ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 440-454

Comparing Estimated and Real Item Difficulty Using Multi-Facet Rasch Analysis *

Ayfer SAYIN ** Sebahat GÖREN ***

Abstract

This study aimed to compare estimated item difficulty based on expert opinion with real item difficulty based on data by utilizing Rasch analysis. For security reasons, some high-stakes tests are not pre-tested and item difficulty is estimated by teachers in classroom assessments, so it is necessary to examine the extent to which experts make accurate predictions. In this study, we developed a 12-item test in the field of measurement and evaluation similar to those used in the Public Personnel Selection Exam. Item difficulty was estimated and compared separately based on 1165 student responses and the opinions of 12 experts. A multi-facet Rasch analysis was conducted to examine the effects of raters on the test scores. The study revealed that the experts had a good ability to estimate item difficulty for items of moderate difficulty. However, they tended to underestimate item difficulty for items.

Keywords: test development, item difficulty, subject matter experts, multi-facet Rasch

Introduction

Item difficulty, a crucial factor for educational assessments and personalized learning resource recommendations, is a concept that measures how difficult a test item is for a given group of test takers. For the exams prepared for these assessments to be effective, item difficulties need to be adjusted. Especially in standardized tests that are used to differentiate between students with different abilities, there is a need to use items of different difficulties: easy, moderate, and hard. This requires writing test items that meet certain quality standards to ensure that achievement on each item is linked to overall test performance, but it is challenging to write an item at a certain difficulty (Yaneva et al., 2020). For this, the item writer needs to have a good experience of the factors that affect item difficulty. In standardized tests, the difficulty of items should be estimated after the item is written and before it is administered. Besides automatic estimation methods, historically, this has been done through expert validation or pre-tests. Pre-testing is resource-intensive and requires considerable time and effort (Lin et al., 2019). Moreover, piloting for classroom assessments is costly and pre-testing is not preferred for safety reasons in high-stakes tests.

Recently, there has been increasing interest in using new methods, such as neural networks, machine learning, exoplanet item response theory, etc. to predict item difficulty (He et al., 2021; Qiu et al., 2019; Yaneva et al., 2020), and here we present findings on predictors of test item difficulty. Chon and Shin (2010) identified potential predictors of test item difficulty, such as response time and paragraph length based on related research and data collected from the College Scholastic Aptitude Test (CSAT). Beinborn et al. (2014) developed a model for the cloze-test difficulty that includes four dimensions: solution difficulty, candidate ambiguity, gap dependency, and paragraph difficulty. The results suggest that all four dimensions contribute to the overall difficulty of the C-test. Stadler et al. (2016) state that item difficulty can be accurately predicted using six key item characteristics, including the use and

* A part of this study was presented at 7th International Congress on Measurement and Evaluation in Education and Psychology. Gazi University, Ankara-Türkiye.

** Assoc. Prof. Dr., Gazi University, Faculty of Education, Ankara-Türkiye, ayfersayin@gazi.edu.tr, ORCID ID: 0000-0003-1357-5674
*** Research Assistant. PhD., Hacettepe University, Faculty of Education, Ankara-Türkiye, sebahatgoren@gmail.com, ORCID ID: 0000-0002-6453-3258

To cite this article:
Sayın, A. & Gören, S. (2023). Comparing estimated and real item difficulty using multi-facet rasch analysis. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 440-454. https://doi.org/10.21031/epod.1310893
Received: 7.06.2023
Accepted: 20.10.2023

number of self-dynamics, the number of input and output variables, the number of input and output variables not related to other variables, and the total number of relationships between all variables. Toyama (2021) found that several features of a passage, including sentence length, word frequency, syntactic simplicity, and temporality have a significant impact on comprehension difficulty. However, all these methods require the inclusion of predictor variables that predict item difficulty in the models they develop to predict item difficulty. However, it is not possible to talk about a variable that affects item difficulty. For example, while many studies have found that longer items tend to be more difficult (Fergadiotis et al., 2019; Lin et al., 2021; Pandarova et al., 2019; Yaneva et al., 2020), it has also been observed that longer items can be easier or item length does not affect item difficulty (Sano, 2015; Toyama, 2021). These findings showed that the difficulty model developed for one test may not be valid for other tests. In addition, in some types of modeling, difficulty features were identified, extracted, and presented as rules by experts (Beinborn et al., 2014; Grivokostopoulou et al., 2014; Perikos et al., 2016; Perkins et al., 1995). Therefore, it is crucial to determine the predictor variables to be used in the construction of models for item difficulties, and in this case, it is important to evaluate the accuracy of experts' difficulty predictions. Hence, expert opinions are often used for estimating item difficulty, but these estimates may differ from students' actual experience (Impara & Plake, 1998).

Expert estimation uncertainty may be due to a variety of factors involved in the cognitive process required to answer a question, as well as the tendency of test creators to overestimate student performance. Moreover, there is no standard that expert estimates accurately reflect item difficulty (Kurdi et al., 2021). Research has examined expert estimates and the cognitive operations involved in test items, but there is no guidance on what experts focus on when making difficulty estimates. Therefore, improving the accuracy of expert estimates of test difficulty requires a better understanding of the relationship between expert estimates and item difficulty (Hamamoto Filho et al., 2020). Attali et al. (2014) found that judges were successful in ranking multiple items in terms of difficulty, this ranking remained consistent among judges and across content areas of the Scholastic Aptitude Test [SAT]. Similarly, experts' ability to estimate item difficulty varied across different studies, with some showing good accuracy (Enright et al., 1993; Le Hebel et al., 2019; Lumley et al., 2012) and others showing limited predictive power (Kibble & Johnson, 2011; Sydorenko, 2011). For this reason, further studies should be conducted to determine the factors underlying the item difficulty of the experts because the predictions of the experts and item difficulty are not only important in the test development process, but also in the interpretation of test scores.

Item difficulty is crucial in setting the standard cut-off for passing or failing an exam. The Angoff method, which involves judges estimating the percentage of average examinees who will answer each test item correctly, is a commonly used criterion-referenced approach in determining the standard cut-off (Afrashteh, 2021; Wyse, 2020). The Angoff method is used to determine the final cut-off score by calculating the average of estimates made by referees for each item. This method is commonly used in high-stakes exams, such as medical exams, as it places a high value on expert opinions (Clauser et al., 2017; Impara & Plake, 1998; Kardong-Edgren & Mulcock, 2016; Wyse, 2018; Yim & Shin, 2020). In the current study, the emphasis was placed on assessing measurement and evaluation items that are similar to those found in the Public Personnel Selection Examination-[PPSE]

Many countries use selection and placement tests forteacher candidates. For example, The Praxis® exams are used to evaluate academic and subject-specific knowledge in the USA, according to the Educational Testing Service[ETS] (Praxis, 2022). It is worth noting that some states with significant teacher populations, such as California, New York, Texas, and Florida, have their own separate licensing exams (Gitomer & Qi, 2010). The Australian Institute for Teaching and School Leadership (AITSL) administers a range of assessments for teacher candidates, including the National Literacy and Numeracy Test for Initial Teacher Education students (AITSL, 2022); The Teaching Council of New Zealand requires all teacher candidates to pass the New Zealand Teachers Council Literacy and Numeracy Professional Skills Test (Ell, 2021). PPSE(Turkish KPSS) in Turkey includes a version specifically for individuals seeking to become teachers in the public school system. This test focuses on education-related subjects, such as pedagogy, educational psychology, and teaching methodologies

(OSYM, 2022). In the teacher certification exam for public institutions, there are 12 items related to assessment and evaluation. The difficulty of these items and the test as a whole is determined through expert opinions.

The aim of this research is to compare the accuracy of expert opinions in estimating item difficulty with real item difficulty based on data, particularly for high-stakes tests. This research lies in providing insights into the accuracy of expert estimates and identifying potential biases that may influence test scores. This information can be useful for improving the reliability and validity of high-stakes tests and ensuring that they accurately measure the knowledge and skills of test-takers.

Methods

Research Model

In this study, the relational survey design, which is a quantitative research method, has been used to demonstrate the relationship between multiple variables without intervention (Büyüköztürk et al.,2020).

Participants

Data were collected from two groups: pre-service teachers and experts who estimate the difficulty of the items. As summarized in Table 1, the first group of participants in the study included 1165 pre-service teachers who were in their third or fourth year of study at a faculty of education in a university. They all took a 14-week course on assessment and evaluation. The second group of participants comprised 12 experts who have either more than five years (5+) or less than five years (0-5) in the field of measurement and evaluation. They were also employed as instructors in education faculties, teaching courses related to assessment and evaluation. Experts who have more than five years of experience are familiar with both the course content and the participant group as they teach the students' courses, and experts who have less than five years of experience are acquainted with both the course content and the participant group as they assist the measurement and evaluation courses at the same universities which the data were collected. While the experts who have more than five years of experience prepare the exams themselves, the experts who have less than five years of experience help to prepare these exams as they assist the courses, and all experts determine the difficulty of the exams themselves. Since the difficulty of the PPSE is determined according to expert opinion, the difficulty of the achievement test developed in this study was also determined based on expert opinion.

Table 1.

Sample Descriptive Statistics

Participants 1 Pre-service teachers			Participants 2 Subject matter experts		
Characteristic	f	%	Characteristic	f	%
Gender			Gender		
Female	752	64.5	Female	10	83.3
Male	413	35.5	Male	2	16.7
Grade			Experiment		
3rdgrade	657	56.4	0-5 years	7	58.3
4th grade	508	43.6	5+ years	5	41.7

Table 2.

Sample Descriptive Statistics (Continued)

Participants 1 Pre-service teachers		-
Characteristic	f	%
Department		
Turkish and Social Sciences Education	531	45,6
Foreign Languages Education	252	21,6
Primary Education	189	16,2
Mathematics and Science Education	98	8,4
Special Education	95	8,2

Instrument

Achievement Test

In this study, we developed a 12-item test in the field of measurement and evaluation similar to those used in thePPSE.. Teacher candidates who apply to the PPSE to be appointed to the Ministry of National Education teacher positions are also required to take the Educational Sciences Test. Among the eight subtests within the Educational Sciences Test, the measurement and evaluation subtest accounts for approximately 6% of the total (OSYM Guide, 2023). In other words, there are 12 items from measurement and evaluation in the PPSE Educational Sciences Test, which consists of 80 items in total. In the achievement test developed in this study, the expert-subject effect of item difficulty perception was also tried to be determined by creating 12 items in five subjects including the most frequently asked subjects (alternative test tools, traditional test tools, item statistics, test statistics, interpretion of test scores). In addition, the fact that the items in the PPSE were prepared according to the university course contents supports the information that the items are appropriate for the course contents.

Prior to conducting factor analysis for the test's construct validity, the researchers examined whether the correlation matrix was suitable for factor analysis. Based on the results of the KMO and Bartlett's sphericity test (KMO=0.82, Bartlett's test=3039.97), exploratory factor analysis (EFA) was conducted. The parallel analysis showed that the difficulty of the responses for the items loaded a single dimension (see Appendix A). The results of the analysis revealed that the items explained 42% of the variance in the students' responses. Additionally, the reliability of the test was found to be 0.80 based on Cronbach's alpha coefficient. The factor loading values obtained from the EFA ranged between 0.484 and 0.792 (see Appendix B). Factor loading values greater than 0.30 for each factor indicate that the items serve the dimension well (Tabachnick & Fidell, 2013).

Expert Opinion Form

The experts were asked to estimate the difficulty of each item in the test. An expert opinion form was used in this process. In this form, the participants were first asked whether they had detailed information about the assessment items in PPSE, and those who answered "yes" to this item were included in the study. Then, the factors affecting the degree of difficulty were explained in detail in the form. Information was given about the semester averages of the participant group and the PPSE success ranking of those who graduated from the same department. Considering this information, 12 measurement and evaluation experts estimated item difficulty by rating each one on a scale of 1 to 5, where 1=very hard, 2=hard, 3=medium, 4=easy, and 5=very easy.

Data Analysis

The study analyzed data from 1165 students using exploratory factor analysis and the Rasch IRT model. The "ltm" package in R studio was used for exploratory factor analysis, and the Rasch IRT model was analyzed using the "TAM" package in R studio. In addition, multi-faceted Rasch analysis was conducted using the Minifac (Facets) package program to analyze data collected from 12 raters who rated 12 items. The study examined four sources of variability, including raters, items, item facets, and rater experience, and assessed model-data fit by examining standardized residual values. The results showed that there were seven values (0.48%) within the ±2 interval and 2 values (0.14%) within the ±3 interval, indicating acceptable model-data fitting. The data met all the assumptions, allowing for the analyses to be conducted.

Results

Real item difficulty (n=1165 students)

According to the results of Rasch IRT analysis, the difficulty parameters for each item are presented in Table 3. As it can be seen the item difficulty parameters in the test vary between -1.645 and 0.899. Furthermore, the test comprises items of varying difficulty levels: easy items (evidenced by negative coefficients), those of medium difficulty (coefficients near zero), and difficult items (marked by positive coefficients). The knowledge function in Figure 1 is a graphical representation of how well the test discriminates individuals with different ability levels. It shows that the test information function has a peak around 0 on the ability axis, indicating that the test is most informative for individuals with ability levels around 0. This means that the test is most accurate in discriminating students whose ability is close to the average ability level required for the test. The results of the Rasch analysis suggest that each item provides useful information about the difficulty parameters and the ability of the test to discriminate between individuals with different ability levels.

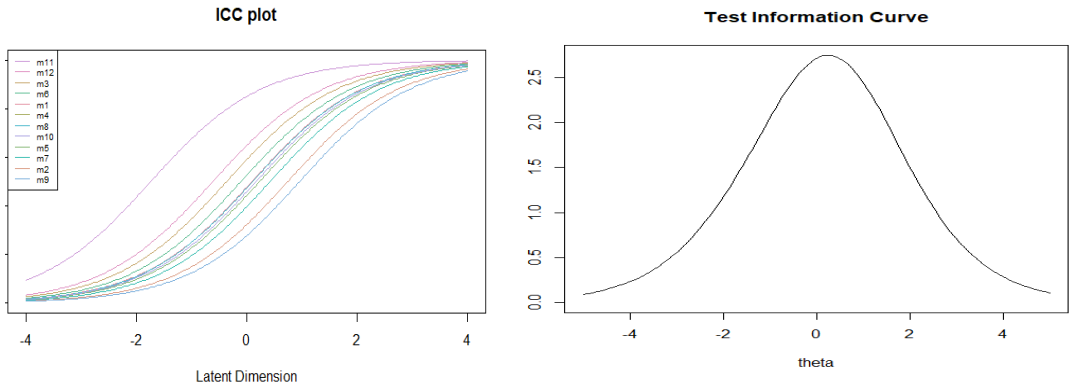
Table 3.

Results of the Rasch Analysis

Item	Difficulty value
I1	0.266
I2	0.899
I3	-0.198
I4	0.280
I5	0.405
I6	0.068
I7	0.591
I8	0.280
I9	1.101
I10	0.342
I11	-1.645
I12	-0.455
Model Summary:	
log.Lik -8167.34	
AIC 16361	
BIC 16426	

Figure 1.

ICC plot and Test Information Curve



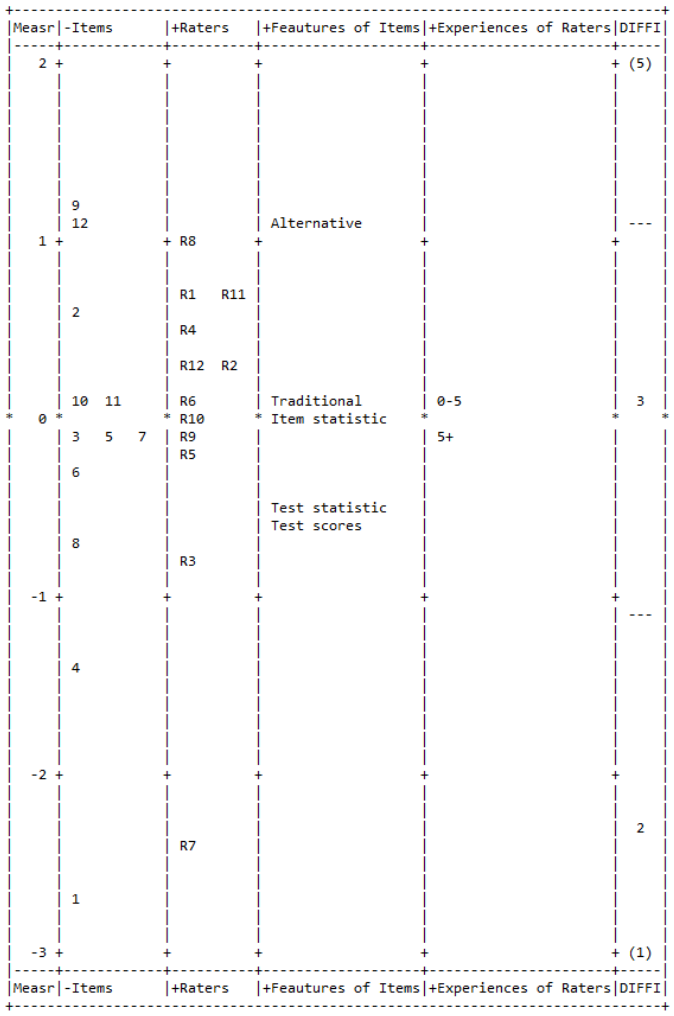
Prediction item difficulty (n=12 expert)

A multi-facet Rasch analysis was conducted to examine the effects of raters on the test scores. The four facets identified in this study were 12 items , 12 raters, four features of items (alternative,traditional, item statistic,test statistic, interpretation of test scores) and two experiences of raters (0-5, 5+ years). The item difficulty of 12 items are determined through the opinions of experts.12 measurement and evaluation experts estimated item difficulty by rating each one on a scale of 1 (very hard) to 5(very easy).

Figure 2 shows the distribution of items, raters, and item features on the same logit scale. The logit map gives general information about the facets and this measure allows for a comparability among variable sources in the study. In this distribution, the item facet is ranked from the most difficult to easiest, the rater facet is ranked from the most generous rater to the strictest rater and features of the item facet is ranked from the easiest subject to the most difficult, from top to bottom.The analysis revealed that I9 was the most difficult item and I1 was the easiest item. Among the raters, , R8 was the most lenient, while R7 was the strictest. Furthermore, items related to test scores and test statistics were found to be difficult, while items related to alternative topics were found to be easy. The multi-facet Rasch analysis is useful for examining the effects of raters on test scores, as it allows for the examination of multiple sources of variation and provides insights into the specific factors that affect the difficulty of test items.

Figure 2.

Logit Map of the Variables in the Study



Measurement Report for Item

The measurement report obtained from the multi-faceted Rasch analysis for the item facet is presented in Table 4. It is observed that the items were differentiated in terms of difficulty/easiness, and the highest and lowest logit values were found to be 1.22 and -2.69, respectively. The reliability index obtained from the Rasch analysis was also acceptable with a value of 0.82. Furthermore, the separation index of 3.22 indicates that the items were significantly different in terms of difficulty. However, it is concerning that only one item, I2, did not meet the criteria for both internal and external consistency. It may be necessary to revise or remove this item from the test to improve its reliability and validity.

Table 4.

Measurement Report for Item

Item	Logit	Std.error	Infit		Outfit	
			MnSq	Zst	MnSq	Zst
I9	1.22	0.46	1.45	1.1	1.34	0.8
I12	1.08	0.46	0.55	-1.2	0.54	-1.2
I2	0.61	0.43	0.23	-2.8	0.23	-2.8
I10	0.10	0.44	0.71	-0.7	0.72	-0.6
I11	0.10	0.44	1.05	0.2	1.06	0.2
I5	-0.08	0.44	1.65	1.5	1.65	1.5
I7	-0.08	0.44	1.00	0.1	1.03	0.2
I3	-0.13	0.44	1.77	1.7	1.79	1.7
I6	-0.27	0.44	0.52	-1.3	0.55	-1.2
I8	-0.70	0.44	0.56	-1.1	0.55	-1.2
I4	-1.45	0.44	1.84	1.8	1.89	1.9
I1	-2.69	0.54	0.73	-0.6	0.74	-0.5
Mean	0.00	0.45	1.00	0.00	1.01	0.00
SD	1.07	0.03	0.55	1.5	0.55	1.5
Model, Sample: RMSE = .45 Standard deviation = .97						
Discrimination ratio=2.17 Discrimination index = 3.22						
Discrimination index of reliability= 0.82						
Model, Fixed (all same) chi square=52.9 df =11 p= .00						
Model, Random (normal) chi square =9.3 df = 10 p= .50						

Measurement Report for Rater

The measurement report resulting from the multi-facet Rasch analysis of the rater facet is displayed in Table 5.The estimated separation ratio, separation index, and separation index reliability for the scoring facet are found to be high. The R8-coded rater is found to be the most generous, while the R7-coded rater is the strictest. When the separation index of 2.71 and the reliability coefficient of 0.76 are evaluated together with the chi-square test result ($\chi 2(df)=41.9(11)$, p=.00) for the fixed effect, it is determined that there is a significant difference among the raters who score the item difficulty of the items in terms of their strictness/generosity. The very low agreement index between the raters (-0.001) indicates that there is no agreement among the raters.

Table 5.

Measurement Report for Rater

Rater	Logit	Std.error	Infit		Outfit	
			MnSq	Zst	MnSq	Zst
R8	1,03	0.44	1.93	1.9	1.87	1.8
R1	0,74	0.44	1.29	0.7	1.32	0.8
R11	0,74	0.44	0.52	-1.3	0.53	-1.3
R4	0,55	0.44	1.21	0.6	1.26	0.7
R2	0,25	0.44	0.67	-0.8	0.66	-0.8
R12	0,25	0.44	0.36	-2.0	0.37	-2.0
R6	0,06	0.44	0.82	-0.3	0.81	-0.3
R10	-0,03	0.44	0.96	0.0	0.96	0.0
R9	-0,13	0.44	0.40	-1.8	0.41	-1.8
R5	-0,23	0.44	1.26	0.7	1.24	0.7
R3	-0,81	0.44	1.09	0.3	1.04	0.2
R7	-2,41	0.48	1.72	1.6	1.63	1.4
Mean	0.00	0.44	1.02	0.00	1.01	0.00
SD	0.91	0.01	0.50	1.3	0.48	1.3

Table 6.

Measurement Report for Rater (Continued)

Model, Sample: RMSE = .44 Standard deviation = .76
Discrimination ratio=1.79 Discrimination index = 2.71
Discrimination index of reliability= 0.76
Model, Fixed (all same) chi square=41.9 df =11 *p*= .00
Model, Random (normal) chi square =9.00 df = 10 *p*= .53
Observed inter-rater agreement: 36.9 %
Expected inter-rater agreement: 37.1%
Kappa inter-rater reliability statistics: -0.001

Measurement Report for Sub-test of the Items

The measurement report obtained through multi-faceted Rasch analysis for the *Features of Items* facet is given in Table 7. It is observed that the separation ratio, separation index, and separation index reliability calculated for item characteristics are high. Accordingly, a significant difference was found in the item difficulty of items based on their characteristics ($\chi 2(df)=18.1(4)$, p=0.00). A negative logit value indicates a low (difficult) score, while a positive logit value indicates a high (easy) score. Accordingly, items related to Interpretation of Test Scores and Test Statistic were found to be difficult, whereas items related to alternative topics were found to be easy.

Table 7.

Features of Items Measurement Report

Item	Logit	Std.error	Infit		Outfit	
			MnSq	Zst	MnSq	Zst
Alternative	1.07	0.34	1.28	1.0	1.19	0.7
Traditional	0.08	0.25	1.28	1.1	1.30	1.2
Item statistic	-0.05	0.30	0.76	-0.8	0.79	-0.7
Test statistic	-0.49	0.31	0.98	0.0	0.94	-0.1
Interpretation of Test Scores	-0.62	0.26	0.78	-0.9	0.77	-1.0
Mean	0.0	0.29	1.01	0.1	1.0	0.0
SD	0.67	0.04	0.26	1.0	0.24	1.0
Model, Sample: RMSE = .29 Standard deviation = .52						
Discrimination ratio=2.03 Discrimination index = 3.04						
Discrimination index of reliability= 0.80						
Model, Fixed (all same) chi square=18.1 df =4 p= .00						
Model, Random (normal) chi square =3.3 df = 3 p= .35						

Measurement Report for Rater's Experiment

The measurement report obtained through multi-faceted Rasch analysis for the rater's experiment facet is given in Table 8. According to the analysis results in Table 8, which evaluates the item difficulty of the items, there was no differentiation according to the experience of the raters, as the discrimination index was 0.99 and the reliability coefficient was 0.19 with a chi-square test result of ($\chi 2(df)=1.2(1)$, p=.27).

Table 8.

Measurement Report for Rater's Experiment

Item	Logit	Std.error	Infit		Outfit	
			MnSq	Zst	MnSq	Zst
0-5	0.14	0.18	0.96	-0.1	1.06	-0.2
5+	-0.14	0.18	1.06	0.3	0.94	0.4
Mean	0.00	0.45	1.01	0.00	1.01	0.1
SD	0.2	0.00	0.07	0.4	0.07	0.4

Model, Sample: RMSE = .18 Standard deviation = .00
Discrimination ratio=0.49 Discrimination index = 0.99
Discrimination index of reliability= 0.19
Model, Fixed (all same) chi square=1.2 df =1 p= .27

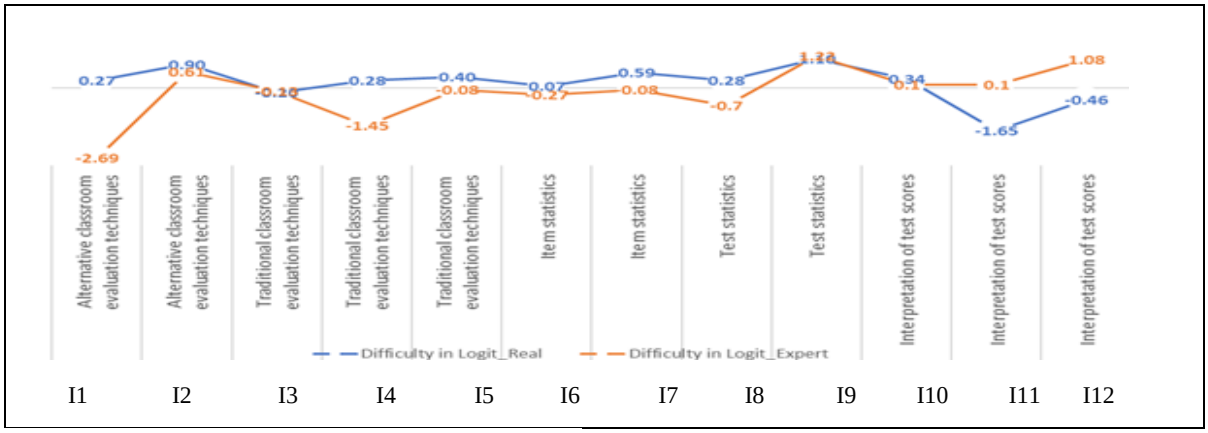
Compared real and prediction item difficulty

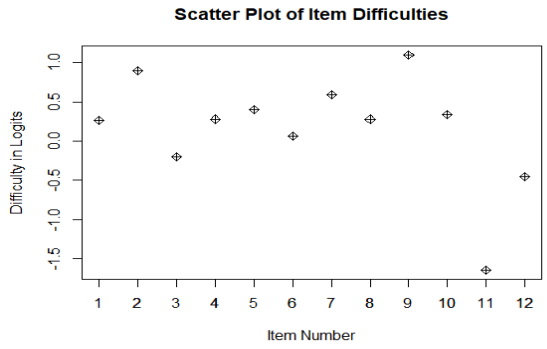
In accordance with the results gathered from 1165 students, the difficulties in logit values of the items in the measurement and evaluation test, which consisted of 12 items, were calculated using Rasch (Figure 3). Multi-facet Rasch analysis was used to estimate the difficulty in logit values for the estimations provided by 12 experts who participated in the study (Figure 3).

The experts predicted that the items were easier, except for the "Interpretation of test scores" subtest, where they estimated that 2 out of 3 items were actually more difficult.

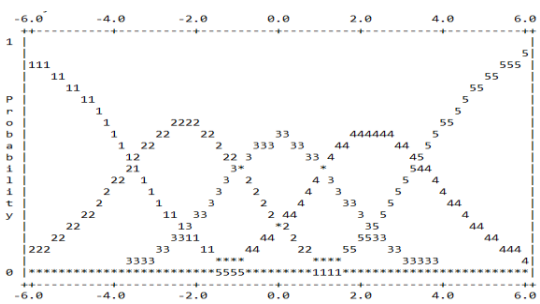
Figure 3.

Estimated and Real Item Difficulty in Logit





Difficulty in Logit_Real

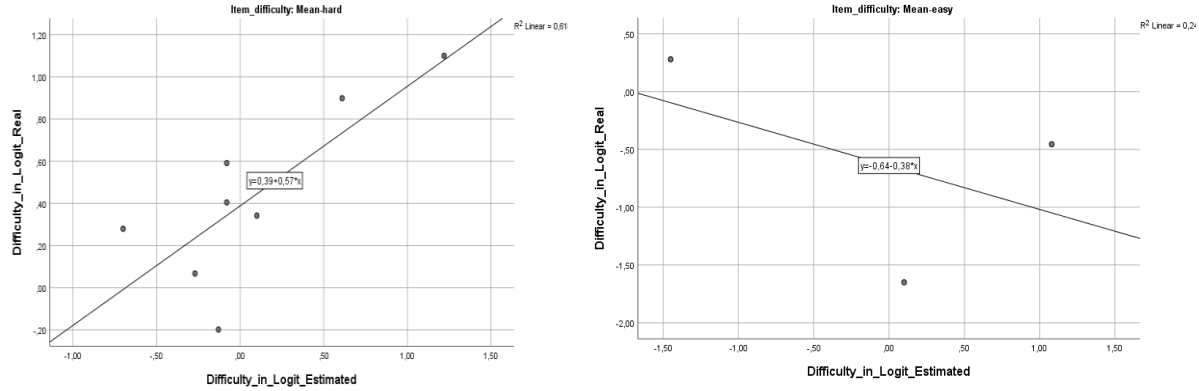


Difficulty in Logit_Estimated

To examine the relationship between the experts' estimates and the real item difficulty index in detail, items with an item difficulty parameter close to or above 0.00 (moderate and hard) and those with a negative value (easy) were studied separately. These comparisons of the difficulty in logit values are presented in Figure 4. As can be seen in Figure 4, while experts made better predictions for moderate items, their difficulty predictions for easy items were not as accurate as for moderate or hard items.

Figure 4.

Comparing the Estimated and Real Item Difficulty in Logit (moderate-hard and easy items)



Discussion

Due to security reasons, some high-stakes tests are not pre-tested, and the item difficulty in-class assessments is estimated by teachers. While there has been increasing interest in using new methods to predict item difficulty, these methods all require the inclusion of predictor variables in the models they build, and the predictors are identified and represented as rules by experts. Furthermore, the difficulty model created for one test may not be applicable to other tests. It is crucial to identify the relevant predictors to create accurate difficulty models and to assess the reliability of the experts' difficulty assessments. The study aimed to compare the experts' estimated item difficulty with the real values based on the data of the high stakes test like the PPSE in Turkey. The present research was to evaluate the accuracy of the experts' difficulty assessments by creating a high-stakes test that resembles the certification exam, and then comparing the results to their estimated item difficulty. The teacher certification exam for public institutions consists of 12 assessment and evaluation-related items. The item difficulty of these items and the overall exam are determined through the opinions of experts.

The results of this study suggested that experts in the field of assessment and evaluation had some bias in predicting item difficulty. Previous studies by Enright et al. (1993) and Wauters et al. (2012) demonstrated a strong positive correlation between expert ratings from science educators and correct rate in forecasting item difficulty. Moreover, they found that there was no significant difference between expert ratings and true value comparisons. However, Lumley et al. (2012) found that experienced experts were able to consistently predict item difficulty on reading tests. Sydorenko (2011) suggested that experts' ability to predict item difficulties may vary depending on the type of test and the specific items assessed. Furthermore, Kibble and Johnson (2011) reported a statistically significant but weak correlation between the intended difficulty of test items and actual student scores.

In our study, it was seen that the experts were adequate in estimating the medium item difficulty. Le Hebel et al. (2019) analyzed the difficulty of science inquiry tasks based on both estimated and real values in relation to students' abilities. The study also examined how accurately teachers predict

students' difficulty in answering Programme for International Student Assesment-[PISA]science questions. Like the previous study, Hamamoto Filho et al. (2020) found that the panel of experts' estimates of item difficulty had 54% correlation with real item difficulty. The study also found that items expected to be easy had significantly lower average difficulty than items expected to be moderate or difficult.

The average score of the students was calculated to be 46 out of 100, indicating that the test generally was of mean difficulty, while the experts estimated that the mean difficulty would be 51. The study found that the experts underestimated the difficulty of the test, with a particular bias towards underestimating items that were easy. Urhahne and Wijnia (2021) reviewed 10 studies that investigated the comparison between teachers' perceived and real difficulty of tasks. In 8 out of 10 studies, it was discovered that teachers often wrongly believed tasks were less challenging than they were, or expected students to perform better than they did. The studies were carried out in various academic subjects, including math, science, language arts, and a combination of language arts and science (Urhahne & Wijnia, 2021). Schult and Lindner (2018) found that teachers tend to underestimate the difficulty of items that require written answers.

The results of the study indicated that the accuracy of the experts' predictions varied across the subtests. The experts believed that the items in the test, excluding the "Interpretation of Test Scores" subtest, would be easier for the students. The accuracy of the experts in predicting the item difficulty in this subtest was lower compared to other topics. However, their predictions for the items in the "Item Statistics" subtopic were found to be more accurate.

The study's results showed that there were differences among experts' estimates of generosity-stinginess, but this variance was not associated with their years of experience. Thus, there is potential for improving the methodology and training used by experts to predict the item difficulty. Wauters et al. (2012) indicated that the inter-rater agreement for the estimation of the item difficulty by experts was good, with an ICC (Intraclass Correlation Coefficient) value of 0.68 for expert rating and for one-to-many comparison. Similarly, Attali et al. (2014) discovered that there was little variability in the quality of judgments across content areas and raters. This means that even new item writers who are not familiar with the items and not exposed to item statistics can perform similarly to more experienced SAT raters. The study implies that the ability to differentiate between the difficulties of the items is less related to test development experience and more linked to the specific difficulty scale used. While experts can assess the item difficulty, there can be variations in their evaluations, indicating room for improvement in their training and methodology. This information can aid in developing effective training programs for item writers and raters involved in test development.

In conclusion, the study highlights the importance of accurately predicting the item difficulty to ensure a fair and valid assessment of student performance. The findings suggest that further research is needed to improve the accuracy of expert estimations of item difficulty in high-stakes tests. The results of this study can be used to improve the accuracy of expert predictions and to refine the methods used for estimating item difficulty in the future. It also suggests the need for more objective and consistent methods that need attention to determine the predictors for predicting item difficulty, such as machine learning and item response theory, which can provide more reliable and accurate estimates of test difficulty. The results of this study can inform future research on item difficulty prediction and help improve the accuracy of expert opinions. It also indicates the need to create an item difficulty guide for item writers and moderators.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: This study was approved by the Ethical Committee of Gazi University dated 23.05.2023 and numbered E-77082166-604.01.02-673852.

References

Afrashteh, M. Y. (2021). Comparison of the validity of bookmark and Angoff standard setting methods in medical performance tests. *Bmc Medical Education*, *21*(1). https://doi.org/10.1186/s12909-020-02436-3

AITSL, A. I. f. T. a. S. L. (2022). *AITSL, Australian Professional Standards for Teachers*. https://www.aitsl.edu.au/tools-resources/resource/australian-professional-standards-for-teachers

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. http://dx.doi.org/10.1002/ets2.12042

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, *2*, 517-530. https://doi.org/10.1162/tacl_a_00200

Chon, Y. V., & Shin, T. (2010). Item difficulty predictors of a multiple-choice reading test. *English Teaching*, *65*(4), 257-282. http://journal.kate.or.kr/wp-content/uploads/2015/02/kate_65_4_11.pdf

Clauser, J. C., Hambleton, R. K., & Baldwin, P. (2017). The Effect of Rating Unfamiliar Items on Angoff Passing Scores.*Educational and Psychological Measurement*, *77*(6), 901-916. https://doi.org/10.1177/0013164416670983

Ell, F. (2021). Teacher education policy in Aotearoa New Zealand: Global trends meet local imperatives. In *Teacher Education Policy and Research: Global Perspectives* (pp. 113-128). Springer.

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://files.eric.ed.gov/fulltext/ED385595.pdf

Fergadiotis, G., Swiderski, A., & Hula, W. D. (2019). Predicting confrontation naming item difficulty. *Aphasiology*, *33*(6), 689-709. https://doi.org/10.1080/02687038.2018.1495310

Gitomer, D. H., & Qi, Y. (2010). Recent Trends in Mean Scores and Characteristics of Test-Takers on" Praxis II" Licensure Tests. *Office of Planning, Evaluation and Policy Development, US Department of Education*.

Grivokostopoulou, F., Hatzilygeroudis, I., & Perikos, I. (2014). Teaching assistance and automatic difficulty estimation in converting first order logic to clause form. *Artificial Intelligence Review*, *42*, 347-367. http://dx.doi.org/10.1007/s10462-013-9417-8

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. http://dx.doi.org/10.1590/1516-3180.2019.0459.R1.19112019

He, J., Peng, L., Sun, B., Yu, L. J., & Zhang, Y. H. (2021). Automatically Predict Question Difficulty for Reading Comprehension Exercises. *2021 Ieee 33rd International Conference on Tools with Artificial Intelligence (Ictai 2021)*, 1398-1402. https://doi.org/10.1109/Ictai52525.2021.00222

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://psycnet.apa.org/doi/10.1111/j.1745-3984.1998.tb00528.x

Kardong-Edgren, S., & Mulcock, P. M. (2016). Angoff Method of Setting Cut Scores for High-Stakes Testing Foley Catheter Checkoff as an Exemplar. *Nurse Educator*, *41*(2), 80-82. https://doi.org/10.1097/Nne.0000000000000218

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

Kurdi, G., Leo, J., Matentzoglu, N., Parsia, B., Sattler, U., Forge, S., Donato, G., & Dowling, W. (2021). A comparative study of methods for a priori prediction of MCQ difficulty. *Semantic Web*, *12*(3), 449-465. https://doi.org/10.3233/Sw-200390

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. http://dx.doi.org/10.13189/ujer.2021.090622

Lin, L. H., Chang, T. H., & Hsu, F. Y. (2019). Automated Prediction of Item Difficulty in Reading Comprehension Using Long Short-Term Memory. *Proceedings of the 2019 International Conference on*

Asian Language Processing (IALP), Shanghai, China, 132-135. https://doi.org/10.1109/IALP48816.2019.9037716.

Linacre, J.M. (2014). A user's guide to FACETS Rasch-model computer programs. Retrieved from http://www.winsteps.com/a/facets-manual.pdf

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012). A framework for predicting item difficulty in reading tests.

OSYM. (2022). *KPSS: Kamu Personel Seçme Sınavı*. https://www.osym.gov.tr/TR,23892/2022-kpss-lisans-genel-yetenek-genel-kultur-ve-egitim-bilimleri-oturumlarinin-temel-soru-kitapciklari-ve-cevap-anahtarlari-yayimlandi-31072022.html

Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R. D., & Brefeld, U. (2019). Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*, *29*, 342-367. https://doi.org/10.1007/s40593-019-00180-4

Perikos, I., Grivokostopoulou, F., Kovas, K., & Hatzilygeroudis, I. (2016). Automatic estimation of exercises' item difficulty in a tutoring system for teaching the conversion of natural language into first-order logic. *Expert Systems*, *33*(6), 569-580. https://doi.org/10.1111/exsy.12182

Perkins, K., Gupta, L., & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language testing*, *12*(1), 34-53.https://doi.org/10.1177/026553229501200103

Praxis, E. T. S. (2022). *ETS, The Praxis Tests*. https://www.ets.org/praxis

Qiu, Z. P., Wu, X., & Fan, W. (2019). Question difficulty prediction for multiple choice problems in medical exams. *Proceedings of the 28th Acm International Conference on Information & Knowledge Management (Cikm '19)*, 139-148. https://doi.org/10.1145/3357384.3358013

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,

Schult, J., & Lindner, M. A. (2018). Judgment Accuracy of German Elementary School Teachers: A Matter of Response Formats? *German Journal of Educational Psychology*, *32*(1-2), 75-87. https://doi.org/10.1024/1010-0652/a000216

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior*, *65*, 100-106. https://doi.org/10.1016/j.chb.2016.08.025

Sydorenko, T. (2011). Item writer judgments of item difficulty versus real 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

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

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

Wyse, A. E. (2018). Equating angoff standard-setting ratings with the rasch model.*Measurement-Interdisciplinary Research and Perspectives*, *16*(3), 181-194. https://doi.org/10.1080/15366367.2018.1483170

Wyse, A. E. (2020). Comparing cut scores from the angoff method and two variations of the hofstee and beuk methods. *Applied Measurement in Education*, *33*(2), 159-173. https://doi.org/10.1080/08957347.2020.1732385

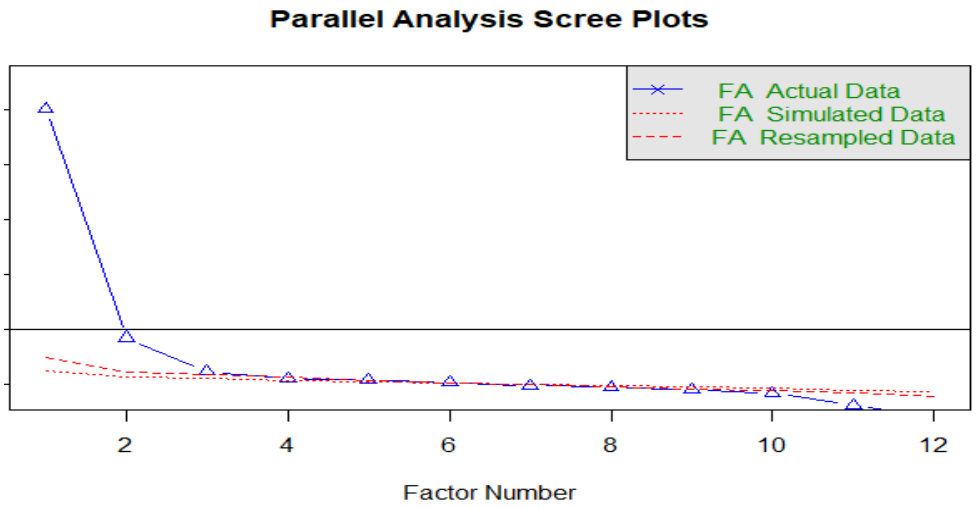
Yaneva, V., Ha, L. A., Baldwin, P., & Mee, J. (2020, May). Predicting item survival for multiple choice questions in a high-stakes medical exam. *Proceedings of the 12th International Conference on Language Resources and Evaluation (Lrec)*, 6812-6818. Marseille, France. https://aclanthology.org/2020.lrec-1.841.pdf

Yim, M. K., & Shin, S. J. (2020). Using the Angoff method to set a standard on mock exams for the Korean Nursing Licensing Examination. *Journal of Educational Evaluation for Health Professions*, *17*(4). https://doi.org/10.3352/jeehp.2020.17.14

Appendix

Appendix A.

Parallel Analysis Scree Plots



Appendix B.

Results of the EFA

Item	Factor loadings
I1	0.579
I2	0.621
I3	0.532
I4	0.604
I5	0.590
I6	0.623
I7	0.639
I8	0.859
I9	0.691
I10	0.792
I11	0.647
I12	0.484
Variance	%42
α	0.80