International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 368–387
https://doi.org/10.21449/ijate.1376160
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

The difference between estimated and perceived item difficulty: An empirical study

Ayfer Sayın[1*], **Okan Bulut**[2]

[1]Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]University of Alberta, Faculty of Education, Department of Educational Psychology, Alberta, Canada

ARTICLE HISTORY
Received: Oct. 15, 2024
Accepted: May 02, 2024

Keywords:
Item difficulty,
Expert prediction,
Feedback,
Language test,
LGS.

Abstract: Test development is a complicated process that demands examining various factors, one of them being writing items of varying difficulty. It is important to use items of a different range of difficulty to ensure that the test results accurately indicate the test-taker's abilities. Therefore, the factors affecting item difficulty should be defined, and item difficulties should be estimated before testing. This study aims to investigate the factors that affect estimated and perceived item difficulty in the High School Entrance Examination in Türkiye and to improve estimation accuracy by giving feedback to the experts. The study started with estimating item difficulty for 40 items belonging to reading comprehension, grammar, and reasoning based on data. Then, the experts' predictions were compared with the estimated item difficulty and feedback was provided to improve the accuracy of their predictions. The study found that some item features (e.g., length and readability) did not affect the estimated difficulty but affected the experts' item difficulty perceptions. Based on these results, the study concludes that providing feedback to experts can improve the factors affecting their item difficulty estimates. So, it can help improve the quality of future tests and provide feedback to experts to improve their ability to estimate item difficulty accurately.

1. INTRODUCTION

Item difficulty is essential not only for test development but also for creating a large item pool (Bock et al., 1988; Segall et al., 1997), providing items of varying difficulty (Huang et al., 2017), creating equivalent test forms (Förster & Kuhn, 2021; Kolen & Brennan, 2004; Van der Linden & Pashley, 2009), developing adaptive testing (Hontangas et al., 2000; Van der Linden & Pashley, 2009), and establishing Angoff standard setting (Berk, 1986; Dalum et al., 2022). The factors affecting item difficulty are first to be examined to determine item difficulty.

Understanding the factors that affect item difficulty can help test developers have better control over the statistical features of the items they create. This knowledge could also help reduce the need for pre-application, improve test statistics control, such as item difficulty distributions, and enhance test specifications (Bejar, 1983; Boldt, 1998). Therefore, there are many studies examining the factors affecting item difficulty. Some research stated that the item difficulty is affected by the item types (Freedle & Kostin, 1993), item length (Lin et al., 2021), readability

*CONTACT: Ayfer SAYIN ✉ ayfersayin@gazi.edu.tr Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

e-ISSN: 2148-7456

(AlKhuzaey et al., 2021; Lumley et al., 2012), taxonomy (Hamamoto Filho et al., 2020), degree of cognitive complexity (Valencia et al., 2017), visual content (Stiller et al., 2016) and several other variables. Despite this examination, predicting item difficulty remains challenging in educational assessment and empirical attempts to explain low variance (El Masri et al., 2017; Ferrara et al., 2022). Because it is difficult to say that any variable is directly effective on item difficulty in every test. Therefore, there is a need to predict item difficulties before each test is administered.

1.1. Item Difficulty Prediction Methods

Several methods are used to predict item difficulty, including pre-testing, automatic estimation methods, and expert opinion (Attali et al., 2014). A pretest is often very costly and time-consuming and can potentially expose the items to test takers. For automatic estimation, factors affecting item difficulty must be defined, but these factors can vary based on item features, content, and the target population of test takers. The third method of estimating item difficulty is the judgement of subject matter experts (SME), which is often subjective and difficult to scale. However, teachers use their judgment in preparing classroom tests, and some testing centers seek expert opinions when developing achievement tests (e.g., licensure examinations). The Angoff standard-setting method, especially for medical education and high-stake examinations, involves consulting a group of experts in the relevant subject area to establish a standard setting that predicts the difficulty of test items and the overall exam (Benton, 2020; Berk, 1986; Dalum et al., 2022).

The information obtained from SMEs can be evaluated together with the information obtained from other sources and can be used for automatic estimation of item statistics (Attali et al., 2014; Mislevy et al., 1993; Swaminathan et al., 2003). If pilot or field testing cannot be performed, the overall test difficulty is usually adjusted based on the SMEs' judgment of item difficulty (Choi & Moon, 2020). Therefore, experts need to know the factors affecting item difficulty. Especially with the recent increase in cognitive diagnostic assessments, the importance of expert prediction on item content has come to the fore (Liu & Read, 2021). Furthermore, the information obtained from SMEs can be evaluated together with those obtained from other sources and can be used to estimate item statistics (Swaminathan et al., 2003).

Predicting item difficulty is multifaceted and influenced by various factors, including text complexity, decision-making processes, test item intricacies, and the diversity of examinee populations. Studies by Embretson and Wetzel (1987) underscore the importance of incorporating a comprehensive approach to accurately gauge item difficulty, emphasizing text-related variables, as further supported by Freedle and Kostin (1993). The utility of response time data, particularly in naming tasks, was demonstrated by Fergadiotis et al. (2018), highlighting the significance of behavioural metrics. Expertise emerges as a crucial factor, with Berenbon and McHugh (2023) showing that trained item writers excel in predicting item difficulty, contrasting with the challenges highlighted by Sydorenko (2011) and Giguère et al. (2022) regarding the limitations of hypothesized difficulty and the uncorrelated nature of difficulty in Rasch models. The work of Kibble and Johnson (2011) and Herzog et al. (2021) further illustrates the challenges in prediction due to significant individual variation and the limited predictive power of certain item characteristics. Therefore, the endeavour to predict item difficulty is difficult because of a multitude of factors, including the intricacies of text and decision processes, the diversity of test items and populations, and the evolving nature of educational standards and curricula, coupled with the essential roles of response time data, expertise, and individual variability.

Items are still commonly written by experts in high-stakes tests, as well as in-class tests, so the experts who write items must have detailed knowledge about the factors that affect the difficulty of the items. Additionally, improving experts' (teacher, item writer, professor, etc.) ability to

predict how students will perform on assessments and how individual items will perform can help ensure greater consistency in the assessments over time. In other words, understanding the factors affecting item difficulty, such as linguistic and cognitive factors, and why certain items are less predictable can guide the writer and practitioners (Davies, 2021). The present research aims to identify the variables that explain SMEs' prediction of item difficulty and provide feedback to improve their predictions in the High School Entrance Examination language test in Türkiye.

1.2. Comparing the Estimated and Perceived Item Difficulty

Previous studies show that the relationships between estimated and perceived item difficulty depend on variables such as subject matter and profession of the predictors. It also shows that many test-related factors affect the difference between estimated and perceived item difficulty. For instance, Hamamoto Filho et al. (2020) investigated the psychometric properties of items used in a progress test, a longitudinal assessment of students' knowledge. The items were classified according to Bloom's taxonomy, and judges' estimates were used to assess their difficulty. The study was conducted in ten medical schools in Brazil. The study suggests that items with high-level taxonomy may better discriminate against students and that a panel of experts can provide coherent reasoning regarding the item's difficulties. Similarly, Choi and Moon (2020) investigated the factors that impact the difficulty of the reading and listening sections of the English test and found high relation difficulties. The predicted difficulties by both native and non-native speakers were significant predictors of observed difficulty. Le Hebel et al. (2019) focused on exploring the abilities of science teachers in predicting the performance of middle-low achieving students in inquiry-based tasks from the PISA science test. The study utilized a questionnaire-based approach with a sample of 125 French science and technology teachers. The study's findings suggest that the teachers could predict the difficulty levels of inquiry tasks for medium-low achieving students. Additionally, they identified potential sources of difficulty or ease in the tasks. Wauters et al. (2012) compared alternative methods to IRT-based calibration for estimating item difficulty used in adaptive item-based learning environments. The research assessed how well seven different ways of estimating something performed. To do this, the estimates produced by each method were compared to item difficulty that was obtained from a larger study conducted by Selor, which is the selection agency for the Belgian government. The larger study involved 2961 participants who took a test. According to the results, learners are more accurate than experts in predicting the item difficulties. However, this difference disappears when learners and experts are asked to rank the items based on their difficulty. Sydorenko (2011) purposed to investigate the accuracy of item difficulty prediction made by item writers and to examine whether factors affecting item writer judgments corresponded to actual item difficulty predictors. The study used online videos containing conversational dialogues centred on pragmatic functions and was completed by 35 students in their second, third, and fourth years of learning Russian. The outcomes revealed that the predicted item difficulty had a weak but significant association with the estimated item difficulty. The study also discovered that the item writer successfully anticipated linguistic focus and response format but did not consider the influence of topical knowledge.

1.3. Giving SMEs Feedback on Item Difficulty

The results of previous research show that based on understanding the underlying reasons for expert opinions, giving feedback or training to the experts for predicting item difficulty leads to improved prediction accuracy. For example, as part of a project, Davies (2021) explored the ability of examiners and item writers to predict the item difficulty in language tests, focusing on Welsh tests. The study aims to identify the factors affecting item difficulty and understand why certain items are less predictable. The method includes a panel of 13 participants who predicted the difficulty of 320 items on a 5-point scale, followed by a workshop and a second prediction round. The research also investigates whether the workshop training improves

predictions and asks panellists to predict their confidence in their judgments for each item. It found that participants' predictions were correlated with estimated value, and the feedback improved the experts' perceived item. Similarly, González-Brenes et al. (2014) introduced a new method called Feature Aware Student Knowledge Tracing (FAST) that integrates general features into Knowledge Tracing, the standard for inferring student knowledge from performance data. It was determined that teachers' predictions of the difficulty of the tasks improved by 25% with the FAST method they used. Fortus et al. (2013) aimed to identify the factors that affect the difficulty level of multiple-choice items, particularly reading comprehension items, in the English test of Israel's Inter-University Psychometric Entrance Test. The researchers found that the vocabulary and grammatical complexity of the reading comprehension text had the greatest impact on item difficulty. Other variables significantly correlated with difficulty in reading comprehension items include the amount of processing, type of item, length of distractors, and level of vocabulary in stem and distractors. The study also aimed to provide feedback to experts in the context of factors affecting item difficulty, and it found that the correlation between raters' predictions of item difficulty and estimated item difficulty significantly improved from .24 to .82 after giving feedback to the experts. In a similar way, Lumley et al. (2012) discussed the importance of understanding the features that influence the difficulty of reading tasks to improve the reliability of a priori estimates of item difficulty in reading tests. This research developed a schema for describing the difficulty reading items used in PISA. This schema includes 10 variables that can be used by trained raters to predict item difficulty with reasonable success. 5 experts who participated in the study found that raters trained on the schema developed in the research showed better agreement in their predictions. Hambleton et al. (2003) aimed to create and evaluate anchor-based judgmental methods allowing LSAT test specialists to predict item difficulty statistics. The results indicated that even though it needed a long process, the specialists believed they could be trained to predict item difficulty accurately. They demonstrated some proficiency in doing so. After the training, the average error in the predictions of item difficulty ranged from about 11-13%. The panellists found the discussions helpful and were able to improve the prediction of item difficulty. Furthermore, the study discovered that test specialists benefited from the descriptions of items and information about the item statistics of many items in the training. Similarly, MacGregor et al. (2008) stated that participants' prediction of item difficulty improved after feedback; the correlation between estimated and perceived item difficulty was .48 to .65.

1.4. Present Study

Previous studies show that the factors that affect the difficulty of items in different tests differ. They also indicated that several variables affect the accuracy of experts' item difficulty perceptions, and experts can provide valuable information in estimating item difficulty. It reveals that feedback provided to experts improves their item difficulty predictions. Previous research in this field has typically concentrated on examining tests within a single content domain, such as exclusively featuring cloze tests or reading comprehension items. The current study marks a significant departure from this trend by investigating a test encompassing three content domains: grammar, reasoning, and reading comprehension. This holistic approach allows for a more comprehensive analysis of item difficulty, considering the varied cognitive skills required across different test items. This study aims to improve expert estimates of the item difficulty in a language test containing three different content domains (reading comprehension, grammar, and reasoning) in a high-stake test. In addition, this study focuses on a test in the Turkish language. Research has shown that language and cultural factors can significantly influence the difficulty of test items. Oliveri and Ercikan (2011) underscore the pronounced effects of culture and language on test performance, particularly in tasks with significant linguistic demands. Allalouf et al. (1999) highlight the role of translation and cultural congruence in item difficulty, attributing disparities in item difficulty and discrimination to translation inaccuracies and cultural relevance. Further research by Masri et

al. (2016) and Noroozi and Karami (2022) illustrates how acknowledging the influence of language on test takers' perceptions can refine our understanding of item evaluation and difficulty estimation. Gao and Rogers (2010) point to the dynamic interaction between test takers and tasks as a pivotal factor in item difficulty, noting variability across language groups and proficiency levels.

This study distinguishes itself by concurrently examining item characteristics like "readability" and the attributes of both the items and the experts involved in difficulty estimation, thereby contributing a novel perspective to the taxonomy of item difficulty in language testing. In the present study, expert features and item features were also examined together using a multi-faceted Rasch analysis. Since the needs of each expert differ, the effect of feedback on the feedback of individual experts was analyzed. In this case, the present study focused on estimated item difficulty based on the data and perceived item difficulty based on the experts' prediction. It investigated the features that affect estimated and expert item difficulty perceptions and, based on the results, gave feedback to the experts. Therefore, the study aims to provide feedback to experts to improve their item difficulty predictions. This study aims to

i. identify variables that experts use to predict item difficulty,

ii. provide feedback to experts to improve their item difficulty predictions.

The study will contribute to understanding the item difficulty of a high-stakes language test that includes reading comprehension, grammar, and reasoning in the domain. Additionally, the study provided feedback to teachers, professors, and test developers- all item writers-. Accurate item difficulty estimation is crucial for developing valid and reliable assessments that align with learning objectives and provide meaningful feedback to experts and policymakers. The feedback from the data is also expected to guide the item-writing process.

2. METHOD

2.1. Research Design

The research was conducted in a semi-experimental design with the current objective of providing feedback to experts to improve their predictions of item difficulty. Experimental research entails studies to test the impact of variations the researcher creates on the dependent variable. The fundamental aim of experimental designs is to examine the cause-and-effect relationships established among variables. In experimental research, causality between variables is investigated, and changes are observed while controlling variables. Experimental studies seek to elucidate relationships between variables, interpret these relationships, and how outcomes may be influenced based on independent variables (Fraenkel & Wallen, 1990). The study received ethical clearance from the Ethics Committee of Gazi University, bearing the reference number 77082166-604.01.02-711551, dated 02.08.2023.

2.2. Participants

The first stage of the study on item difficulties was estimated based on 20,000 students who attended LGS and took the A booklet. In the second stage, 32 experts predicted the item difficulty, and in the third stage, 24 experts who had at least 3 correct predictions were selected and were given individual feedback to them. The same 24 experts predicted item difficulty again in the 4th stage of the study. Table 1 shows some information about the participants of the research.

Table 1. *Participants.*

Stage1 Estimated item difficulty			Stage2 First prediction			Stage 3 Feedback			Stage 4 Second prediction		
Characteristic	*f*	%	Characteristic	*f*	%	Characteristic	*f*	%	Characteristic	*f*	%
Students			Experts			Experts			Experts		
Test year			*Gender*			*Gender*			*Gender*		
2018	10,000	50.0	Female	18	56.3	Female	13	54.2	Female	13	54.2
2019	10,000	50.0	Male	14	43.8	Male	11	45.8	Male	11	45.8
Gender			*Years of experience*			*Years of experience*			*Years of experience*		
Female	9,913	49.6	<1 year	3	9.4	<1 year	3	12.5	< 1 year	3	12.5
Male	10,087	50.4	1-5 years	8	25.0	1-5 year	5	20.8	1-5 year	5	20.8
School Type			5-10 years	13	40.6	5-10 year	11	45.8	5-10 year	11	45.8
Public	18,366	91.8	10+ years	8	25.0	10+years	5	20.8	10+years	5	20.8
Private	1,634	8.2	*Profession*			*Profession*			*Profession*		
			Professor	13	40.6	Professor	8	33.3	Professor	8	33.3
			Teacher	10	31.3	Teacher	8	33.3	Teacher	8	33.3
			Test developer	9	28.1	Test developer	8	33.3	Test developer	8	33.3
Total	20,000	100	Total	32	100	Total	24	100	Total	24	100

2.3. Process

This study was carried out in four stages as an experimental design. In the first stage, item difficulties were estimated for 40 items based on the data in the High School Entrance Examination (known as LGS) in Türkiye, and item features that affect item difficulty were determined. In the second stage, 6 items were determined from the 40 items with different item features. Items were selected based on different content domains (reading comprehension, reasoning and grammar), some of which include visual and some non-visual content. While some items are very long, some are short; some are easy, and some are moderate or hard. In this stage, 32 experts predicted the difficulty of the same 6 items. The factors that affect experts' item difficulty predictions were studied and the experts' predictions were compared with the actual item difficulty in the second stage. In the third stage, experts who had at least 3 correct predictions were determined and gave individual feedback to experts based on the results. In the fourth stage, 24 experts predicted 34 items' difficulty on a 5-point scale in a nested way. It means that in this stage, experts predicted the item difficulty of 6 items and did not see all items. Each item was predicted by at least 3 experts. After that, the factors that affect experts' item difficulty predictions were identified, and the experts' perceptions were compared with the estimated item difficulty again.

2.4. Predictors

In the current study, certain variables that contribute to the estimation of item difficulty within the Turkish test were analyzed. This analysis encompassed several item characteristics, including item length (word count), readability, visual content, content domain, and question prompt for the item features. Additionally, attributes of the raters themselves were considered to explore factors influencing SMEs' predictions of item difficulty in the Turkish test. Specifically, the analysis took into account the gender of the raters, their years of experience in test development, and their professional backgrounds. The findings about these features are delineated in Table 2 and Table 3. The features of items of visual content, question prompt, and content domain were scrutinized based on the assessments of two experts with backgrounds in Turkish language education and item writing. These experts independently identified the attributes of the items, and their findings were subsequently synthesized for analysis. The

textual properties of the items, including item length and readability, were calculated utilizing Python software. Determining item length involved computing the word count, while readability was assessed by implementing the Ateşman (1999) formula.

Table 2. *Item features to determine affecting estimated and perceived item difficulty.*

Item Features	Min	Max	*M*	*S*	*n*	%
Visual content						
Yes					7	17.5
No					33	82.5
Question prompt						
Positive phrased					31	77.5
Negative phrased					9	22.5
Content domain						
Reading comprehension					24	60.0
Grammar					10	25.0
Reasoning					6	15.0
Textual features						
Item length (word count)	24.0	416.0	113.5	77.8		
Readability	36.2	84.9	62.0	11.7		

Table 3. *Rater features to determine affecting perceived item difficulty.*

Rater Features	1st prediction		2nd prediction	
	n	%	*n*	%
Gender				
Female	18	56.3	13	54.2
Male	14	43.8	11	45.8
Years of experience				
<1 year	3	9.4	3	12.5
1-5 years	8	25.0	5	20.8
5-10 years	13	40.6	11	45.8
10+ years	8	25.0	5	20.8
Profession				
Professor	13	40.6	8	33.3
Teacher	10	31.3	8	33.3
Test developer	9	28.1	8	33.3

2.5. Feedback Process

In the second stage, 24 experts provided feedback on the difficulty of the items. For this, an instructor group was established. It consisted of three professors, two of them working in the field of Turkish education and one of them working in measurement and evaluation at the university. While preparing the feedback, the factors affecting the difficulty of the 6 items in the first stage were determined in detail by the instructor group. Based on the first stage results, they examined the purpose of the items, the formal and content features, and the order of the options together. Then, the accuracy and inaccuracy of the experts' predictions in the first stage

were deduced, and the tutorial group conducted online interviews with each expert individually. The feedback was presented personally by comparing the factors that the experts paid attention to during the prediction process with the actual item statistics.

2.6. Data Collection Tool

The data collection process comprised four sequential stages within an experimental design framework. In the initial phase, item difficulty estimates were derived for 40 items based on the High School Entrance Examination data in Türkiye, with concurrent identification of item features influencing difficulty levels. The annual exam by the Ministry of National Education serves as a pivotal placement test for approximately 1 million students seeking admission to high schools. Subsequently, six items were selected from the initial pool, each characterized by distinct features such as content domain (e.g., reading comprehension, reasoning, grammar), visual or non-visual elements, varying lengths, and differing difficulty levels. Data were collected from the experts using an item difficulty estimation form. The form included the items and the item difficulty that the expert could mark the answer next to each item. Expert predictions of item difficulty on a 5-point scale (1=very difficult to 5=very easy) were obtained for these six items in the second stage, involving 32 experts. The third stage involved providing individual feedback to experts who demonstrated at least three correct predictions. Finally, in the fourth stage, 24 experts, following a nested design, predicted the difficulty of 34 items on a 5-point scale, with each item assessed by at least three experts.

2.7. Analysis

In the first stage, based on the answers of 10,000 students who participated in LGS in 2018 and 2019 and received booklet A, item difficulty was estimated based on the CTT for 40 items. Then, hierarchical regression analysis was performed to determine the features affecting the difficulty of the items by using the item length (word count), readability, visual content, content domain, and question prompt as predictors. In the second stage, the features affecting the item difficulty predictions of 32 experts were analyzed by multi-faceted Rasch analysis. Multi-faceted Rasch analysis is a statistical method used to examine the influence of different factors, such as experts and items, on expert predictions. This analysis provides individual and group-level statistics on a single comparable scale, the logit scale. The logit scale allows for meaningful comparisons and interpretations of the estimates (Myford & Wolfe, 2003). This study performed analyses using the Minifac (Facets) Rasch software program. The analysis included 6 item facets (item difficulty, item visual, question prompt, content domain, item length and item readability), and 4 rater facets (experts, experts' gender, profession, and year of experience). In the third stage, during the feedback process, the points that the experts paid attention to while predicting the difficulty of the items were determined and compared with the estimation of the items. In the fourth and final stage, 24 experts predicted the item difficulty of the remaining 34 items in the tests. In line with the experts' prediction, the difficulty of the 34 items was analyzed. In the multifaceted Rasch analysis, a 10-facet crossed design was used as items (6) x expert (24) x gender (2) x profession (3) x years of experience (4) x item visual (2) x question prompt (2) x content domain (3) x item length (2) x readability (2). In the analysis after the second prediction, predictions were similarly made based on the 10-factor crossed design. The model in the second prediction is as follows: items (34) x expert (24) x gender (2) x profession (3) x years of experience (4) x item visual (2) x question prompt (2) x content domain (3) x item length (2) x readability (2). The Spearman correlation coefficient was estimated to examine the relationship between the estimated and perceived values.

3. FINDINGS

3.1. Estimated Item Difficulty

After the estimated item difficulties, the average difficulty of the Turkish items in 2018 was estimated as 0.63. The item difficulties varied between 0.23 and 0.91. In 2019, the item difficulties varied between 0.34 and 0.75; the average difficulty was 0.59. Hierarchical regression analysis was performed to determine the extent to which item features explained the item's difficulties, and the results are shown in Table 4. As a result of the analysis, it was determined that 27% was explained by only the content domain feature. It was found that there is positive and moderate relationship between reading comprehension items and item difficulty (β=0.519; p<0.01). It shows that reading comprehension items are easier than grammar and reasoning items. However, it was determined that the item length (word count), readability, visual content and question prompt do not have a direct effect on the item difficulty (p>0.01).

Table 4. *Results of the regression analysis.*

Model		Unstandardized coefficients		Standardized coefficients		
		B	Std. Error	Beta	*t*	*p*
1	(Constant)	.519	.031		16.679	.000
	Cognitive_domain	.150	.040	.519	3.746	.001
2	(Constant)	.513	.139		3.689	.001
	Cognitive_domain	.145	.046	.502	3.184	.003
	Length	2.458E-5	.000	.013	.073	.942
	Readability	.000	.002	.015	.095	.925
	Visual_content	-.023	.077	-.061	-.295	.770
	Question_prompt	-.002	.051	-.007	-.044	.965

a. Dependent Variable: Item difficulty

3.2. First Round of Item Difficulty Prediction

In the second stage of the study, 32 experts predicted the difficulty of 6 items. The item difficulty predictions of the experts were analyzed by multi-faceted Rasch analysis with the experts and the items' features. All facet vertical rules are shown in Appendix 1, and the measurement report is shown in Table 5.

When Appendix 1 was examined, the experts indicated that the most difficult item was the 6th, and the easiest item was the 1st. It is seen that the experts' predictions of the difficulty/ease of the items were significantly divided into two categories approximately (reliability=0.70; strata=2.35; χ^2=16.1; p<0.05). It is also seen that R27 is the most generous (predicting that the items are easier), while R18 and R26 are the most rigid (predicting that the items are more difficult) experts. However, it was determined that the item difficulty predictions did not differ significantly in terms of strictness/generosity (reliability=0.27; χ^2=41.2; p>0.05). It was also determined that the item difficulty predictions of the experts did not differ significantly according to their gender (reliability=0.00; χ^2=0.4; p>0.05), profession (reliability=0.00; χ^2=0.7; p>0.05), and years of experiment (reliability=0.34; χ^2=6.4; p>0.05). When the item difficulty predictions were analyzed according to the item features, the experts tended to predict items with visual text more difficult than those with nonvisual text (the discrimination reliability values are high (>0.70) for the discrimination ratio (separation=1.74) and the discrimination index (strata=2.65); χ^2=4.0; p<0.05). Experts' item difficulty predictions also varied according to the length (number of words) of the item, and experts predicted items with more than 150 words to be more difficult than items with less than 150 words (reliability=0.71; χ^2=3.4;

p<0.05). However, it was determined that the predictions did not show a significant difference according to the positive-negative question prompt (reliability=0.00; χ^2=0.0; p>0.05), content domain (reading comprehension, grammar, reasoning) (reliability=0.00; χ^2=0.8; p>0.05) and readability (reliability=0.00; χ^2=0.0; p>0.05).

Table 5. *Measurement report of the first prediction.*

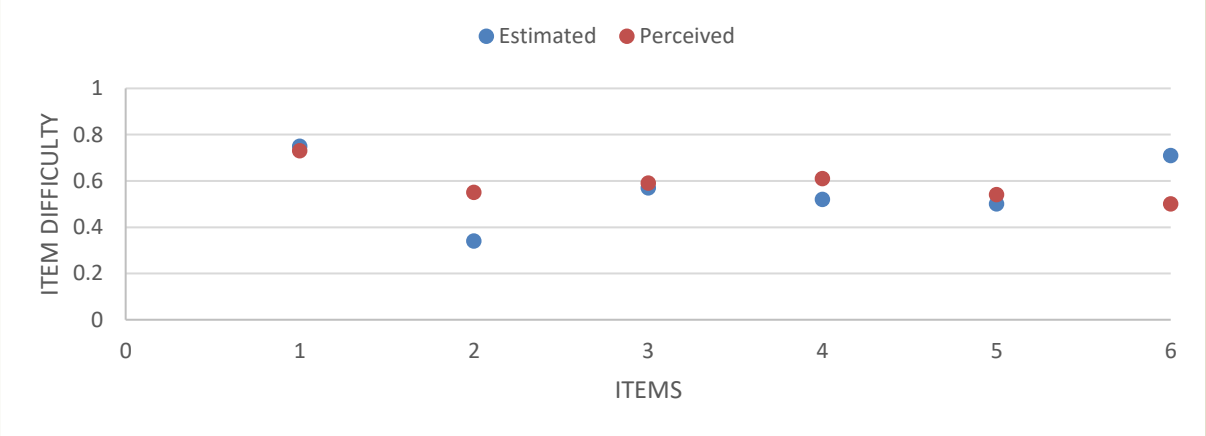
Model Sample	Items*	Raters	Rater Features			Item Features				
			Gender	Profession	Years of experience	Visual*	Question Prompt	Content Domain	Length*	Readability
RMSE	0.26	0.60	0.15	0.19	0.24	0.15	0.16	0.22	0.15	0.20
Adj (True) S.D.	0.39	0.37	0.00	0.00	0.18	0.26	0.00	0.00	0.23	0.00
Separation	1.51	0.61	0.00	0.00	0.72	1.74	0.00	0.00	1.56	0.00
Strata	2.35	1.15	0.33	0.33	1.30	2.65	0.33	0.33	2.41	0.33
Reliability	0.70	0.27	0.00	0.00	0.34	0.75	0.00	0.00	0.71	0.00
X^2	16.1	41.2	0.4	1.1	6.4	4.0	0.0	1.4	3.4	0.0
(sig.)	(0.01)	(0.10)	(0.51)	(0.58)	(0.09)	(0.04)	(0.93)	(0.50)	(0.04)	(0.93)

* Separated variables

3.3. Second Round of Item Difficulty Prediction

In the second round, 24 experts predicted the difficulty of the remaining 34 items. For this purpose, tests consisting of 6 items were prepared for the experts. For example, R4 predicted the difficulty of items 1, 5, 9, 12, 14, and 15, while R12 predicted the difficulty of items 4, 9, 23, 24, 26, and 27. In other words, a nested method was followed, not a cross method. So, each expert predicted the difficulty of 6 items, and at least 3 experts examined one item. It is shown in Figure 1.

Figure 1. *Compare the estimated and perceived item difficulty of the first prediction.*



The item difficulty predictions were analyzed using a multi-faceted Rasch analysis with the experts and item features. All facet vertical rules are shown in Appendix 2, and the measurement report is in Table 6. As a result of analyses, the raters indicated that the most difficult item was the 29th, and the easiest item was the 15th. When the item measurements are examined, the discrimination reliability values are high (>0.70) for the discrimination ratio (separation=1.88) and the discrimination index (strata=2.84). Accordingly, it is seen that the experts significantly categorized the difficulty/ease predictions of the items into approximately three categories (χ^2=189.3; p<0.05). When the estimated values are also examined, the tests do not have very easy and very difficult items. Therefore, it can be said that the experts' item difficulty predictions are similar to the estimates. It is seen that R4 is the most generous (predicting that the items are easier), while R19 and R12 are the strictest (predicting that the items are more difficult) experts. It was determined that the experts' predictions differed significantly in terms of strictness/generosity (reliability=0.76; χ^2=41.2; p>0.05). This is likely because the experts predicted 34 items using a nested method during the second prediction

process. The second predictions did not differ significantly according to their gender (reliability=0.00; χ^2=0.8; p>0.05), profession (reliability=0.00; χ^2=0.6; p>0.05) and seniority (reliability=0.19; χ^2=4.1; p>0.05). When the predictions were analyzed according to the item features, it was determined that the item difficulty predictions varied according to the content domain (reliability=0.86; strata=3.57; χ^2=21.2; p<0.05). Accordingly, the experts stated that the most difficult items belonged to the grammar content domain, followed by the reasoning content domain. They stated that the reading comprehension items were easier than the items in the other content domain. Experts' item difficulty predictions also varied according to the length (number of words) of the item, with more than 150 words being more difficult than items with fewer than 150 words (reliability=0.84; strata=3.44; χ^2=6.4; p<0.05). Experts' predictions were also affected by the readability; as the readability of the items increased, experts tended to evaluate the items more difficult (reliability=0.85; strata=3.51; χ^2=12.0; p<0.05). However, it was determined that the item difficulty predictions did not show a significant difference according to the visual content (reliability=0.06; χ^2=1.1; p>0.05) and positive-negative question prompt (reliability=0.52; χ^2=2.1; p>0.05).

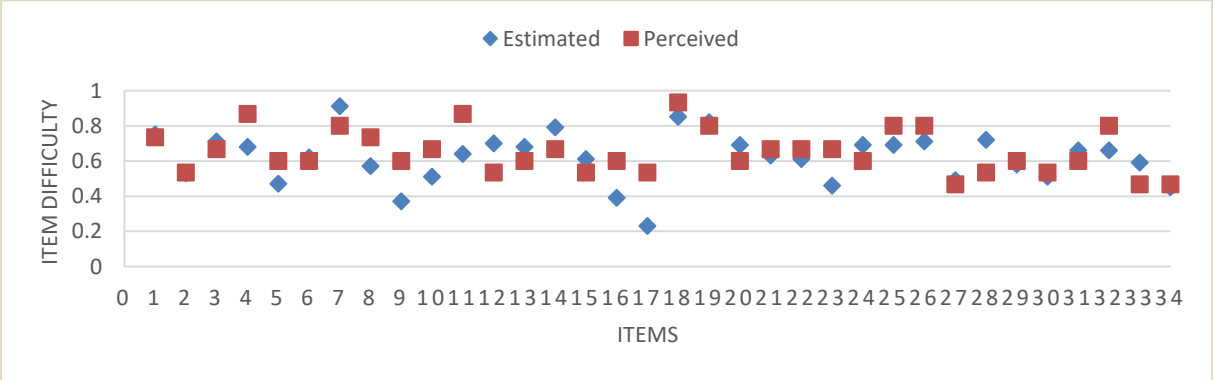
In the fourth stage of the study, after giving feedback to the experts, it was also found a positive and moderate correlation between the estimated and perceived item difficulty (r=0.410; p<0.01). It was observed that the experts tended to predict the items as easily as they were (Figure 4).

Table 6. *Measurement report of the second prediction.*

Model Sample	Rater Features				Item Features					
	Items*	Raters*	Gender	Profession	Years of experience	Visual	Question Prompt	Content Domain*	Length*	Readability*
RMSE	1.79	1.65	.41	.47	.62	.54	.48	.54	.44	.55
Adj (True) S.D.	3.37	2.97	.00	.00	.30	.13	.50	1.31	1.03	1.30
Separation	1.88	1.80	.00	.00	.48	.24	1.04	2.43	2.33	2.38
Strata	2.84	2.74	.33	.33	.98	.66	1.72	3.57	3.44	3.51
Reliability	0.78	0.76	.00	.00	.19	.06	.52	.86	.84	.85
X^2	189.3	98.7	.8	.6	4.1	1.1	2.1	21.2	6.4	12.0
(sig.)	(.00)	(.00)	(.36)	(.73)	(.25)	(.30)	(.15)	(.00)	(.01)	(.00)

* Separated variables

Figure 2. *Compare the estimated and perceived item difficulty of the second prediction.*



r=0.410; p<0.01

4. DISCUSSION and CONCLUSION

This study aimed to increase the accuracy of experts' item difficulty estimates by focusing on estimated item difficulty based on data and perceived item difficulty based on expert estimates. All results of the current study are summarized in Table 7.

Table 7. *Summary of the results.*

Turkish test	Estimated item difficulty	1st prediction	2nd prediction
LGS 2018	0.34 - 0.75	√	
LGS 2019	0.23 - 0.91		√
Rater features			
Rater (strictness/generosity)	---	x	√
Gender	---	x	x
Profession	---	x	x
Years of experience	---	x	x
Item features			
Visual content	x	√	x
Question prompt	x	x	x
Content domain	√	x	√
Item length	x	√	√
Readability	x	x	√

It was determined that 27% of the difficulty of the items in the Turkish test was significantly explained by only the content domain features. Although a limited number of studies have established models that explain a significant portion of the variation in item difficulty (53.5% (Sung et al., 2015), research showed that a significant variance in item difficulty is not explained by the models. For instance, despite identifying many explanatory predictors, they explained 23% of the variance in item difficulty in a science test (El Masri et al., 2017). The difficulty of 214 reading and listening comprehension items was modeled as a function of 12 predictor variables with item and text interaction. Seven of the 12 variables in the model explained approximately 31% of the variance in item difficulty (Rupp et al., 2001). In another study examining how task features affect item difficulty in EFL listening tests, regression analyses were conducted by using 20 predictors. As a result of the research, it was determined that item features explained 31.6% of the difficulty. (Ying-hui, 2006). The reason why a significant portion of the item difficulties were not explained may be that the difficulty varies according to the field, language, purpose, item types and other different structures of the test (Sydorenko, 2011).

The present study found that the reading comprehension items were easier than the grammar and reasoning items. The results also showed that the length of the items (word count), readability, visual or non-visual content, and positive or negative phrasing did not directly affect the item difficulty. Some research showed that longer items (i.e. length of distractors, item length) could be more difficult because they required more cognitive effort to process and comprehend (Fortus et al., 2013; Freedle & Kostin, 1993; Gorin & Embretson, 2006; Lin et al., 2021; Stenner, 2022; Stiller et al., 2016; Trace et al., 2017), and some studies also indicated that as the readability of items increases, their difficulty also increased (AlKhuzaey et al., 2021; Choi & Moon, 2020; Toyama, 2021). However, similar to the present research, some studies

found that item length or readability might not always affect item difficulty directly (Aljehani et al., 2020). In this case, it is important to consider the specific context in which item length and readability are being considered. For example, in a language test where the primary goal is to assess reading comprehension skills, longer passages may be easier, even if they need more time to read, because they provide more information, and it might be easy to find the main idea or other indicators. The test also included grammar, reading comprehension and reasoning items in this research. Although grammar items were the shortest in the test, reading comprehension items were the easiest. For all these reasons, although experts thought length and readability are affected, the textual features (length and readability) examined in the study may not have effectively affected the item difficulty. In general, visual content can affect item difficulty by either aiding or hindering the test-taker's ability to comprehend the item. For example, if a test item includes a visual aid that effectively illustrates the content of the item, it may be easier for test-takers to understand the item and answer the item correctly. Conversely, if the visual aid is confusing or it is necessary to read the information in the visual and compare it with the information in the text and reach an inference, it may make the item more difficult for test-takers to understand and respond correctly (Santi et al., 2015; Stiller et al., 2016). In this study, it was determined that the visual content did not directly affect the difficulty of the item. The students had enough time to solve the items, the visual items were carefully designed in the item writing, the visual content was clearly expressed, the visuals were designed by the level of the students, and the students were familiar with the items in the visual content. Question prompting, another variable examined in this study, can also affect item difficulty. Research showed that negatively worded items can be more difficult than positive ones for test-takers to understand and answer correctly compared to positively worded items (Haladyna et al., 2002). However, some studies found that visual content or question prompts might not affect item difficulty (Caldwell & Pate, 2013). This study, conducted on a Turkish test, found that question prompts did not directly affect item difficulty. However, as with item length, it is important to consider other factors that may have influenced this finding. The findings that reading comprehension items were easier than grammar and reasoning items may indicate that students encounter greater challenges with grammar and reasoning items, which likely demand higher cognitive efforts. Reading comprehension items, relying on the ability to understand and interpret text, may enable students to locate answers more easily using information that is directly related to and retrievable from the text. In contrast, grammar and reasoning items might require more complex cognitive processes such as abstract thinking, knowledge of rules, and problem-solving skills. The result that the length of items (word count), readability, presence of visual or non-visual content, and the use of positive or negative phrasing did not directly impact item difficulty suggests the complexity of factors determining item difficulty, indicating that these features alone may not significantly influence the challenge level of an item. This implies that other variables, such as the cognitive abilities of the students being tested, their pre-existing knowledge, and their familiarity with the text or type of items, might be more determinative in influencing item difficulty. Although some research indicates that longer items might be more challenging due to the increased cognitive effort required to process and understand them, the findings of this study could suggest that students may have developed strategies to manage these lengths and remain unaffected in their question-solving process. Moreover, features like readability and visual content may not significantly affect item difficulty if they contain information that students are already familiar with or can easily understand.

In the first prediction, while the visual content in the items affected the experts' prediction, it did not affect the second estimation. This is consistent with the real situation. While the content domain of the items did not affect the experts' predictions in the first prediction, it did in the second one. Experts stated that reading comprehension items were easier. This is exactly consistent with the estimated situation. The length of the items was effective in both predictions

of the experts. The readability of the items was also effective in the experts' second prediction. The changes, which impact the experts' item difficulty predictions, are consistent with the estimates. In other words, there has been an improvement in the factors affecting the experts' predictions in line with the feedback given to the experts. A positive and moderate correlation was also found between the experts' perceptions and the estimated item difficulty (r=410; p<0.01). This finding is generally consistent with the results in the literature. For example, a study by Choi and Moon (2020) determined that the experts' prediction and estimated item difficulty were moderately or highly correlated in the reading comprehension items. Le Hebel et al. (2019) found that teachers could identify relevant potential sources of difficulty or easiness in the items that come from the PISA science test. Similarly, Attali et al. (2014) discovered that judges could accurately rank various items according to their difficulty level, and this trend remained consistent across multiple judges and subject areas in the SAT. Impara and Plake (1998) also stated that experts could predict item difficulty with 54% accuracy. Some research also showed that experts predict item difficulties significantly (Enright et al., 1993; Hamamoto Filho et al., 2020; Wauters et al., 2012), whereas some research showed the opposite of these results. For example, Sydorenko (2011) found a low correlation (r = .30) between the estimated and perceived difficulty, which could be due to the item writer not taking into account the difficulty of the topic and the similarity of intermediate and advanced items (Sydorenko, 2011). Kibble and Johnson (2011) stated that there is a significant but relatively low correlation between the perceived and estimated item difficulty in multiple-choice items (r=-0.19; p<0.01). Therefore, research suggests that experts should be aware of their potential biases and take steps to mitigate them, such as seeking feedback. In this research, it was found that there was an improvement in item difficulty prediction after giving feedback to the experts. It was consistent with research results that feedback or training on item difficulty improves experts' predictions (Davies, 2021; Fortus et al., 2013; González-Brenes et al., 2014; Hambleton & Jirka, 2011; Lumley et al., 2012; MacGregor et al., 2008).

In this study, it was also observed that the experts tended to predict the items as easily as they were. Urhahne and Wijnia (2021) reviewed 10 studies that examined the correlation between teachers' perceptions of task difficulty and the actual difficulty of those tasks with meta-analysis. The review found that in 8 out of the 10 studies, teachers tended to underestimate the level of challenge posed by the tasks or overestimate the expected performance of their students.

4.1. Limitation and Future Research

The study focuses on the High School Entrance Examination in Türkiye, which limits the generalizability of the findings to other contexts or examinations. Furthermore, 40 items can also be considered relatively small, potentially affecting the representativeness of the findings. In addition, the study primarily examines the factors that influence experts' item difficulty predictions and does not consider other potential sources of variability, such as test-taker characteristics. Based on the outcomes of this research, the practical implications for test developers, item writers, and educational practitioners are substantial and can significantly enhance the development and evaluation process of test items. The improvement in experts' predictions of item difficulty following feedback underscores the value of continuous training and development for item writers. Implementing feedback mechanisms and training programs that focus on the nuanced aspects of item design, such as the influence of visual content, content domain, item length, and readability on item difficulty, can empower item writers to make more accurate predictions. Similarly, the fluctuating impact of the content domain on expert predictions across different estimations highlights the importance of iterative review processes in accounting for various factors that may influence item difficulty. Furthermore, the findings suggest that training programs for item writers should cover the technical aspects of item construction and include modules on cognitive psychology and how test-takers interact with different item types. Such comprehensive training can enhance item writers' awareness of their potential biases and improve their ability to predict item difficulty accurately. In other words,

the results may also serve as a source of guidance for item writers. It highlights the importance of validating expert judgments and using multiple sources of information when assessing item difficulty or other constructs in research.

Acknowledgments

This study was presented as an oral presentation at the NCME (National Council on Measurement in Education) conference held in Chicago, USA, from May 12-15, 2023.

The authors would like to thank the blind reviewers and SMEs for their useful comments and insightful suggestions. We also thank the Ministry of National Education (MEB) for providing the data necessary for this research.

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. **Ethics Committee Number**: Gazi University Ethics Committee, E77082166-604.01.02-711551, dated 02.08.2023.

Contribution of Authors

Ayfer Sayın: Design, Data Collection and/or Processing, Materials, Analysis and Interpretation, Literature Review, Writing. **Okan Bulut**: Conception, Design, Supervision, Writing, Critical Review.

Orcid

Ayfer Sayın https://orcid.org/0000-0003-1357-5674
Okan Bulut https://orcid.org/0000-0001-5853-1267

REFERENCES

Aljehani, D.K., Pullishery, F., Osman, O., & Abuzenada, B.M. (2020). Relationship of text length of multiple-choice questions on item psychometric properties–A retrospective study. *Saudi J Health Sci*, *9*, 84-87. https://doi.org/10.4103/sjhs.sjhs_76_20

AlKhuzaey, S., Grasso, F., Payne, T.R., & Tamma, V. (2021). A Systematic Review of Data-Driven Approaches to Item Difficulty Prediction. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova, *Artificial Intelligence in Education* Cham. https://doi.org/10.1007/978-3-030-78292-4_3

Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of dif in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198. https://doi.org/10.1111/j.1745-3984.1999.tb00553.x

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series*, *2014*(2), 1-8. https://doi.org/10.1002/ets2.12042

Bejar, I.I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*(3), 303-310. https://doi.org/10.1002/j.2333-8504.1981.tb01274.x

Benton, T. (2020). How Useful Is Comparative Judgement of Item Difficulty for Standard Maintaining? *Research Matters*, *29*, 27-35.

Berenbon, R., & McHugh, B. (2023). Do subject matter experts' judgments of multiple-choice format suitability predict item *quality*?. *Educational Measurement Issues and Practice, 42*(3), 13-21. https://doi.org/10.1111/emip.12570

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*(1), 137-172. https://doi.org/10.3102/00346543056001137

Bock, R.D., Murakl, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*(4), 275-285. https://doi.org/www.jstor.org/stable/1434961

Boldt, R.F. (1998). GRE analytical reasoning item statistics prediction study. *ETS Research Report Series*, *1998*(2), i-23. https://doi.org/10.1002/j.2333-8504.1998.tb01786.x

Caldwell, D.J., & Pate, A.N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, *77*(4). https://doi.org/10.5688/ajpe77471

Choi, I.-C., & Moon, Y. (2020). Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, *17*(1), 18-42. https://doi.org/10.1080/15434303.2019.1674315

Dalum, J., Christidis, N., Myrberg, I.H., Karlgren, K., Leanderson, C., & Englund, G.S. (2022). Are we passing the acceptable? Standard setting of theoretical proficiency tests for foreign-trained dentists. *European Journal of Dental Education*. https://doi.org/10.1111/eje.12851

Davies, E. (2021). Predicting item difficulty in the assessment of Welsh. Collated Papers for the ALTE 7th International Conference, Madrid, Spain.

El Masri, Y.H., Ferrara, S., Foltz, P.W., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal*, *28*(1), 59-82. https://doi.org/10.1080/09585176.2016.1232201

Embretson, S., & Wetzel, C. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*(2), 175-193. https://doi.org/10.1177/014662168701100207

Enright, M.K., Allen, N., & Kim, M.I. (1993). A Complexity Analysis of Items from a Survey of Academic Achievement in the Life Sciences. *ETS Research Report Series*, *1993*(1), i-32. https://doi.org/10.1002/j.2333-8504.1993.tb01529.x

Fergadiotis, G., Swiderski, A., & Hula, W. (2018). Predicting confrontation naming item difficulty. *Aphasiology, 33*(6), 689-709. https://doi.org/10.1080/02687038.2018.1495310

Ferrara, S., Steedle, J.T., & Frantz, R.S. (2022). Response Demands of Reading Comprehension Test Items: A Review of Item Difficulty Modeling Studies. *Applied Measurement in Education*, *35*(3), 237-253. https://doi.org/10.1080/08957347.2022.2103135

Förster, N., & Kuhn, J.-T. (2021). Ice is hot and water is dry: Developing equivalent reading tests using rule-based item design. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000691

Fortus, R., Coriat, R., & Fund, S. (2013). Prediction of item difficulty in the English Subtest of Israel's Inter-university psychometric entrance test. In *Validation in language assessment* (pp. 61-87). Routledge.

Fraenkel, J.R. & Wallen, dan Norman E. (2006). *How to Design and Evaluate Research in Education.* McGraw-Hill Education, USA.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. *ETS Research Report Series*, *1993*(1), i-48. https://doi.org/10.1002/j.2333-8504.1993.tb01524.x

Gao, L., & Rogers, W. (2010). Use of tree-based regression in the analyses of l2 reading test items. *Language Testing, 28*(1), 77-104. https://doi.org/10.1177/0265532210364380

Giguère, G., Brouillette-Alarie, S., & Bourassa, C. (2022). A look at the difficulty and predictive validity of ls/cmi items with rasch modeling. *Criminal Justice and Behavior, 50*(1), 118-138. https://doi.org/10.1177/00938548221131956

González-Brenes, J., Huang, Y., & Brusilovsky, P. (2014). General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. The 7th international conference on educational data mining (pp. 84–91), London. https://doi.org/pdfs.semanticscholar.org/0002/fab1c9f0904105312031cdc18dce358863a6.pdf

Gorin, J.S., & Embretson, S. E. (2006). Item diffficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394-411. https://doi.org/10.1177/0146621606288554

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Hamamoto Filho, P.T., Silva, E., Ribeiro, Z.M.T., Hafner, M.d.L.M.B., Cecilio-Fernandes, D., & Bicudo, A.M. (2020). Relationships between Bloom's taxonomy, judges' estimation of item difficulty and psychometric properties of items from a progress test: a prospective observational study. *Sao Paulo Medical Journal*, *138*, 33-39. https://doi.org/10.1590/1516-3180.2019.0459.R1.19112019

Hambleton, R.K., & Jirka, S.J. (2011). Anchor-based methods for judgmentally estimating item statistics. In *Handbook of test development* (pp. 413-434). Routledge.

Hambleton, R.K., Sireci, S.G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). *Anchor-Based Methods for Judgmentally Estimating Item Difficulty Parameters.* LSAC Research Report Series, Newtown, PA.

Herzog, M., Sari, M., Olkun, S., & Fritz, A. (2021). Validation of a model of sustainable place value understanding in Turkey. *International Electronic Journal of Mathematics Education, 16*(3), em0659. https://doi.org/10.29333/iejme/11295

Hontangas, P., Ponsoda, V., Olea, J., & Wise, S.L. (2000). The choice of item difficulty in self-adapted testing. *European Journal of Psychological Assessment*, *16*(1), 3. https://doi.org/10.1027/1015-5759.16.1.3

Hsu, F.-Y., Lee, H.-M., Chang, T.-H., & Sung, Y.-T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, *54*(6), 969-984. https://doi.org/10.1016/j.ipm.2018.06.007

Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., & Hu, G. (2017). Question Difficulty Prediction for READING Problems in Standard Tests. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10740

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*(1), 69-81. https://doi.org/10.1111/j.1745-3984.1998.tb00528.x

Kibble, J.D., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in physiology education*, *35*(4), 396-401. https://doi.org/10.1152/advan.00062.2011

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking methods and practices*. Springer New York, NY. https://doi.org/10.1007/978-1-4939-0317-7

Le Hebel, F., Tiberghien, A., Montpied, P., & Fontanieu, V. (2019). Teacher prediction of student difficulties while solving a science inquiry task: example of PISA science items. *International Journal of Science Education*, *41*(11), 1517-1540. https://doi.org/10.1080/09500693.2019.1615150

Lin, C.-S., Lu, Y.-L., & Lien, C.-J. (2021). Association between Test Item's Length, Difficulty, and Students' Perceptions: Machine Learning in Schools' Term Examinations. *Universal Journal of Educational Research*, *9*(6), 1323-1332. https://doi.org/10.13189/ujer.2021.090622

Liu, X., & Read, J. (2021). Investigating the Skills Involved in Reading Test Tasks through Expert Judgement and Verbal Protocol Analysis: Convergence and Divergence between the Two Methods. *Language Assessment Quarterly*, *18*(4), 357-381. https://doi.org/10.1080/15434303.2021.1881964

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012). *A framework for predicting item difficulty in reading tests* Proceedings of the annual meeting of the American educational research association (AERA), Vancouver, BC, Canada.

MacGregor, D., Kenyon, D., Christenson, J., & Louguit, M. (2008). Predicting item difficulty: A rubrics-based approach. *American Association of Applied Linguistics. March, Washington, DC*. https://doi.org/10.1109/FIE.2015.7344299

Masri, Y., Baird, J., & Graesser, A. (2016). Language effects in international testing: the case of pisa 2006 *science* items. *Assessment in Education Principles Policy and Practice, 23*(4), 427-455. https://doi.org/10.1080/0969594x.2016.1218323

Mislevy, R.J., Sheehan, K.M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*(1), 55-78. https://doi.org/www.jstor.org/stable/1435164

Noroozi, S., & Karami, H. (2022). A scrutiny of the relationship between cognitive load and difficulty estimates of language test items. *Language Testing in Asia, 12*(1). https://doi.org/10.1186/s40468-022-00163-8

Oliveri, M., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions?. *Applied Measurement in Education, 24*(4), 349-366. https://doi.org/10.1080/08957347.2011.607063

Rupp, A.A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, *1*(3-4), 185-216. https://doi.org/10.1080/15305058.2001.9669470

Sano, M. (2015). Automated capturing of psycho-linguistic features in reading assessment text. Annual meeting of the National Council on Measurement in Education, , Chicago, IL, USA.

Santi, K.L., Kulesz, P.A., Khalaf, S., & Francis, D.J. (2015). Developmental changes in reading do not alter the development of visual processing skills: an application of explanatory item response models in grades K-2. *Frontiers in Psychology*, *6*, 116. https://doi.org/10.3389/fpsyg.2015.00116

Segall, D.O., Moreno, K.E., & Hetter, R.D. (1997). Item pool development and evaluation. In *Computerized adaptive testing: From inquiry to operation.* (pp. 117-130). American Psychological Association. https://doi.org/10.1037/10244-012

Septia, N.W., Indrawati, I., Juriana, J., & Rudini, R. (2022). An Analysis of Students' Difficulties in Reading Comprehension. *EEdJ: English Education Journal*, *2*(1), 11-22. https://doi.org/10.55047/romeo

Stenner, A.J. (2022). Measuring reading comprehension with the Lexile framework. In *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner* (pp. 63-88). Springer. https://doi.org/10.1007/978-981-19-3747-7_6

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, *41*(5), 721-732. https://doi.org/10.1080/02602938.2016.1164830

Sung, P.-J., Lin, S.-W., & Hung, P.-H. (2015). Factors Affecting Item Difficulty in English Listening Comprehension Tests. *Universal Journal of Educational Research*, *3*(7), 451-459. https://doi.org/10.13189/ujer.2015.030704

Swaminathan, H., Hambleton, R.K., Sireci, S.G., Xing, D., & Rizavi, S.M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*(1), 27-51. https://doi.org/10.1177/0146621602239475

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, *8*(1), 34-52. https://doi.org/10.1080/15434303.2010.536924

Toyama, Y. (2021). What Makes Reading Difficult? An Investigation of the Contributions of Passage, Task, and Reader Characteristics on Comprehension Performance. *Reading Research Quarterly*, *56*(4), 633-642. https://doi.org/10.1002/rrq.440

Trace, J., Brown, J.D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing*, *34*(2), 151-174. https://doi.org/10.1177/0265532215623581

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Valencia, S.W., Wixson, K.K., Ackerman, T., & Sanders, E. (2017). Identifying text-task-reader interactions related to item and block difficulty in the national assessment for educational progress reading assessment. In: San Mateo, CA: National Center for Education Statistics.

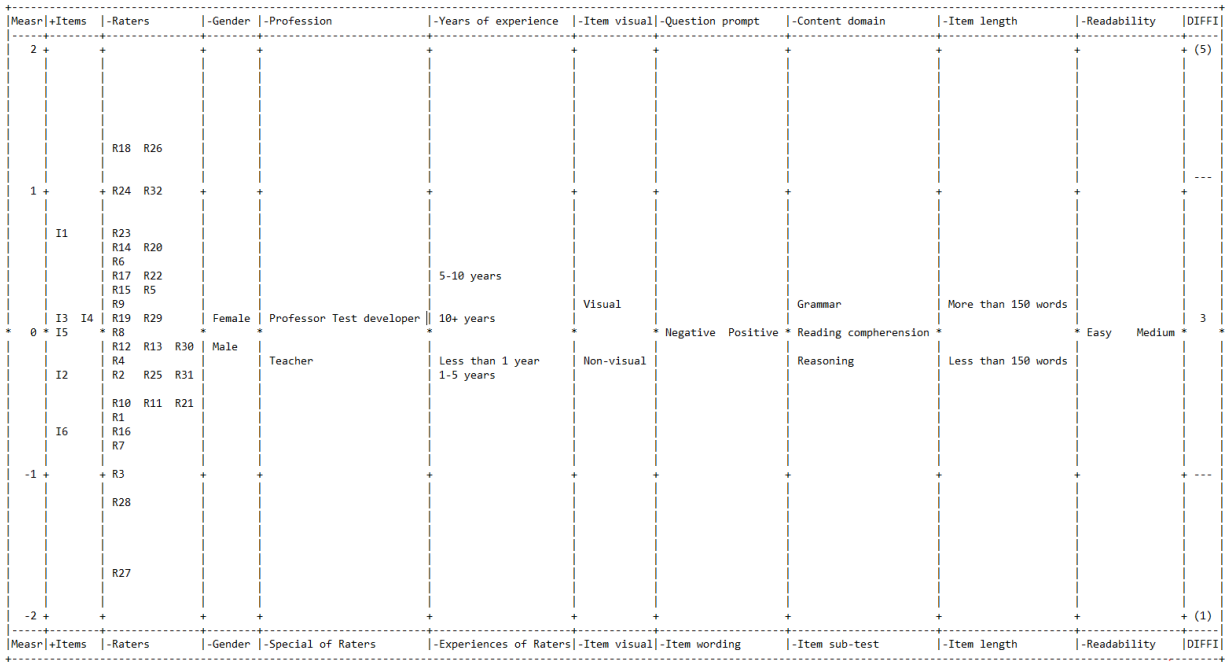
Van der Linden, W.J., & Pashley, P.J. (2009). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing* (pp. 3-30). Springer, New York, NY. https://doi.org/10.1007/978-0-387-85461-8_1

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, *58*(4), 1183-1193. https://doi.org/10.1016/j.compedu.2011.11.020

Ying-hui, H. (2006). An investigation into the task features affecting EFL listening comprehension test performance. *The Asian EFL Journal Quarterly*, *8*(2), 33-54.

APPENDIX

Appendix 1. *All facet vertical "rulers" of the first prediction.*



Appendix 2. *All facet vertical "rulers" of the second prediction.*