New Metrics for Analysing Multiple-choice Questions: A Window into Examination Design and Curriculum Alignment

Joshua Pople, Di Warren and Rashika Agarwal

Corresponding author: Diana Warren (diana.warren@sydney.edu.au)
School of Mathematics and Statistics, University of Sydney, Sydney NSW 2006, Australia

Keywords: multiple-choice questions, curriculum alignment, discrimination

Abstract

While the ideal of constructive alignment in curriculum is well established, and the importance of evaluating learning and teaching is well known, evaluating assessments remains a complex task and gaps can arise between learning outcomes, learning activities and assessments.

Our study outlines an innovative way of analysing multiple-choice question (MCQ) examinations, which reveals possible weaknesses in the examination design and gaps in the alignment of curriculum. Individual examination questions are analysed through traditional metrics of the Discrimination Index (DI) and Difficulty Index (DIFF), combined with two novel metrics called the Grade Inversion Score (GIS) and Association with Total Score (ATS).

Focusing on examination marks from a data science examination from The University of Sydney, we perform two investigations. First, we identify poorly designed questions using DI, DIFF, GIS and ATS. Second, as multiple questions arise as deficient in these metrics, we explore specific areas in the curriculum where there may be misalignment of course material with the learning outcomes.

Our analysis provides a simple visual way for examiners to inspect the validity of an examination design and encourages an evidence-based approach to exploring curriculum alignment using multiple-choice assessments.

Introduction

The principle of constructive alignment has been well established as best practice in curriculum design, since the seminal works of Tyler (1949) and Biggs (1996, 1999, 2014), and the corpus of literature that has followed. In practice, there are many challenges to implementing and maintaining constructive alignment, especially in a large first-year cohort, with team teaching and a culture of continual improvement. Moreover, evaluating constructive alignment is challenging. For example, writing a final examination paper that is aligned with curriculum, while remaining interesting and discriminating, is a complex task.

The importance of evaluating teaching and learning is clear in the literature, with Kelder, Carr & Walls (2017) arguing for a “Curriculum Evaluation and Research (CER) framework” to establish a “scholarly regime” for evaluation. However, there appears to be less work on concrete methods for evaluating assessment, with respect to curriculum design. For example, recent work by Kirschner, Henrick & Heal (2022), surveys 30 seminal works under the following six separate sections - (1) Teacher Effectiveness, Development, and Growth, (2) Curriculum Development/Instructional Design, (3) Teaching Techniques, (4) Pedagogical Content Knowledge, (5) In the Classroom, and (6) Assessment - with no suggested tools for evaluating the integration of (2) with (6).

In Chemistry Education, Schmid et. al. (2016) proposes a tool to evaluate the use of Chemistry Threshold Learning Outcomes (CTLOs) in assessment tasks, which involves a lengthy iterative process of review by academic peers. Interestingly, their study revealed that "faculty over-estimate the ability of their assessment items to confirm achievement of CTLOs", suggesting the need to refine their work with data capturing students' performance in the tasks.

Multiple-choice assessments

Irrespective of curriculum design, two well established paired constructs for evaluating an assessment are reliability and validity. Although reliability has an established statistical framework - including test-retest reliability, inter-rater reliability, classical test theory, item response theory, Cronbach's alpha, and Kuder–Richardson Formula 20 - the notion of validity is harder to assess, with added complexities such as construct underrepresentation or irrelevance (APA, 2020).

> *Content-oriented evidence of validation is at the heart of the process in the educational arena known as alignment, which involves evaluating the correspondence between student learning standards and test content. Content-sampling issues in the alignment process include evaluating whether test content appropriately samples the domain set forward in curriculum standards, whether the cognitive demands of test items correspond to the level reflected in the student learning standards (e.g., content standards), and whether the test avoids the inclusion of features irrelevant to the standard that is the intended target of each test item.* (AERA, 2014)

In investigations of multiple-choice assessments, the most common metrics are the Difficulty Index (DIFF) and the Discrimination Index (DI) (Salkind, 2017).

For each question in a multiple-choice examination, the DIFF focuses on the proportion of the cohort who answered the question correctly and has been used to assess question appropriateness and whether a question should be omitted from a future examination (Hingorjo & Jaleel, 2013). Mahjabeen et al. (2018) propose the following schema: if the DIFF of a question is below 30% then it is too difficult, if above 70% then the question is too easy, and anything between 30% and 70% is reasonable. However, if an examination is designed to differentiate between different abilities, with questions intentionally set at different levels, this schema becomes less informative.

Other studies use the DI, which seeks to measure how well a question distinguishes between higher and lower achieving students. The DI has been used to investigate MCQ design and areas of improvement (Hingorjo & Jaleel, 2013; Dixon, 1994). There are different versions of the DI, which typically compares the difference in proportions of correct answers between the lowest 50% and highest 50% of students based on their total score (Salkind, 2017), but it can be generalised into the point-biserial correlation coefficient (Essen & Akpan, 2018). It is commonly accepted that a question with a DI of less than 0.2 is deemed to have poor discriminating power (Belay et al., 2022; Taib et al., 2014; Mahjabeen et al., 2018). Again, as a decision-making tool, use of the DI can be confounded with the intended level of the question by the examination setter, and the categorisation of the cohort into two parts is somewhat artificial at the boundaries– that is the bottom student in the top group and the top student in the bottom group may have very similar performances.

Given the constraints of using and interpreting the Difficulty and Discrimination Indices alone, we suggest two ways ahead:

(1) To investigate **combinations** of the metrics, set in the context of the marker's intent for each question. Following Warren (2023), we study a simple and informative visualisation using three variables: the Difficulty Index, the Discrimination Index, and the Marker Grade.
(2) To develop **new statistics,** which are based on multiple aspects of the data, and hence more finely tuned to how performance on an individual question relates to the student's Total Score. We propose two new metrics - the Grade Inversion Score and the Association with Total Score, and methods based on them.

Our analysis aims to provide supporting evidence for evaluating MCQ examinations, and to allow setters to use the results to consider gaps in curriculum alignment.

Methodology

Context and Data

Our study focuses on a first-year data science unit (DATA1001: Foundations of Data Science) at the University of Sydney with a large, diverse cohort. The unit has 10 learning outcomes (www.sydney.edu.au/units/DATA1001/2023-S1C-ND-CC) covering experimental design, modelling data, sampling data and hypothesis testing, addressing both statistical theory and computational skills. In what follows, learning outcome x is referred to as LOx.

In Semester 1 2023, there were N=1234 students sitting the final examination. The final examination was worth 60% of the students' overall grade, with 50% of the examination mark comprising of 20 multiple-choice questions (Q1-Q20). Each question had four options (a, b, c, d) with a single correct answer.

When the MCQs were being written, the examiner assigned a **Marker Grade** to each question, which indicated their assessment of its expected difficulty for students. The Marker Grade was based on the examiner's academic judgment of the knowledge level of each question, due to their experience of teaching the unit for over five years. Additionally, the Marker Grade was later tested against the team of markers, who were asked to independently assign a grade to each MCQ, based on their academic judgment from tutoring the unit.

The Marker Grade consisted of four categories as follows: P: Pass (easiest), CR: Credit, D: Distinction and HD: High Distinction (hardest). Table 1 shows the number of questions for each category, where half of the questions were designed to be at the Pass level, while only a few were HD level.

Table 1 – Distribution of Marker Grades in MCQs

Marker Grade	**Pass (P)**	**Credit (CR)**	**Distinction (D)**	**High Distinction (HD)**
Number of Questions	10	6	2	2

The initial data from the examination consisted of a 1234 x 20 matrix, containing each student's answers (rows) for each MCQ (columns). In addition, we added a quantitative and a qualitative variable.

- The **Total Score** for each student was calculated, which was an integer ranging from 0 (no correct answers) to 20 (all correct answers); our dataset had a minimum of 2.
- Each student was assigned a **Student Grade** depending on their **Total Score,** as outlined in Table 2.

Table 2 - Student Grade categorisation based on Total Score

Total Score (as a %)	**0 - 49**	**50 - 64**	**65 - 74**	**75 - 84**	**85 - 100**
Student Grade	F	P	CR	D	HD

The resulting data was a 1236 x 22 matrix, as seen in Table 3, with

- Row 1: the **Correct Answers** for each MCQ (a, b, c, d).
- Row 2: The **Marker Grades** for each MCQ (P, CR, D, HD).
- Rows 3-1236: the anonymized matrix of student answers for each MCQ (Q1-20), Total Score and Student Grade.

Table 3 - Example MCQ examination data for DATA1001 S1 2023

ID	**Q1**	**...**	**Q20**	**Total Score**	**(Student) Grade**
Correct answer	d		b		
Marker Grade	P		CR		
Student 1	d		b	14	CR
...					
Student 1234	d		b	10	P

Four Metrics

Using the examination data, we construct four metrics. First, two traditional indices (DIFF, DI) were considered.

Using the **Total Score,** the cohort was ranked from 0 to 1234, with multiple students on each level ranking. For example, there were two students scoring 20 and none scoring 0. The students were then split in half, representing the high and low achieving groups. These groups were used to calculate some of the metrics in this analysis.

The Difficulty Index (DIFF, or Actual Level)

The Difficulty Index is the percentage of students who answered the question correctly. It represents the 'Actual Level' of the cohort for a particular question. For example, Q1 had a DIFF of 87.52%.

N_T = Number of students in the top 50% that answered the question correctly
N_L = Number of students in the bottom 50% that answered the question correctly
N = Total number of students

$$\text{Difficulty Index (\%)} = \frac{N_T + N_L}{N} \times 100$$

Following Mahjabeen et al. (2018), the DIFF should be investigated if outside the range (30%, 70%).

The Discrimination Index (DI)

The Discrimination Index is the difference between the number of students who answered the question correctly in the top 50% of the cohort and the bottom 50% as a scaled proportion of the size of the cohort.

N_T = Number of students in the top 50% that answered the question correctly
N_L = Number of students in the bottom 50% that answered the question correctly
N = Total number of students

$$\text{Discrimination Index} = \frac{N_T - N_L}{0.5 \times N}$$

The DI ranges from –1 and 1 and is commonly categorised as per Table 4.

Table 4 - Discrimination Index with its categorisation (Mahjabeen et al., 2018)

Discrimination Index	< 0.2	0.2 – 0.24	0.25 – 0.35	> 0.35
Categorisation	Poor	Acceptable	Good	Excellent

A score of 0 indicates that the question did not discriminate between the two groups (top 50% and bottom 50%), and a negative score denotes that a higher proportion of students answered the question correctly in the bottom 50% compared to the top 50%. At the extremes, a score of –1 and 1 indicates a very poorly designed question and a highly discriminating question, as they were only answered correctly by the bottom or top 50% of the cohort respectively. Note that good and excellent discriminating questions paired with a suitable DIFF should reflect a correctly functioning question in the absence of other confounding variables. For example, Q15 had a DIFF of 58.27% and a DI of 0.40, making it an excellent question in the examination.

Next, two new indices were developed.

The Grade Inversion Score (GIS)

The Grade Inversion Score is based on the proposition that students' abilities can be reflected through their Total Score categorised into grades (F, P, CR, D, HD), and that as the academic level increases, the proportion of correct answers in each grade category (for a particular question) should increase.

For each of the 20 MCQs, we calculate the percentage of correct answers across the five levels of Student Grade, as seen in Table 5 for Q1 and Q17.

Table 5 - Grade categorisation for Q1 and Q17

Student Grade	F	P	CR	D	HD
Correct answers (%) for Q1	79.61	91.15	90.05	90.51	100.00
Correct answers (%) for Q17	19.90	21.13	13.74	24.82	41.79

Given the natural order of F, P, CR, D, HD, we say an "inversion" has occurred if the proportion of correct answers in a lower Grade is higher than that of a higher Grade. For example, an inversion would be said to have occurred if 50% of the "F" cohort answered the question correctly compared to only 40% of the "P" cohort. More formally: Given an array A, let A[i] be the value at index i. An inversion is when A[i] > A[j] such that i < j. We count the inversions across the 10 possible grading pairs (F/P, F/CR, F/D ... D/HD).

The GIS is the number of inversions present for a selected question. The GIS ranges between 0 to 10, which represents an ascending and descending order of proportions of correct answers from the "F" to "HD" groups respectively. The GIS is "high" when the total inversions are greater than or equal to 2, as the question is performing differently to the assumed ascending trend. For example, the GIS for Q1 and Q17 were 2 and 3 respectively.

Association with Total Score (ATS)

The Association with Total Score (ATS) metric uses a one-sided two sample t-test for the mean Total Score (excluding Question Score) of students who answered the question correctly against those who answered incorrectly. The ATS is premised on the assumption that a student answering a question correctly has a higher academic ability than a student who answers incorrectly, and thus should associate with a higher Total Score. Hence, a significant p-value is expected to be produced from all questions, indicated by a YES result. When this is not the case, that question is flagged for investigation (NO result).

Given ATS uses a one-sided t-test, all assumptions for a t-test must hold. A Bonferroni adjustment is also made, considering 20 t-tests were conducted for this analysis, so that the 5% significance level effectively becomes 0.25%.

More formally, given that μ_C and μ_{IC} denote the mean Total Score of students who answered a question correctly and incorrectly respectively, we are testing the null hypothesis, H_0, that both means are the same against the alternate hypothesis, H_1, that μ_C is greater than μ_{IC}.

Including the four metrics, results in the following 1240 x 22 data matrix (Table 6).

Table 6 - MCQ examination data, with 4 indices for Q1 and Q20

ID	**Q1**	**...**	**Q20**	**Total Score**	**(Student) Grade**
Correct answer	d		a		
Marker Grade	P		CR		
Student 1	d		b	14	CR
...					
Student 1234	d		b	10	P
DIFF	87.52		58.83		
DI	0.09		0.29		
GIS	2		0		
ATS	YES		YES		

Methods used for the post-completion analysis of MCQ items

To consider the relationship between different variables and indices, and what insights they bring concerning individual questions, we suggest three statistical methods:

(1) A visualisation combining three variables: the Difficulty Index (DIFF), the Discrimination Index (DI) and the Marker's Grade.
(2) A statistical analysis of Question Score vs Total Score (ATS).
(3) A statistical analysis of Grade Inversion Score (GIS).

Results

Method 1: The Discrimination Index (DI) vs The Difficulty Index (DIFF, or Actual Level) across Marker Grades.

Figure 1 is a simple visualisation proposed by Warren (2023), which shows the relationship of three variables: Discrimination Index (DI), the Difficulty Index (DIFF) and Marker Grades. Note here we use the terminology "Actual Level' for the DIFF to make clearer the comparison with the "Expected Level" designated by the Markers Grade. The plot helps identify MCQs which are not functioning as expected in the examination – that is, which questions are functioning differently to what the examiner expected.

Figure 1 shows that although Q4 and Q10 were set to be Pass level questions (green), less than 25% of students answered them correctly (see x-axis). In contrast, Q13 was set to be a Distinction level question (purple) and was answered correctly by more than 60% of students. Q4 and Q10 could be investigated for why they were harder in practice than the setter intended, but as Q10 was effective in discriminating students (DI > 0.2), no further analysis was done on Q10.

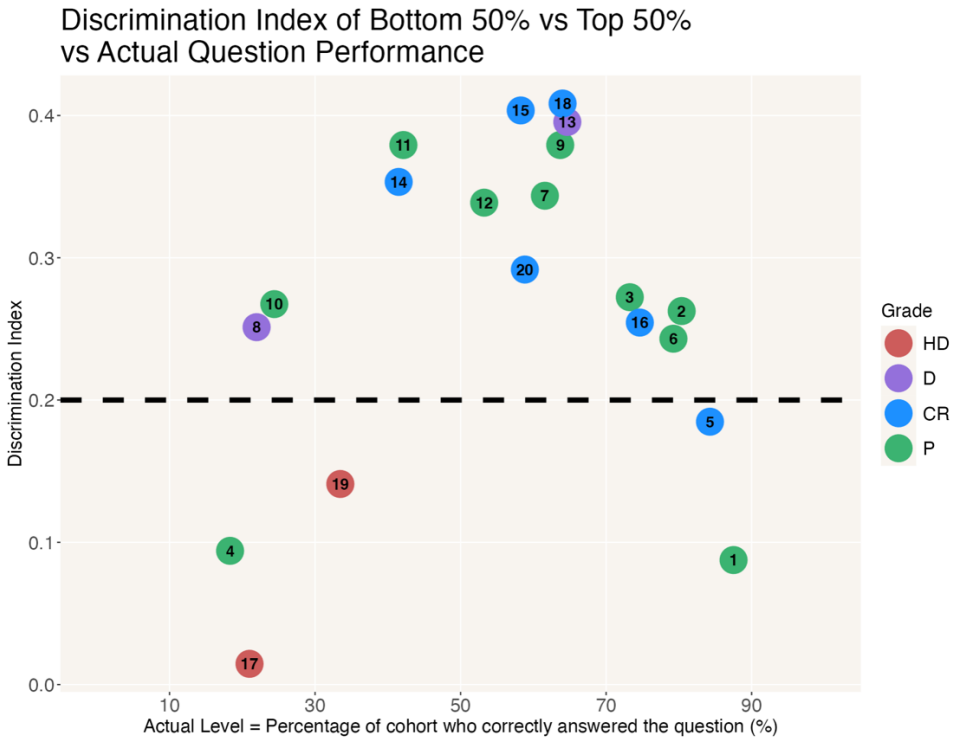


Figure 1 - Relationship between Marker Grade/Expected Level (colour legend), Difficulty Index/Actual Level (x axis) and Discrimination Index (y axis) for a 20 MCQ examination

The mean/standard deviation (SD) of the DI for the MCQs was 0.27/0.11. Questions identified as having a DI below 0.2 are Q1, Q4, Q5, Q17, Q19, which are summarised alongside the DIFF and Marker Grade in Table 7. These five questions also have a DIFF that is outside the acceptable range of 30% to 70%, except for Q19.

Table 7: Identifying questions with a poor Discrimination Index

Question	DI	DIFF (%)	Marker Grade	MCQ Investigation
1	0.09	87.52	P	Too easy, or functions as easy mark.
4	0.09	18.31	P	Too hard but categorised as P by the setter.
5	0.18	84.28	CR	Too easy but categorised as CR by the setter.
17	0.01	20.99	HD	Possibly too difficult.
19	0.14	33.47	HD	Functioning well as HD question.

Interrogating each question, we notice the following observations from Table 7.

Q1 is a Pass graded question with a DIFF of 87.52%. The high DIFF of this question limits its ability to discriminate between the top 50% and bottom 50% of students, thus leading to a DI of 0.09. Since this question had a DIFF > 70%, it should be investigated for testing trivial content, having inefficient distractors, or whether it is purposely functioning as an easy starter question for all students. Alternatively, the concept assessed by Q1 could have been grasped well by this cohort or covered more in depth in this iteration of the course. Similarly, Q5 has a high DIFF (84.28%), which is well above 70% - this contributes to its lack of ability to discriminate (0.18), but interestingly it was categorised Credit level by the examiner (Marker Grade).

In contrast, Q4 was the most poorly answered question in the examination with a DIFF of 18.31% which is a striking mismatch with the Pass designation by the examiner. Again, it is not able to discriminate (DI of 0.09).

Q17 and Q19 both have a low DI and DIFF, with Q17 having a Difficulty Index of below 30%. A DIFF of 20.99% is intuitive here because it is a HD graded question, however, it could be argued that this is too difficult. A possible confounding factor is that questions that are very difficult may lead to many students guessing the answer, which contributes to a low Discrimination Index.

Method 2: Association with Total Score: Comparing mean Total Score of the cohort divided into correct and incorrect answers (ATS)

Using the ATS testing framework, we look for insignificant p-values (we use p-value > 0.0025 with the Bonferroni adjustment) to identify questions that do not have better overall student performance – excluding the question score - for the correct cohort compared to the incorrect cohort, which is deemed unusual. All the assumptions for the one sided two-sample t-test were satisfied, namely each student was independent from one another, and diagnostic plots showed that both groups had similar variance. Due to the sufficiently large cohort, the central limit theorem was relied on to satisfy the normality assumption.

We find that all questions yielded a significant p-value, except Q17 and Q19, as shown in Table 8. Note the p-value for Q19 is very close to the threshold (0.0025), hence our main interest is Q17.

Studying Q17, we find that the mean Total Score for students who answered correctly and incorrectly was 10.56 and 10.92 respectively, which goes against the assumed trend. The null hypothesis is hence retained (given p-value = 0.934) and we conclude that there is no statistical evidence that answering Q17 correctly contributes to a higher mean Total Score. This result is further affirmed by comparing Q17 to Q19, which was also classified as an HD level question by the examination setter. Q19 had an insignificant p-value for ATS (p-value=0.003), despite the mean Total Score for the correct cohort being only slightly larger than the incorrect cohort.

As Q17 has an insignificant association with Total Score, this suggests that academic ability (Total Score) does not influence a student's ability to answer this question correctly. An intuitive deduction would be that most students guessed the answer to this question.

Table 8 - Identifying questions with an insignificant p-value (Q17, Q19) based on ATS

Question	p-value	Mean Total Score (Correct Cohort)	Mean Total Score (Incorrect Cohort)
17	0.934	10.56	10.92
19	0.003 (close to threshold)	11.1	10.54

Method 3: Grade Inversion Score (GIS)

Studying the GIS allows us to find anomalies in how a particular question is functioning, in terms of the different Student Grades in the cohort – that is, how do the HD students perform compared to the D students? A GIS of 0 or 1 indicates that the question is functioning as expected.

Table 9 shows the three MCQs that had a GIS of 2 or above – namely, Q1, Q2 and Q17.

Table 9: Identifying questions with a high GIS

Question	GIS
1	2
2	2
17	3

Q1 had very slight differences in the proportions between P, CR and D students - namely 91.15%, 90.05%, 90.51% respectively - and thus resulted in a GIS of 2 due to chance error. Due to the high DIFF of this question (see Table 7), the ordering of proportions is largely inconsequential indicating that no further investigation was needed regarding curriculum

alignment issues. Q2 produced similar results with the CR, D, and HD groups having the proportions 95.26%, 94.89% and 94.03%, requiring no further analysis.

Q17 emerged as having a GIS of 3. Further investigation in Figure 2, which visualises the proportions of correct responses within each grade group, reveals that the percentages of correct responses for F and P students were 19.90% and 21.13%, compared to 13.74% for CR students. Furthermore, D level students had a proportion of 24.82%, which is a small increase from the lower groups, suggesting that this question was answered poorly by many students across all academic levels. Note, if students from all Grade categories guessed randomly then proportions of roughly 25% should be seen in all categories. In contrast, the other HD graded question (Q19) had no inversions - despite being a difficult question (DIFF = 33.47%), it displayed the assumed ascending trend of grade proportions. This further indicates that Q17 should be reviewed.

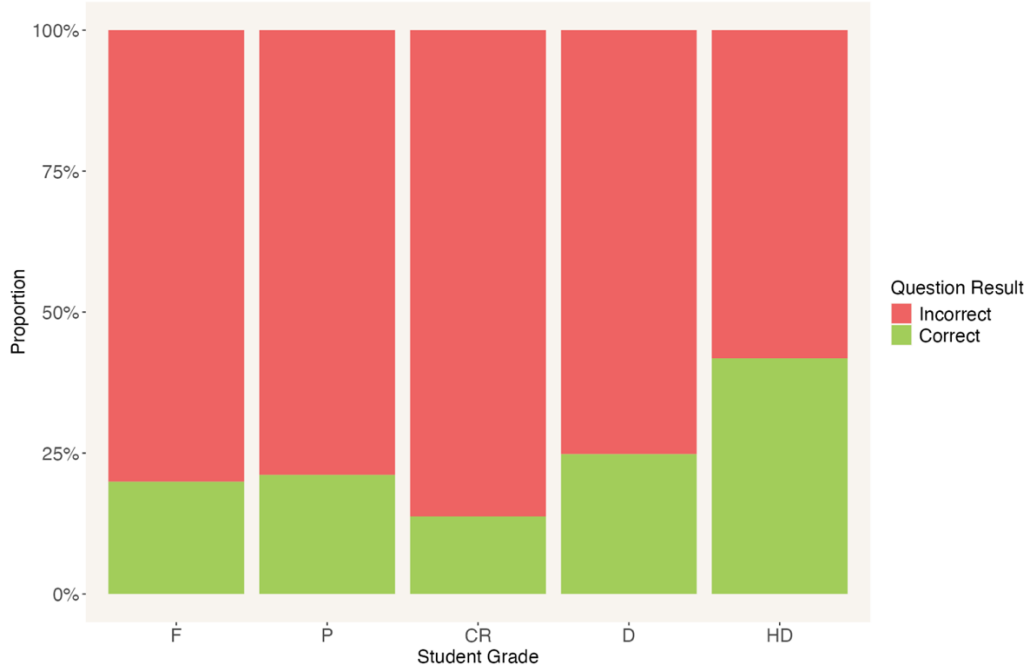


Figure 2 - Proportion of correct answers vs Student Grade categories for Q17

Discussion of curriculum alignment

Combining the results leads to Table 10, which flags the MCQs to be investigated regarding curriculum alignment. Four questions emerge for attention – namely Q1, Q4, Q5 and Q17, with Q17 appearing in all three methods.

Table 10 - Identified and flagged questions using the 3 methods

Method	Questions investigated	Questions flagged for investigation of curriculum alignment
1	1, 4 ,5, 17, 19	1, 4, 5, 17
2	17, (19)	17
3	1, 2, 17	17

Q1 and Q5 were both very easy questions (see Figure 3) with a DIFF of 87.52% and 84.28% (see Table 7), despite Q5 being designated a Credit level question by the examination setter. A very large DIFF could reveal gaps in curriculum alignment as it suggests that the question did not align well with the common areas of confusion.

Q1 related to the fundamental concept of challenges faced when working with data (LO1). The three complexities presented as alternatives (see Figure 3) were all sensible and equally valid, making "All of the other answers" a natural choice for most students. The exam setter also confirmed that this question was not trivial content, but rather meant to be easy for students.

Q5 assessed the students' ability to interpret numerical summaries of quantitative data (LO3). The question did not simply test recall of factual knowledge considering negative standard deviation was not covered in the lecture materials which affirms its CR level grading. Students were expected to apply the definition of these numerical summaries and choose the best answer.

In summary, both Q1 and Q5 did not present any curriculum gaps. However, Q5 indicates that students understand properties of common numerical summaries and hence similar future questions on this learning outcome can be Pass graded.

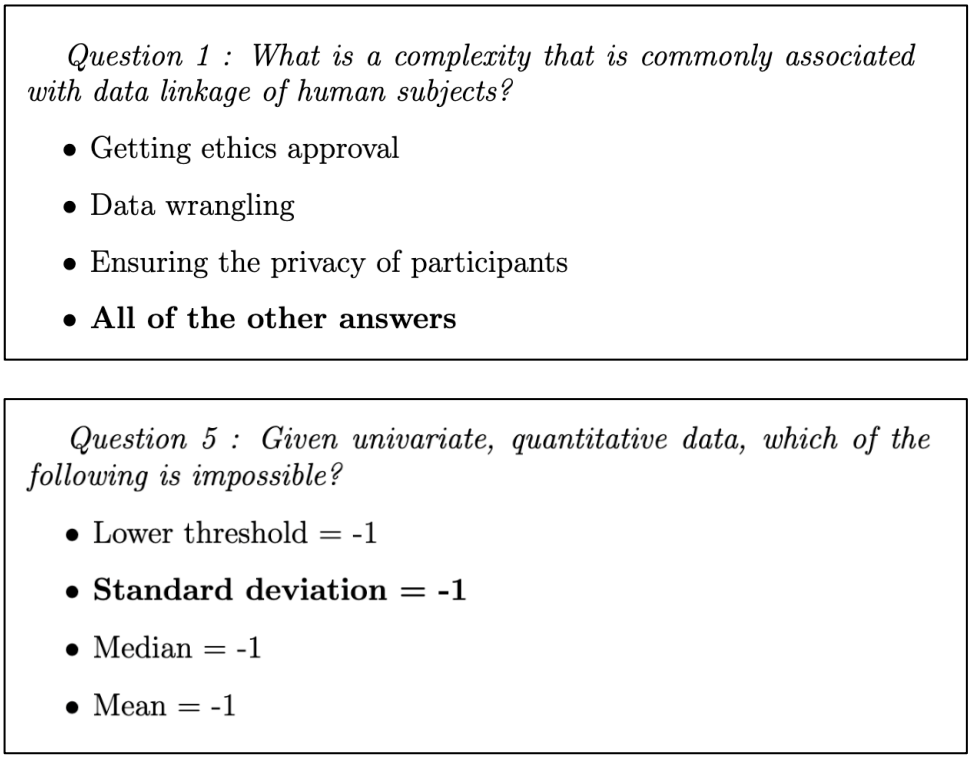


Figure 3 - Q1 and Q5 from the DATA1001 Final Examination 2023 S1

Q4 (see Figure 4) emerged as having the lowest DIFF (18.31%) of all questions, despite being designated a P level question by the examiner (see Table 7). A contradiction between the Marker Grade and DIFF could be a strong indication of a gap in curriculum design among other possible confounders.

Investigation into the learning outcome that Q4 assessed (namely LO3, and associated learning outcomes) found that this question presented an abstraction on the topic of standard deviation. While the question could be solved algebraically, the effect of scaling data on statistics was not

explicitly covered in lectures or tutorials in 2023 S1, and hence could possibly explain why students performed poorly.

Checking this hypothesis against the 2024 S1 cohort, who had a similar MCQ in their final examination, but who also had new learning activities to specifically address the gap in curriculum, we found that 75.29% of the students answered the question correctly, which is consistent with a P level question.

Question 4 : A company decreases all of their food prices by 2%. By how much will the mean and standard deviation of food prices change, respectively?

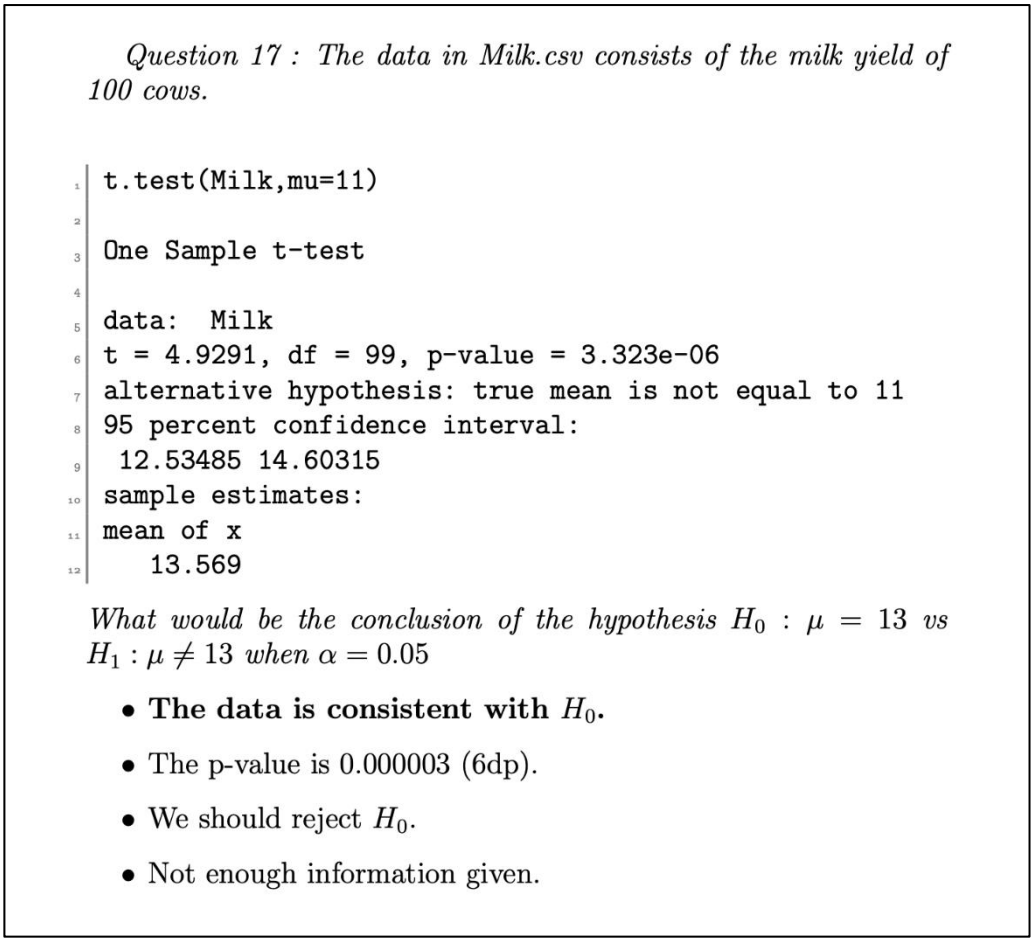
- 0% and 2%
- 2% and 0%
- **2% and 2%**
- 2% and 4%

Figure 4 - Q4 from the DATA1001 Examination 2023 S1

Q17 emerged as deficient in all metrics (DI = 0.01, DIFF = 20.99%, GIS = 2, ATS = NO) strongly suggesting that further investigation needed to occur. To begin with Q17 covered content from the last and most difficult 'capstone' module of the course, as "students regard Hypothesis Testing as the most important and the most difficult threshold concept" (Swanepoel, Engelbrecht, Harding & Fletcher, 2015; p.1), consistent with the seminal GAISE (Carver et.al, 2016) findings.

The structure of Q17 (see Figure 5) required students to interpret *R* output to answer the question. Part of the output included extensively covered course content (p-values) which was unrelated to the specific question. It seems that students from all levels saw this information and wrongly used it to answer the question, contributing to a low score across all academic levels.

Whilst being answered poorly by students, Q17 involved concepts that were all covered in lecture material, albeit separately, and thus did not present a gap in curriculum alignment. Rather it requires a critical synthesis of multiple learning outcomes. Hence, in terms of curriculum alignment it raises two questions: (1) how can the connections between multiple learning outcomes be better reinforced in tutorial activities, and (2) what is the place of a high-level discriminator which assesses synthesis rather than applications of learning activities? More specifically, should synthesis of concepts be better incorporated in further iterations of the course, or is this type of question solely functioning as a high-level discriminator?



Question 17 : The data in Milk.csv consists of the milk yield of 100 cows.

```
t.test(Milk,mu=11)

One Sample t-test

data:  Milk
t = 4.9291, df = 99, p-value = 3.323e-06
alternative hypothesis: true mean is not equal to 11
95 percent confidence interval:
 12.53485 14.60315
sample estimates:
mean of x
   13.569
```

What would be the conclusion of the hypothesis $H_0 : \mu = 13$ *vs* $H_1 : \mu \neq 13$ *when* $\alpha = 0.05$

- **The data is consistent with H_0.**
- The p-value is 0.000003 (6dp).
- We should reject H_0.
- Not enough information given.

Figure 5 - Q17 from the DATA1001 Examination 2023 S1

Summary: Using DI, DIFF, GIS and ATS, unsatisfactory questions in the MCQ examination were highlighted. Specifically, Q1, Q4, Q5 and Q17 emerged as being subpar and needed further investigation into question design and curriculum alignment. Each of these metrics provides different ways to identify potential flaws in question design and misalignments in the curriculum.

Since the questions flagged by the Association with Total Score (Q19, Q17) and the Grade Inversion Score (Q1, Q17) also appeared during Discrimination Index Analysis (Q1, Q17, Q19), our new metrics appear to complement the common DI metric.

Confounding Variables

There are many possible confounding variables, which can distort the findings. Firstly, variation in student ability between different cohorts can lead to the misgrading of questions by the examiner. For example, Q13 was denoted as a Distinction level question but 64.67% of the students answered it correctly. This gap between the examiner's expectation and the students' performance could arise from the 2023 S1 cohort being academically stronger than previous cohorts, or more specifically, understanding the concept tested by Q13 better than previous cohorts, which could be caused by a change in lecturer or other uncontrollable factors.

Secondly, the Marker Grade provided by the examiner may have an inherent bias, due to their assessment of student ability. However, deploying standards-based assessment should mean

that an academic's evaluation is based solely on the difficulty of the question, and the Discrimination Index provides one way of cross-checking the Marker Grade.

Thirdly, the way that a question is written (orthographic encoding), and distractor effectiveness, can strongly influence a student's ability to answer a MCQ correctly. For example, confusing language or culturally specific examples can discriminate against students from non-English speaking backgrounds. Additionally, using ineffective distractors can allow a high percentage of the cohort to answer a question correctly, even if the question is designated a high Marker Grade.

Limitations

Our analysis necessitates several limitations. Firstly, the Discrimination Index (DI), Difficulty Index (DIFF), Grade Inversion Score (GIS) and Association with Total Score (ATS) only identified potentially unsatisfactory questions in the MCQ examination. Neither the GIS nor ATS can assess good quality MCQs.

Secondly, the DI functions as if all questions are the same level – that is, when splitting the cohort up we do not consider which questions they answered correctly. This may be particularly pertinent for those scoring around 8 –14. This may result in some students being misclassified into the two groups, producing unreliable DI scores. Should a DI be developed that allows students who answered more D or HD questions to be ranked higher?

Thirdly, our current analysis is based on one examination with only 20 questions. Considering that there are 10 learning outcomes, on average there are two questions per learning outcome. Examinations from other iterations of this course will allow for further investigation of curriculum alignment of certain learning outcomes.

Fourthly, the current analysis only studies the MCQ questions. Evaluating the extended response questions could further refine the grading of students, and student understanding. All metrics used in the study will need to be modified to account for partial marks in extended response questions.

Conclusion and Further Work

Our analysis provides a simple visual way for examiners to inspect the validity of an examination design and encourages an evidence-based approach to exploring curriculum alignment using multiple-choice assessments.

Adding the new metrics (GIS and AIS) creates a suite of four metrics, which enables three statistical methods for analysing MCQ data. Each metric or method may identify questions for investigation, but those flagged multiple times clearly require special attention.

Examiners can use this simple and robust construct to analyse a desired MCQ examination and determine which questions need further investigation into their design where curriculum gaps can potentially be identified. All conclusions need to be nuanced by the limitations, confounders and uncontrollable factors.

In terms of further work, we suggest:
(1) Further investigation can be conducted into functional distractors along with DI, DIFF, GIS, and ATS, searching for any correlation between them. Additionally, distractor analysis can provide insight into the variability of difficulty among questions, and which groups of students are selecting these distractors. Considering that the four metrics were explored individually, it would be worth looking into relationships between them to see if either one contributes to another.

(2) Other subsets of the DI (e.g. Top 27% vs Bottom 27%) can be investigated to see if more MCQs emerge as weak discriminators. Furthermore, other calculations of the DI, like point-biserial correlation, can be explored to identify other questions needing investigating.

(3) Conducting more research on functional distractors and Discrimination Indices may allow for a more robust model for identifying questions which need investigation. This will ultimately lead to examiners reviewing more questions and evaluating how well their examination aligns with the curriculum.

(4) Investigation into the alignment of Marker Grade and the proportion of students in each Grade category that answered the question correctly. This could allow for insights into the validity of Marker Grades and suggest which questions have been Graded inappropriately. For example, if a P graded question is only answered correctly by 60% of P students, this may lead to reconsideration of the Marker Grade to a CR.

References

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

American Psychological Association. (2020). *APA Guidelines for Psychological Assessment and Evaluation. Retrieved from https://mindremakeproject.org/wp-content/uploads/2023/08/APA-Guidelines-for-*Psychological-Assessment-Evaluation.pdf

Belay, L. M., Sendekie, T. Y., & Eyowas, F. A. (2022). Quality of multiple-choice questions in medical internship qualification examination determined by item response theory at Debre Tabor University, Ethiopia. BMC Medical Education, 22(1). https://doi.org/10.1186/s12909-022-03687-y

Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher Education, 32(3), 347–364. https://doi.org/10.1007/bf00138871

Biggs, J.B. (1999, 2001). Teaching for Quality Learning at University. Maidenhead, Berkshire, England: Open University Press.

Biggs, J. (2014). Constructive alignment in university teaching. HERDSA Review of Higher Education, 1, 5-22.

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Rowell, G., Velleman, P., Witmer, J., & Wood, B. (2016). Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016. Retrieved from: https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports

Dixon, R. A. (1994). Evaluating and improving multiple choice papers: true-false questions in public health medicine. Medical Education, 28(5), 400–408. https://doi.org/10.1111/j.1365-2923.1994.tb02551.x

Essen, C., & Akpan, G. (2018). Analysis of difficulty and point-biserial correlation indices of 2014 Akwa Ibom State Mock Multiple Choices Mathematics Test. International Journal of Education and Evaluation, 4(5), 1-11. Retrieved from https://ijee.io/

Hingorjo MR, Jaleel F. (2013). The Difficulty Index, Discrimination Index and Distractor Efficiency. Pakistan Medical Association, *62*(2). Retrieved from https://www.researchgate.net/

Kelder, J.-A., Carr, A., & Walls, J. (2017). Evidence-based Transformation of Curriculum: a Research and Evaluation Framework. Retrieved from https://www.researchgate.net/

Kirschner, P., Hendrick, C., & Heal, J. (2022). *How teaching happens: Seminal works in teaching and teacher effectiveness and what they mean in practice*. London, England: Routledge.

Mahjabeen et al. (2018). Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions. Annals of Pims, *13*(4). Retrieved from https://www.apims.net/

Salkind, N.J. (2017) *Tests & Measurements for people who (think they) hate tests & measurements* (3rd ed., pp. 161-172). Thousand Oaks, Calif: SAGE Publications.

Schmid, S., Schultz, M., Priest, S. J., Glennys O'Brien, Pyke, S. M., Bridgeman, A., Lim, K. F., Southam, D. C., Bedford, S. B., & Jamie, I. M. (2016). Assessing the Assessments: Development of a Tool To Evaluate Assessment Items in Chemistry According to Learning Outcomes. ACS Symposium Series, 1235, 225–244. https://doi.org/10.1021/bk-2016-1235.ch013

Swanepoel, A., Engelbrecht, J., Harding, A., & Fletcher, L. (2015). Identifying and Evaluating Threshold Concepts in First Year Statistics courses at a large university in South Africa. Retrieved from https://2015.isiproceedings.org/Files/CPS469-P1-S.pdf

Taib, F., & Yusoff, M. S. B. (2014). Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. J*ournal of Taibah University Medical Sciences*, 9(2), 110–114. https://doi.org/10.1016/j.jtumed.2013.12.002

Tyler, R.W. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

Warren, D. (2023). Disrupting the Past Paper Pandemic - Developing New Question Banks. *Proceedings of the Australian Conference on Science and Mathematics Education (2023)*, Perth: Australian Council of Deans of Science.