connections in mathematics textbook for 6th graders. *Elementary Education Online, 19*(1), 310-323. https://doi.org/10.17051/ilkonline.2020.656880

Korkmaz, E., Tutak, T., & İlhan, A. (2020). Ortaokul matematik ders kitaplarının matematik öğretmenleri tarafından değerlendirilmesi [Evaluation of secondary school mathematics textbooks by mathematics teachers]. *Avrupa Bilim ve Teknoloji Dergisi, 18,* 118-128. https://doi.org/10.31590/ejosat.667689

Krouse, S. (2016, Jan 8). Why do we need to learn this. *Medium*. https://medium.com/@stevekrouse/why-do-we-need-to-learn-this-3ba1d42bd08a.

Kul, Ü., Sevimli, E., & Aksu, Z. (2018). A comparison of mathematics questions in Turkish and Canadian school textbooks in terms of synthesized taxonomy. *Turkish Journal of Education, 7*(3), 136-155. https://doi.org/10.19128/turje.395162

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook.* Sage.

Miller, D.M., Linn, R.L, & Gronlund, N.E. (2013). *Measurement and assessment in teaching* (11th ed.). Pearson Education, Inc.

MoNE (Ministry of National Education) (2019). *PISA 2018 ulusal ön raporu [PISA 2018 Preliminary National Report]*. Eğitim Analiz ve Değerlendirme Raporları Serisi,10.

OECD (2009). *Learning mathematics for life: A view perspective from PISA* OECD Publishing.

OECD (2019a). *PISA 2018 results (Volume I): What students know and can do.* PISA OECD Publishing. https://doi.org/10.1787/5f07c754-en

OECD, (2019b). PISA 2018 assessment and analytical framework. PISA OECD Publishing. https://doi.org/10.1787/b25efab8-en

Osterlind, S.J. (2002). What is constructing test items? In S. J. Osterlind (Ed.), *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (pp. 1–16), Springer. https://doi.org/10.1007/0-306-47535-9_1

Peeters, M.J., Beltyukova, S.A., & Martin, B.A. (2013). Educational testing and validity of conclusions in the scholarship of teaching and learning. *American Journal of Pharmaceutical Education, 77*(9), 1-9. https://doi.org/10.5688/ajpe779186

Rahimah, D. & Visnovska, J. (2021). Analysis of mathematics textbook use: An argument for combining horizontal, vertical, and contextual analyses. *Journal of Physics: Conference Series, 1731*, 1-5. https://doi.org/10.1088/1742-6596/1731/1/01204

Rea-Dickson P. & Germania, K. (2001). Evaluating curriculum change. In D. Hall & A. Hewimng (Eds.), *Innovation in English language teaching: A reader*. British Library Catalogue.

Royal, K.D. & Stockdale, M.R. (2017). The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations. *Journal of Advances in Medical Education & Professionalism, 5*(2), 84-89.

Rush, B.R., Rankin, D.C & White, B.J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education, 16*(250), 1-10. https://doi.org/10.1186/s12909-016-0773-3

Schwarzkopf, R. (2007). Elementary modeling in mathematics lessons: The interplay between real-world knowledge and mathematics structures. In W. Blum, P. L. Galbraith, H.W. Henn, & M. Niss (Eds.), *Modelling and applications in mathematics education: The 14th ICMI study* (pp. 209–216). Springer.

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. Frontiers in Psychology, 10, 1-14. https://doi.org/10.3389/fpsyg.2019.00825

Simsek, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education, 4*(4), 477-489.

Törnroos, J. (2005). Mathematics textbooks, opportunity to learn and student achievement. *Studies in Educational Evaluation, 31*, 315-327. https://doi.org/10.1016/j.stueduc.2005.11.005

Valverde, G., Bianchi, L, Wolfe, R., Schmidt, W. & Houang, R. (2002). *According to the book: Using TIMSS to investigate the translation of policy into practice through the world of textbooks*. Kluwer Academic Publishers.

Van den Heuvel-Panhuizen, M. (2005). The role of context in assessment problems in mathematics. *For the Learning of Mathematics, 25*(2), 2-23.

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Swets & Zeitlinger.

Wijaya, A., van den Heuvel-Panhuizen, M., & Doorman, M. (2015). Opportunity-to-learn context-based tasks provided by mathematics textbooks. *Educational Studies in Mathematics, 89*(1), 41-65. https://doi.org/10.1007/s10649-015-9595-1

Yam, H. (2005). What is contextual learning and teaching in physics? Retrieved from http://www.phy.cuhk.edu.hk/contextual/approach/tem/brief_e.html

APPENDIX

Appendix 1. Context Analysis Form (CAF)

Sub-category (Code)	Explanation	
No context (A1)	Contains only mathematical symbols or structures	
Camouflage context (A2)	Daily life experiences and reasoning are not required. The mathematical operations required to give answer to the problems are already clear. The results can be found by combining the numbers given in the question text.	
Relevant and essential context (A3)	To provide answer to problem, common sense of reasoning within the context is necessary. The mathematical operation necessary for solving the problem is not obvious Mathematical modeling is necessary.	The item is included in the '*personal*' category if the item is related to students' families, their lives, such as shopping, games, personal life and so on (A3.1). The item is included in the '*occupational*' category if the item is related to the job/profession such as measuring, architecture, job-related decision-making and so on (A3.2). The item is included in the '*societal*' category if the item focuses on community perspectives, such as public transport, government, public policies and so on (A3.3). The item is included in the '*scientific*' category if the item context is related to science and technology, such as the weather, medicine, ecology and so on (A3.4).

Appendix 2. Checklists for Evaluating Item Quality (CEIQs)

Criteria list for short-answer items (B1.1)

Criteria
1.Can the items be answered with a number, symbol, word, or brief phrase? (B1.1.1)
2.Has textbook language been avoided? (B1.1.2)
3.Are the answer blanks equal in length? (B1.1.3)
4. Do the answers blank place at the end of the items? (B1.1.4)
5.Has the degree of precision been indicated for numerical answers? (B1.1.5)
6.Have the units been indicated when numerical answers are expressed in units? (B1.1.6)
7.Have the items been phrased so as to minimize spelling errors? (B1.1.7)
8. Do the items contain any clues? (B1.1.8)

Criteria list for true-false items (B1.2)

Criteria
1.Can each statement be clearly judged true or false? (B1.2.1)
2.Have specific determiners (e.g., usually, always) been avoided? (B1.2.2)
3.Have negative statements (especially double negative) been avoided? (B1.2.3)
4.Have the items been stated in simple, clear language? (B1.2.4)
5.Are the true and false items approximately equal in length? (B1.2.5)
6.Is there an approximately equal number of true and false items? (B1.2.6)
7.Has a detectable pattern of answers (e.g., T, F, T, F) been avoided? (B1.2.7)

Criteria list for matching items (B1.3)

Criteria
1.Is the material in the two lists homogeneous? (B1.3.1)
2. Do the responses rank alphabetically or numerically? (B1.3.2)
3.Do the directions indicate the basis for matching? (B1.3.3)
4. Do the directions specify the number of times each response may be used? (B1.3.4)
5.Is all of each matching item on the same page? (B1.3.5)
6.Is the list of responses longer or shorter than the list of premises? (B1.3.6)

Criteria list for multiple-choice items (B1.4)

Criteria
1. Is each item stem meaningful? (B1.4.1)
2. Do the item stems contain irrelevant material? (B1.4.2)
3. If used, has negative wording been given special emphasis (e.g., capitalized)? (B1.4.3)
4. Are there any grammatical consistency between the alternatives and the item stem? (B1.4.4)
5. Are the alternatives answers brief and free of unnecessary words? (B1.4.5)
6. Do the length and form of the alternatives similar? (B1.4.6)
7. Are the distracters plausible to low achievers? (B1.4.7)
8. Do the items contain any verbal clues to the answer? (B1.4.8)
9. Do the verbal alternatives rank alphabetically? (B1.4.9)
10. Do the numerical alternatives rank numerically? (B1.4.10)
11. Have none of the above and all of the above been avoided? (B1.4.11)

Criteria list for open-ended items (B1.5)

Criteria
1.Is the material to be interpreted appropriate to the students reading level? (B1.5.1)
2. Have pictorial materials been used whenever appropriate? (B1.5.2)
3. Does the material to be interpreted contain some novelty (to require interpretation? (B1.5.3)
4. Are the test items based directly on the introductory material (cannot be answered without it)? (B1.5.4)
5. Are the questions designed to measure higher-level learning outcomes? (B1.5.5)
6. Does each question specify the response expected? (B1.5.6)