Asian Journal of Research in Education and Social Sciences
e-ISSN: 2682-8502 | Vol.7, No. 2, 209-222, 2025
https://asianscholarsnetwork.com/asnet-journals
SPECIAL ISSUE: International Conference on Future and Sustainable Education (ICFSE) 2024

Lecturers' Question Difficulty Estimation: Measuring Difference Between Table of Specification Test and Students' Results of Human Sciences Programme Courses

Shamsuddin Abddullah[1], Jefri Khairi Zabril[1*], Aimi Aqilah Kamaruzzali[1], Wan Ahmad Khairi Wan Ahmad[1], Nabilah Diyana Abdul Aziz[1]

[1] Department of Human Sciences, Centre for Foundation Studies, International Islamic University Malaysia

*Corresponding Author: jefrikhairil@iium.edu.my

Received: 15 July 2024 | Accepted: 1 December 2024 | Published: 1 March 2025

DOI: https://doi.org/10.55057/ajress.2025.7.2.21

Abstract: *Questions in educational assessments must be valid, reliable and fair to students to ensure minimal error in measuring student learning outcomes. One of the methods to ensure fairness of questions used in educational assessments is that to plan for appropriate proportion of questions of different difficulties. Lecturers of Human Sciences programme in the Department of Human Sciences, Centre for Foundation Studies, International Islamic University Malaysia, design Table of Specification (TST) to distribute different difficulties of questions for an educational assessment and lecturers estimate the questions' difficulty based on expert opinion. Currently, the accuracy of these estimations has yet to be verified as part of the department's assessment practices. This situation impedes the continual quality improvement of the questions. The purpose of this paper is to verify the accuracy of lecturers' estimation through measuring the difference between the question difficulty stated in the TST and the difficulty stated in the results of item analysis from students' results. Pen and paper tests of all five courses in the Human Sciences programme were selected and the comparison of question difficulty between TST and item analysis from students' results was made. Results suggested that there is a difference between question difficulty stated in TST and item analysis from students' results. The explanation of results and the limitation of the study are described in the discussion section. In conclusion, the paper indicated that there should be a review of TST and the questions to improve the quality of the educational assessment.*

Keywords: Educational Assessment, Table of Specification Test, Question Difficulty Estimation, Question Difficulty Based on Students' Results, Assessment Fairness

1. Introduction

Research Background

The core ideas of assessments are validity, reliability, and fairness. According to Kane (2013), validity is the fundamental tenet of assessments, meaning that they should measure the things they are supposed to measure. It is further separated into two categories: content validity and construct validity. According to Cronbach and Meehl (1955), construct validity refers to how well an assessment measures the theoretical construct it is supposed to examine. On the other hand, content validity is concerned with how well the subject being measured is covered and if it follows the curriculum requirements (Lane, 2014). Meanwhile, fairness in assessments refers to the equal opportunity given to students in demonstrating their skills and knowledge.

(American Educational Research Association et al., 2014). This goes beyond the aspects of the assessment design itself. It covers elements that come before the assessment (e.g.: resources and access) as well as its consequences (like impact and outcome interpretations) (Gipps & Stobart, 2009). Camili (2006) emphasised that fairness should be the guiding principle throughout the assessment process, from creating test items to analysing test results. In addition, there is a key research gap in understanding how fairness and validity work together to create equitable assessments across diverse educational settings. While existing research covers validity and fairness as essential principles, it provides limited guidance on how to apply these concepts, especially regarding question difficulty. Appropriate question difficulty is essential for fairness, as poorly calibrated questions can skew results, impacting students' scores for reasons beyond their actual knowledge and skills. This emphasises how crucial it is for the lecturer to estimate the question difficulty to guarantee that the three core assessment principles are met and the assessment reach its standard. But the most important factor in deciding the degree of question difficulty is assessment fairness. Failure to design a fair level of question difficulty may affect students' results. Therefore, further research into balanced assessment design is essential to improve accuracy and fairness in measuring student performance.

Research Objective

The main objective of this paper is to investigate the difference between the lecturer's estimation of question difficulty, as outlined in the Test Specification Table (TST), and the calculated Difficulty Index (DI) of the questions.

2. Literature Review

In this sub-section, existing literature on the methodologies for determining question difficulty and the evaluation of the accuracy in lecturers' estimations of question difficulty will be reviewed. This review aims to highlight key findings and approaches used in the field, ultimately leading to the identification of a relevant research gap.

Assigning question difficulty

The methods to determine item difficulty in assessments are essential for ensuring the validity, reliablity, and fairness of the test. Accordingly, the quality and validity of an assessment cannot be compromised as it is heavily dependent on the item's quality, or in this case, the quality of the questions assessed in an examination. Therefore, the lecturers have a huge responsibility in assigning item difficulty as item difficulty estimation is heavily prominent in determining the quality of an assessment (Al-Khuzaey, Grasso, Payne & Tamma, n.d.). In addition, the traditional method of determining question difficulty during assessment planning is that the questions are constructed based on the lecturers' expertise on the subject matter, but changes need to be made. According to Rafatbakhsh & Ahmadi (2023), it is important for the lecturers to construct questions with an accurate level of difficulty, and it can be done by measuring the discrepancies between difficulty levels and students' results.

In addition to delivering lectures and grading examinations, lecturers possess the duty of formulating diverse assessments, which encompasses the development of examination questions. The process of creating test questions that are deemed valid, reliable, and fair involves the categorization of item difficulty into easy, moderate, and hard levels. Several research advocate for the incorporation of a combination of easy, moderate, and hard questions to ensure student engagement and precise evaluation (Haladyna & Rodriguez, 2013), with easy questions primarily focusing on recall, while more difficult questions necessitate critical analysis or evaluation (Webb, 2002). Furthermore, educators utilize Bloom's Taxonomy to classify questions based on their level of complexity.

Assessing the accuracy of lecturer’s estimation of difficulty

Impara and Plake (1998) mentioned that in ranking questions by difficulty, lecturers tend to underestimate the difficulty of hard items and overestimate the difficulty of easy items. This somehow indicates that students’ results do not match the difficulty level set by lecturers. One possible explanation for this may stem from the influence of the teacher's proficiency. According to Goodwin (1999), instructors tend to possess advanced knowledge within their respective domains. Their extensive expertise and depth of understanding in the specific subject area could hinder them from aligning themselves with the level of understanding of their students. However, scarce numbers of studies are available in measuring the distance between a lecturer’s estimation of the difficulty of questions with the difficulty index of the questions.

Van de Watering and Van der Rijt (2006) iterates that lecturers exhibited the capacity to accurately gauge the levels of difficulty for merely a small portion of the assessment items but displayed a tendency to overrate the difficulty of the most assessment items. Conversely, students tend to underestimate their own achievements. One possible explanation provided was the difficulty in visualizing the proficiencies and skills of the students due to the high expectation set by the lecturers. The instructors predominantly overestimated the students' performances for most items, indicating that said items proved to be more difficult for the students compared to the lecturers' expectations. Thus, it is suggested that an emphasis on discussion and training might enhance the accuracy of the estimations for the item's difficulties.

Research Questions (RQ)

In this research, we aim to answer the following research questions:

RQ1: How are the questions distributed based on their difficulty extracted from the Test Specification Table (TST) that is based on lecturer’s estimation (LE), and the one calculated through Difficulty Index (DI)?
RQ2: What is the percentage of LE being the same with DI?
RQ3: How much is the distance between LE and DI?
RQ4: Is the distance between the LE and the DI statistically significant?
RQ5: Is there a significant difference between the courses in terms of their distance between LE and DI?

3. Methodology

In this section, we elaborate the research methodology which includes the research design, research model, sample and procedure.

Research Design

This research employed a quantitative descriptive research design in which data from the TST is extracted and compared with the data gathered from DI.

Research Model

Figure 1 shows the research model employed in achieving its research objective. Investigating the distance between the LE and DI will help improve the LE for future assessments planning.

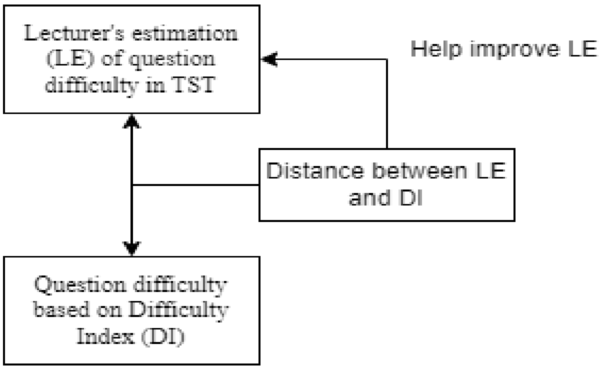
Lecturer's estimation (LE) of question difficulty in TST
Help improve LE
Distance between LE and DI
Question difficulty based on Difficulty Index (DI)

Figure 1: Research Model.

Sample For Lecturers' Estimation (LE)

Test-Specification-Table (TST) were collected from the five courses that were offered during Cohort 2023/2024.

Courses	TYPE OF TST
ITC	EOSE
ITH	EOSE
ITP	EOSE
ITPOL	TEST
ITSA	EOSE

**ITC=Introduction to Communication; ITH=Introduction to History; ITP=Introduction to Psychology; ITPOL= Introduction to Political Science; ITSA= Introduction to Sociology & Anthropology*

How the TSTs were developed

The development of a TST is a structured approach to ensure assessment alignment with learning objectives and content coverage (Fives & DiDonato-Barnes, 2013). The process begins with defining clear learning objectives and analysing course content (Nitko & Brookhart, 2020). Cognitive levels are then determined, often using frameworks like Bloom's Taxonomy (Krathwohl, 2002).

A matrix is created, mapping content areas against cognitive levels, with allocations for item numbers and weights (Downing & Yudkowsky, 2009). Item types are specified for each cell, ensuring variety in assessment methods (Haladyna & Rodriguez, 2013). The table undergoes review and validation to ensure content validity and curriculum alignment (Kane, 2013).

The TST serves as a blueprint for test assembly, guiding item selection or creation (Schmeiser & Welch, 2006). The process is iterative, with refinements based on test results and feedback (Brookhart et al., 2016). This systematic approach enhances the quality and fairness of assessments while maintaining alignment with educational goals (Oermann & Gaberson, 2021).

How TST represents lecturer's estimation

Lecturer's estimation (LE) influences various aspects of the TST, including content weighting, cognitive level distribution, item type selection, and difficulty level estimation (Nitko & Brookhart, 2020; Krathwohl, 2002; Haladyna & Rodriguez, 2013). The TST incorporates the lecturer's judgments on time allocation, emphasis on learning objectives, and predictions of student performance (Oermann & Gaberson, 2021; Fives & DiDonato-Barnes, 2013; Südkamp et al., 2012).

The table can be adjusted based on past performance data and aligned with successful teaching strategies (Brookhart et al., 2016; McMillan, 2013). Contextual factors estimated by the lecturer are also considered in the TST structure (Bennett et al., 2012).

This integration ensures the TST accurately reflects the course's emphases and challenges as perceived by the instructor, aligning formal assessment with professional judgment and course experience.

Sample For Difficulty Index (DI)

In determining the size of assessment sample size, McCoach et al. (2013) emphasize the importance of a representative sample (i.e it should reflect the target population). This is due to the classical test theory fact that larger samples provide more accurate estimates of item difficulty and discrimination while collecting small samples may lead to unstable or misleading item statistics (Thorndike & Thorndike-Christ, 2010). To conduct a meaningful item analysis, the number of samples required can vary depending on the specific statistical methods used and the desired level of precision. A commonly cited rule of thumb is to have at least 5-10 sample per item for initial item analysis (Nunnally & Bernstein, 1994). For more stable results, larger sample sizes are recommended. According to Crocker and Algina (2008), a sample size of at least 200 sample is generally recommended for item analysis in classical test theory. This sample size helps ensure more stable item statistics and reduces the impact of sampling error. Based on the recommendation and to strengthen the sample, this research collected 60 samples of answer scripts from four courses, which totalling to 240 assessment samples used for the DI analysis.

This research focuses on Human Sciences-based courses offered in all three semesters of cohort 2023-2024 for HS students at the Department of Human Sciences. The courses studied for this research are Introduction to Communication (ITC) (SSHF 0314), Introduction to Sociology and Anthropology (ITSA) (HSSF 0314), Introduction to Political Science (ITPOL) (HSSF 0334), Introduction to Psychology (ITP) (HSSF 0344), and Introduction to History (ITH) (HSSF 0354). Apart from the full name and/or course code, abbreviations that refers to the respective courses are also used (i.e ITC, ITSA, ITPOL, ITP, ITH) in certain parts of writing of this research from here on.

For this research, one pen and paper-based assessment of the aforementioned courses is selected, and marks of every question are extracted and analysed. The following are the details of the number of questions for each course:

Type of Questions	**MCQ**	**SAQ**	**SEQ**	**TOTAL**
Courses				
ITC	14	11	-	**25**
ITH	10	10	2	**22**
ITP	67	-	3	**70**
ITPOL	8	8	-	**16**
ITSA	19	-	9	**28**
TOTAL	**118**	**29**	**14**	**161**

**ITC=Introduction to Communication; ITH=Introduction to History; ITP=Introduction to Psychology; ITPOL= Introduction to Political Science; ITSA= Introduction to Sociology & Anthropology*
MCQ= Multiple-Choice Questions; SAQ= Short Answer Questions; SEQ= Short Essay Questions

Tools

i. Lumen Ex Machina 4 100x100 (LEM4) template by Carlo Excels

LEM4 is a free all-in-one item analysis template for teachers (Microsoft Excel tool) developed by Carlo Excels. The function of the template is as a tool to conduct item analysis for educational assessment. Based on the entered test data, the template will provide users extractable processed item analysis data (i.e. Items' Reliability, Item Difficulty analysis, Item Discrimination analysis, and Multiple-Choice Distracter analysis).

ii. Jeffreys's Amazing Statistics Program (JASP version 0.18.3.0)

JASP is a free, user friendly, also open-source statistical analysis program developed by the University of Amsterdam. It offers both classical and Bayesian analysis methods. This application is useful for this research as 1) one-sample t-test is needed to compare the result of the mean distance of LE and Di for all courses and 2) ANOVA test is needed determine whether there is any significant difference between the LE and DI of all courses.

Procedure

The marks of pen and paper-based assessments of the aforementioned courses are extracted and analysed using Lumen Ex Machina 4 100x100 (LEM4) template by Carlo Excels. The course's general information (i.e Class, Teacher, Test Name, etc.) are entered and is followed by the item types also marks. Afterwards, the student's individual marks for each question is then key-ed in the RawData sheet of the excel and the assessment's item difficulty and discrimination are generated.

Afterward, the notation of question difficulty (Easy, Medium, Hard) estimated by lecturers (LE) were extracted from their TST and the DI were calculated based on students' results and the difficulty of the questions were determined based on the following standards: $DI < 0.25$ = Hard; $DI > 0.75$ = Easy; and $0.25 < DI < 0.75$ = Medium. Both distribution of difficulty of every question from LE and DI are then compared with one another to determine the percentage of hit. A comparison of a question will be considered as a 'hit' if both (from LE and DI) score the same level of difficulty (i.e Easy). The percentage is calculated by summing all hits divided by total questions times by 100; thus, will show either consistency of inconsistency between LE and DI for all courses

Then, the calculation of distance between LE and DI is done by assigning numbers to the difference of difficulty observed for a question. If there is no difference between the difficulty of a question based on LE and DI, the distance was recorded as 0. If there is one-level difference (i.e. Easy-Medium, Medium-Hard), the distance was recorded as '1'. If there is two-level difference (i.e. Easy-Hard), the distance was recorded as '2'.

The distance for all the questions in all the courses were recorded and the mean distances were calculated based on the courses. Since the distance value ranges between 0-2, mean values closer to 0 indicates low distance while mean value closer to 2 indicates higher distance between LE and DI.

Thereafter, the result of the mean distance of LE and Di for all courses are then compared using one-sample t-test. The test was ran using inferential statistical analysis software (JASP version 0.18.3.0) and the result of the test will inform whether the mean distance is significantly more than "0" or not. In other words, whether the lecturers' estimation of difficulty of the questions were consistent or inconsistent with students results. Last, an ANOVA test was conducted to obtain the result of whether there is any significant difference between the LE and DI of all courses.

4. Results and Discussion

In this section, we elaborate the results, discussion and limitation of this research and implications of the research.

Results

This sub-section we present the results of the research based on the RQs.

RQ1: How are the questions distributed based on their difficulty extracted from the Test Specification Table (TST) that is based on lecturer's estimation (LE), and the one calculated through Difficulty Index (DI)?
To answer RQ1, the notation of question difficulty estimated by lecturers (LE) were extracted from their TST and the DI were calculated based on students' results and the difficulty of the questions were determined based on the following standards: DI < 0.25 = Hard; DI > 0.75 = Easy; and 0.25 < DI < 0.75 = Medium. The distribution of difficulty of questions from LE and DI are presented in Table 1.

Table 1: Distribution of difficulty of questions from LE and DI.

Sources	LE from TST			TOTAL	DI			TOTAL
Difficulty	Easy	Medium	Hard		Easy	Medium	Hard	
Courses								
ITC	14 (56%)	7 (28%)	4 (16%)	25	10 (40%)	15 (60%)	0 (0%)	25
ITH	10 (46%)	8 (36%)	4 (18%)	22	7 (32%)	15 (68%)	0 (0%)	22
ITP	32 (45%)	23 (33%)	15 (22%)	70	37 (53%)	33 (47%)	0 (0%)	70
ITPOL	9 (56%)	3 (19%)	4 (25%)	16	6 (37%)	10 (63%)	0 (0%)	16
ITSA	14 (50%)	5 (18%)	9 (32%)	28	10 (36%)	15 (53%)	3 (11%)	28
TOTAL	79 (49%)	46 (29%)	36 (22%)	161	70 (43%)	88 (55%)	3 (2%)	161

There was a total of N=161 questions analysed for this research from all courses related to the Human Sciences programme. The results showed that from the TST, the questions were allocated to all difficulties for all courses. However, in the DI, the results showed that for all courses, there were many questions that were estimated to be hard were found to be distributed in easy and medium.

The results also showed that the number of questions for the assessment across all the courses are different ranging from the least, 16 questions (ITPOL) to the most, 70 questions (ITP). It is important to note that different types of questions are used in all courses which mainly fell under three types of questions namely multiple-choice questions (MCQ), short answer questions (SAQ) and short essay questions (SEQ). It was observed that in ITP, all the questions were MCQ which made the question count high. This is due to the nature of the course which emphasised on mass content coverage.

RQ2: What is the percentage of LE being the same with DI?

To answer RQ2, every question in each course were compared in terms of its difficulty based on LE and DI. For example, if in ITC, Question 1 in LE is Easy and in DI is also Easy, this will count as one 'hit'. All questions were compared. The percentage is calculated by summing all hits divided by total questions.

Table 2: Number and percentage of hits.

Courses	Number of hits / total questions	Percentage of hits
ITC	15/25	60%
ITH	13/22	59%
ITP	26/70	37%
ITPOL	7/16	44%
ITSA	10/28	28%
TOTAL	71/161	44%

The results showed that ITC and ITH have the highest percentage of hits, indicating roughly a nearer distance between LE and DI, indicating a more accurate estimation of difficulty of questions. ITP, ITPOL and ITSA have hit percentage that are lower than 50% with ITSA being the lowest at 28%. The results suggested that there is high inconsistency between LE and DI for ITP, ITPOL and ITSA. Overall, for the percentage of hits against the total number of questions, it was 44%, lower than 50% is considered low.

RQ3: How much is the distance between LE and DI?

To answer RQ3, the distance between LE and DI is calculated by assigning numbers to the difference of difficulty observed for a question. If there is no difference between the difficulty of a question based on LE and DI, the distance was recorded as 0. If there is one-level difference (i.e. Easy-Medium, Medium-Hard), the distance was recorded as '1'. If there is two-level difference (i.e. Easy-Hard), the distance was recorded as '2'. For example, in ITC, the difficulty of Question 1 from LE is Easy, and from the DI is Medium, therefore the distance for Question 1 recorded as '1'. The distance for all the questions in all the courses were recorded and the mean distances were calculated based on the courses. Since the distance value ranges between 0-2, mean values closer to 0 indicates low distance while mean value closer to 2 indicates higher distance between LE and DI. The mean distances according to courses are presented in Table 3.

Table 3: Mean distance between LE and DI according to different courses.

Course	N	Mean	SD	SE
ITC	25	0.400	0.500	0.100
ITH	22	0.409	0.503	0.107
ITP	70	0.771	0.685	0.082
ITPOL	16	0.688	0.704	0.176
ITSA	28	0.786	0.630	0.119

**N=total number of questions.*

The results showed that ITC (N=25) had the lowest mean distance between LE and DI which is 0.4. On the other hand, ITSA (N=28) had the hight mean distance between LE and DI which is 0.786. ITC and ITH have lower mean distance between LE and DI as compared to ITP, ITPOL and ITSA which are values near to 0.7. However, it is important to also note that ITP, ITPOL and ITSA have quite high SD for their mean, indicating that the distances of questions for these courses were not as consistent as ITC and ITH which have SD value of 0.5. Table 4 presents the mean distance between LE and DI for all courses in Human Sciences.

Table 4: Mean distance between LE and DI for all courses in Human Sciences Programme.

Descriptives

	N	Mean	SD	SE
Distance	161	0.658	0.643	0.051

The results showed that the mean distance between LE and DI for all courses is below 1, the middle point of 0-2, roughly indicating a low distance between LE and DI. However, it must be corroborated with other tests to have a better interpretation of data.

RQ4: Is the distance between the LE and the DI statistically significant?

To answer RQ4, a one-sample t-test was conducted to compare the mean distance of LE and DI with the value of no difference, which is “0”. The result is presented in Table 5.

Table 5: Results of one sample t-test of mean distance between LE and DI.

One Sample T-Test

	t	df	p
Distance	12.986	160	< .001

Note. For the Student t-test, the alternative hypothesis specifies that the mean is greater than 0.

The results showed that the mean distance between LE and DI (0.658) is significantly larger than “0” at p < .001. Therefore, the results showed that the distance between the LE and DI is statistically significant. The results suggested that the lecturers’ estimation of difficulty of the questions were not as consistent with students results.

RQ5: Is there a significant difference between the courses in terms of their distance between LE and DI?

To answer RQ5, an ANOVA test was conducted to compare the mean distances between LE and DI of all courses to find out if there are significant differences between them. The result is presented in Table 6.

Table 6: Results of ANOVA test.

ANOVA - Distance

Cases	Sum of Squares	df	Mean Square	F	p
Course	4.398	4	1.100	2.775	0.029*
Residuals	61.813	156	0.396		

Note. Type III Sum of Squares
**p < .05*

The results showed that according to the ANOVA test, the mean difference between the mean distances of all courses were significantly different at p < 0.05. The results suggested that the courses should not be treated the same in terms of their distances between the LE and DI. Since the ANOVA test indicated a statistically significant difference, there is a need to find out which of the courses were significantly different from another course. Therefore, a post hoc comparison was conducted. The Games-Howell post hoc comparison test was used because of the unequal group size between the courses. The results are presented in Table 7.

Table 7: Results of Games-Howell Post Hoc Comparisons
Games-Howell Post Hoc Comparisons - Course

Comparison	Mean Difference	SE	t	df	p_{tukey}
ITC - ITH	-0.009	0.147	-0.062	44.168	1.000
ITC - ITP	-0.371	0.129	-2.875	57.877	0.043*
ITC - ITPOL	-0.287	0.202	-1.420	24.639	0.621
ITC - ITSA	-0.386	0.155	-2.481	50.346	0.111
ITH - ITP	-0.362	0.135	-2.685	47.633	0.071
ITH - ITPOL	-0.278	0.206	-1.350	25.682	0.663
ITH - ITSA	-0.377	0.160	-2.350	47.978	0.147
ITP - ITPOL	0.084	0.194	0.432	21.960	0.992
ITP - ITSA	-0.014	0.144	-0.099	53.839	1.000
ITPOL - ITSA	-0.098	0.213	-0.462	28.541	0.990

* $p < .05$

Note. Results based on uncorrected means.

The results showed that there is only one significant difference between ITC and ITP where the p < 0.05. It is interesting to note that the mean difference between ITC - ITP is 0.371 which is not the highest among the other comparisons such as ITC - ITSA (0.386) and ITH - ITSA (0.377) which have higher mean differences. This is probably because of ITP having many questions than other courses. The results suggested that there is significant difference between ITC and ITP in terms of the mean difference between LE and DI. However, caution needs to be taken when interpreting this significance as it may be influenced by the number of questions by ITP being more than the rest of the courses.

Discussion

The study analysed a total of 161 questions from all courses related to the Human Sciences program. The number of questions varied across courses, ranging from 16 questions in ITPOL to 70 questions in ITP. Different types of questions were used across all courses, primarily multiple-choice questions (MCQ), short answer questions (SAQ), and short essay questions (SEQ).

The RQ1 investigated how the difficulty levels of questions, as estimated by LE and calculated through the DI were distributed. The DI results indicated that many questions initially estimated to be hard were found to be easy or medium. This is probably due to overestimation by HS lecturers towards the DI. The results for RQ1 are consistent with the previous studies, discussed by Impara and Plake (1998) that teachers could generally rank order items by difficulty, however they tended to overestimate the difficulty of easy items and underestimate the difficulty of hard items. It is likely that HS lecturers' accurate estimation is pivotal to obtain correct difficulty levels of questions and to be able to construct questions according to teaching guidelines.

The RQ2 examined the percentage of agreement between the lecturers' estimations of questions difficulty and the DI. A match or 'hit,' was recorded if both LE and DI classified the question at the same difficulty level. The results found that ITC and ITH had the highest percentage of hits, indicating a closer arrangement between LE and DI, showing higher accuracy estimations of question difficulty. The results shown are consistent with the previous studies in which the satisfactory outcomes of the assessment can be used to improve the quality of examinations questions for the purpose of attaining the target of teaching and learning outcomes (Juridah, Jaafar, Dzuraidah, Shahrum, Shahrir, Mohd Zaidi & Norhamidi, 2011). However, subjects such as ITP, ITPOL and ITSA had shown lower percentage of hit, with ITSA being the lowest at 28%. Overall, the hit percentage across all courses was 44%, which is considered low, indicating a general inconsistency between LE and DI. This is probably due to several factors including students' incomprehension of topics and weak fundamentals onto those topics being assessed which HS lecturers are unaware about. This is supported by Shikha and K. Subramaniam (2012) that awareness of students' thinking ability is an essential part of teacher education.

The RQ3 explored the distance between LE and DI. The distance was calculated by assigning numerical values to the differences in difficulty level which are (0) for no difference, (1) for one-level difference, and (2) for two-level difference. For example, if a question in ITC was rated as easy by LE and medium by DI, the distance was recorded as 1. Overall, courses like ITC and ITH exhibited lower mean distances (0.4), indicating closer alignment between LE and DI, while ITSA showed the highest mean distance (0.786), suggesting significant inconsistencies.

The RQ4 assessed the statistical significance of distance between LE and DI. The results found a statistically significant difference, highlighting an important division between lecturers' anticipation towards questions' difficulty and respond received from students. The possible reason is due to lecturers' inability to anticipate what students are likely to think and what students will find confusing in answering the questions. The results are contradicted with the study done by Llinares, Fernandez and Sanchez-Matamoros (2016) that anticipation of the possible responses from students with different characteristics of conceptual understanding is crucial for teachers in teaching practice.

The RQ5 evaluated the differences between courses in distance between LE and DI. The findings specified notable differences among HS programme courses, indicating that courses should not be treated uniformly in terms of agreement between lecturers' estimation and actual difficulty perceived by students. According to Post Hoc tests, specifically the Games-Howell comparison, a notable difference was found only between ITC and ITP. This is probably due to ITP having large number of questions compared to other courses, which influencing the perceived difficulty discrepancies. It is believed that different ways of designing questions and different numbers of questions across HS programme courses led to the significant distance between LE and DI.

To sum up, the results from the RQ1 - RQ5 suggest that there may be a need for a more comprehensive approach to question design and difficulty calibration. This could involve piloting questions with a sample group of students to gauge their perceived difficulty before including them in formal assessments. Furthermore, the different numbers and types of questions across courses highlight the need for a balanced assessment approach that suits each course's specific goals and content. While multiple-choice questions (MCQs) are good for covering a lot of material, using a variety of question types can better assess students' higher-order thinking skills.

Limitations

Due to the limited size of our sample, this study was not able to generate results that adequately represent the broader population, leading to general findings for this research. Consequently, it is imperative for future researchers to focus on expanding the pool of participants. By increasing the number of respondents, we can enhance the reliability and generalizability of our results, thereby enabling more robust conclusions. Furthermore, the study's analysis of the 'Easy', 'Medium', and 'High' categories lacks precision, making it challenging to accurately depict the true nature of the data.

Implications Of the Research

The use of data from the Test Specification Table (TST) and Difficulties Index (DI) as variables to assess fairness in educational evaluations introduces a novel theoretical framework that enhances our understanding of how these elements influence the equity of assessments. Practically, this analysis can be seamlessly integrated into current practices by leveraging existing departmental resources, thereby allowing educational institutions to refine and improve their assessment methods without the need for significant additional investments. Methodologically, the innovative approach of measuring the distance between TST and DI, a concept not previously explored in academic research, opens new avenues for further investigation and validation, potentially establishing a new standard for evaluating the effectiveness and fairness of educational assessments.

5. Conclusion

The study found that while the test-specification tables (TST) suggested a balanced distribution of easy, medium, and hard questions as required by lecturers for assessment planning, the difficulty index (DI) revealed that many questions were categorized as easy or medium. Overall, 44% of the questions showed a match between lecturers' estimations of difficulty (LE) and DI. The mean distance between LE and DI across all five courses was 0.658, with values ranging from 0 to 2, indicating a generally low level of inconsistency and high consistency between LE and DI. However, a one-sample T-test showed that the mean difference of 0.658 was significantly greater than zero, suggesting a significant misalignment between lecturers' estimated question difficulty and the actual difficulty experienced by students. Additionally, an ANOVA test revealed significant differences between courses ($p < 0.05$), with a post-hoc analysis highlighting a notable discrepancy between the ITC and ITP courses in terms of the distance between LE and DI. From these findings, the study concluded that there is a notable difference between LE and DI for questions used in HS course. Thus, study will provide insight to the educators and exam questions provider in ensuring the precise calibration of question difficulty is essential for creating equitable assessments that accurately reflect students' abilities, offering fair opportunities for success while enhancing the validity of educational evaluations and fostering more inclusive, reliable assessments across diverse learning contexts.

Acknowledgments

The researchers extend their gratitude to the Department of Human Sciences and the Centre for Foundation Studies at the International Islamic University Malaysia. They would like to express their appreciation to all individuals who provided support, whether directly or indirectly, in the completion of this project.

References

Al-Khuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (n.d.). A Systematic Review of Data-Driven Approaches to Item Difficulty Prediction. Springer Link.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (2012). Influence of behavior perceptions and academic skills on teachers' judgments of students' academic achievement. Journal of Educational Psychology, 104(3), 756-775. https://doi.org/10.1037/a0027958

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. Review of Educational Research, 86(4), 803-848. https://doi.org/10.3102/0034654316672069

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 221-256). Praeger.

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. Cengage Learning.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302.

Downing, S. M., & Yudkowsky, R. (2009). Assessment in health professions education. Routledge.

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. Practical Assessment, Research, and Evaluation, 18(1), 3. https://doi.org/10.7275/cztt-7109

Gipps, C., Stobart, G. (2009). Fairness in Assessment. In: Wyatt-Smith, C., Cumming, J.J. (eds) Educational Assessment in the 21st Century. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-9964-9_6

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. Applied Measurement in Education, 12(1), 13–28.

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. Routledge.

Hambleton, R. K., & Jones, R. W. (1993). Educational Measurement: Issues and Practice, 12(3), 38-47.

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.

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

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory Into Practice, 41(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Lane, S. (2014). Validity evidence based on testing content. Psicothema, 26(1), 100-107.

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing? Instructional Science, 41(3), 621-634.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Instrument development in the affective domain. Springer.

McMillan, J. H. (Ed.). (2013). SAGE handbook of research on classroom assessment. SAGE Publications.

Nathan, M. J., & Petrosino, A. (2003). Expert blind spot among preservice teachers. American Educational Research Journal, 40(4), 905-928

Nitko, A. J., & Brookhart, S. M. (2020). Educational assessment of students (8th ed.). Pearson.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

Oermann, M. H., & Gaberson, K. B. (2021). Evaluation and testing in nursing education (6th ed.). Springer Publishing Company.

Rafatbakhsh, E., Ahmadi, A. Predicting the difficulty of EFL reading comprehension tests based on linguistic indices. Asian. J. Second. Foreign. Lang. Educ. 8, 41 (2023). https://doi.org/10.1186/s40862-023-00214-4

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 307-353). Praeger Publishers.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. Journal of Educational Psychology, 104(3), 743-762. https://doi.org/10.1037/a0027627

Thorndike, R. M., & Thorndike-Christ, T. (2010). Measurement and evaluation in psychology and education (8th ed.). Pearson.

Van de Watering, G., & Van der Rijt, J. (2006). Teachers’ and students’ perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. Educational Research Review, 1(2), 133–147. https://doi.org/10.1016/j.edurev.2006.05.001

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. Language Arts.