Gaining Insights into Course Difficulty Variations Using Item Response Theory

Frederik Baucks*
Ruhr-Universität Bochum
Bochum, Germany
frederik.baucks@ini.rub.de

Robin Schmucker*
Carnegie Mellon University
Pittsburgh, PA, USA
rschmuck@cs.cmu.edu

Laurenz Wiskott
Ruhr-Universität Bochum
Bochum, Germany
laurenz.wiskott@ini.rub.de

ABSTRACT

Curriculum analytics (CA) studies curriculum structure and student data to ensure the quality of educational programs. To gain statistical robustness, most existing CA techniques rely on the assumption of time-invariant course difficulty, preventing them from capturing variations that might occur over time. However, ensuring low temporal variation in course difficulty is crucial to warrant fairness in treating individual student cohorts and consistency in degree outcomes. We introduce item response theory (IRT) as a CA methodology that enables us to address the open problem of monitoring course difficulty variations over time. We use statistical criteria to quantify the degree to which course performance data meets IRT's theoretical assumptions and verify validity and reliability of IRT-based course difficulty estimates. Using data from 664 Computer Science and 1,355 Mechanical Engineering undergraduate students, we show how IRT can yield valuable CA insights: First, by revealing temporal variations in course difficulty over several years, we find that course difficulty has systematically shifted downward during the COVID-19 pandemic. Second, time-dependent course difficulty and cohort performance variations confound conventional course pass rate measures. We introduce IRT-adjusted pass rates as an alternative to account for these factors. Our findings affect policymakers, student advisors, accreditation, and course articulation.

CCS CONCEPTS

• **Applied computing** → *Education*; *Learning management systems.*

KEYWORDS

curriculum analytics, item response theory, fairness.

ACM Reference Format:
Frederik Baucks, Robin Schmucker, and Laurenz Wiskott. 2024. Gaining Insights into Course Difficulty Variations Using Item Response Theory. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3636555.3636902

1 INTRODUCTION

Student grade point average (GPA) scores are a summary of *individual* course grades and are viewed as a measure of students' performance in an educational program. Notably, GPA scores play a central role in decision processes of employers and academic institutions and are known to be correlated with students' future career success (e.g., [16, 28]). This makes the monitoring and controlling of potential course difficulty variations–that can affect student GPAs–an important task for policymakers of academic and professional degree programs. It is essential for ensuring fairness in treatment of individual student cohorts and consistency in GPA scores.

The field of Curriculum Analytics (CA) studies educational program structure and student data to assess the quality of individual courses inside a curriculum and to provide insights to various stakeholders (e.g., program administrators and student advisors). Still, CA methodologies for analyzing variations that can occur inside programs over time are currently underexplored. For example, existing CA approaches that rely on process mining and simulation techniques to monitor student activities inside a curriculum struggle with issues of concept drift [10] because they are unable to capture differences between *individual offerings* of the *same course* (e.g., CS1 in winter 2020 and CS2 in winter 2021), that can occur over time and that can result in changes in the process while being measured. Similarly, CA approaches that make curriculum structure-based predictions assume a stationary data generation process and are unable to quantify the effects of distribution shifts over time.

This paper addresses the open problem of monitoring variations in course difficulty inside educational programs over time. We introduce item response theory (IRT)–originally proposed for high-stakes assessments [14]–as a promising new CA methodology. We assess the suitability of IRT for analyzing multi-year course performance data and show how IRT can yield valuable insights by revealing temporal course difficulty variations inside a Computer Science (CS) and a Mechanical Engineering (ME) Bachelor's program. Our analysis demonstrates the importance of IRT as a CA instrument that enables us to warrant temporal consistency and fairness inside educational programs. Our findings prompt policymakers to implement feedback mechanisms to verify that policies achieve intended effects and to detect and investigate unintended variations, such as those induced by COVID-19. Key contributions of this paper include:

- **Certifying the need for temporal modeling**: We show that course characteristics, such as difficulty, are subject to significant variations over time by conducting a log-likelihood ratio test comparing two nested course grade models. This highlights the need for CA methodologies capable of monitoring such fluctuations.

[*] Both authors contributed equally to this research.

LAK '24, March 18–22, 2024, Kyoto, Japan
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636902

- **IRT as temporal CA methodology**: We assess the suitability of IRT for CA by studying to what degree course performance data meets IRT's theoretical assumptions. We further evaluate the validity and reliability of IRT-based course difficulty estimates. By estimating difficulty values for *different offerings* of the *same course*, we can monitor variations in course difficulty over time.
- **Case study**: Using 10 years of course grade data from a CS and a ME Bachelor's program. We estimate course difficulty values for individual course offerings, revealing substantial variations over time. We further observe a systematic shift in course difficulty during the COVID-19 pandemic.

2 BACKGROUND AND RELATED WORK

2.1 Curriculum Analytics

Curriculum Analytics (CA) is a subfield of Learning Analytics that studies curriculum-related data (e.g., information describing when individual students take different courses and how well they perform in them) intending to understand, modify, and improve educational programs such as college degrees and professional certification programs [10].

Existing CA approaches can be classified into three major categories based on their underlying methodology: (i) process mining, (ii) process simulation, and (iii) curriculum structure-based prediction. Process mining techniques have been proposed to create representations of the educational process focusing on the order of interactions with individual curriculum elements (e.g., [10, 32]). As an extension to process mining, simulation approaches have been explored to estimate the effects of potential curriculum changes (e.g., [6, 23]). Lastly, prediction techniques have been developed to predict future student performance [27] and to make personalized curriculum recommendations [5, 18].

This paper addresses the open problem of how to monitor course difficulty variations inside a curriculum over time, which is crucial for ensuring fairness in treatment of individual student cohorts and consistency in GPA scores. Existing process mining and simulation approaches assume that individual courses behave the same over time and are known to suffer from concept drift issues [10]. Similarly, prior prediction studies assume a stationary data generation process and are unable to quantify the effects of distribution shifts. While descriptive statistics such as course *pass rates* (PR), student retention [35], and curriculum coherence [22] can be used to monitor courses over time, they provide limited information regarding underlying factors–i.e., is a metric change due to a variation in the course or cohort?

2.2 IRT in Curriculum Analytics

IRT has been proposed in the context of high-stakes assessments to address fundamental limitations of classical test theory (i.e., (i) the inability to compare scores obtained from different tests and (ii) the dependence of item parameters on the test taker cohort) [14]. Outside the domain of standardized testing, IRT-based approaches have, for example, been used for adjusting high school GPAs based on subject difficulty [17] and for health assessments [31]. Related to CA, multiple IRT-based approaches have been proposed to model students' university course satisfaction in a single year (e.g., [4]) and over multiple years (e.g., [29, 30]) based on students' teaching evaluation (SET) surveys.

Closest to the spirit of this paper is a work by Bacci et al. [3], which proposed a multidimensional latent class IRT (LC-IRT) model to assign first-year students into different performance groups using exam enrollment and exam grade data. They studied data from 861 incoming Economics and Business students going through six courses during the *single academic year* 2013/2014. Students were split by last name into four groups, and each group was taught courses by different lecturers. As part of their work, Bacci et al. [3] pointed out variations in course difficulty between individual groups. In contrast, our work focuses on accurately monitoring variations in course difficulty over *multiple years* using data from a CS Bachelor's program consisting of 19 courses over 9 years and a ME Bachelor's program consisting of 17 courses over 10 years. We show that IRT can yield valuable insights for CA using multi-year performance data. Bacci et al. [3] trained a comparatively more complex IRT model but reported difficulties fitting course discrimination parameters even when working with a small number of courses. In our work, we employ the simpler Rasch model [14] as it yielded the highest confidence regarding course difficulty parameter fit.

3 METHODOLOGY

First, we formulate a statistical procedure to verify whether the time-invariant course difficulty assumptions apply to CA datasets. Second, we introduce IRT in the CA context. Third, we define a multi-step IRT-based methodology (i.e., (i) dimensionality assessment, (ii) model selection, and (iii) validity/reliability assessment) for monitoring course difficulty variations over time, which we later use to analyze datasets from a CS and a ME Bachelor's program.

3.1 Testing Time-Invariance in Course Properties

To emphasize the need for CA methodologies capable of capturing variations in course difficulty over time, we conduct a likelihood ratio test [9] to determine if course properties, such as course difficulty, vary over time. The test is based on the null hypothesis that a simple model assuming that course properties are time-invariant is sufficient to describe the course grade data. If the test result is significant, we would reject this null hypothesis, indicating that a model that accounts for temporal variations in course properties may be a better fit. This would also suggest that changes in course properties occur over time.

Formally, we fit two models with nested model parameter spaces. First, we fit a simple regression model M_0 that models student-course grade ($g_{s,c}$) additively using student (θ_s) and course (θ_c) parameters, and an intercept (b): $g_{s,c} = b + \theta_s + \theta_c$. Here, the course parameters θ_c come from a parameter set Θ_0. The more complex model, M_1, also models the course grades additively but uses one parameter per course-semester combination: $g_{s,c} = b + \theta_s + \theta_{c \times t}$. Here, $\theta_{c \times t}$ represents the properties of course c in semester t. The parameters $\theta_{c \times t}$ now come from a superset $\Theta_1 \supseteq \Theta_0$, which allows us to formulate the null hypothesis H_0 and the alternative hypothesis H_1 for the likelihood ratio test as:

$$H_0 : \forall\,(c \times t) : \; \theta_{c \times t} \in \Theta_0, \quad H_1 : \exists\,(c \times t) : \; \theta_{c \times t} \in \Theta_1 \setminus \Theta_0. \quad (1)$$

3.2 Item Response Theory

In the following, we assume a curriculum consisting of several courses offered repeatedly in different semesters with dichotomous grades ("pass"/"fail"). We use the term *course offering* (CO) to refer to one course in one semester. Focusing on CA, the idea underlying IRT is to assign each student a latent trait value that explains the probabilities with which the student passes individual COs. The relationship between student trait values and CO pass rates (PRs) can be modeled by fitting a sigmoid function for each CO known as the item response function (IRF). The inverse image (x-axis) of the IRF is the student's trait value, and the image (y-axis) is the student's probability of passing a specific CO.

For CO j, the position of its IRF on the x-axis (i.e., the value with the largest IRF slope) indicates the CO difficulty denoted as δ_j. The slope of the IRF describes the CO discrimination property denoted as α_j. Given student trait θ_i, CO difficulty, and discrimination, we define the probability of passing a CO j as

$$\mathbb{P}(X_{i,j} = 1 \mid \theta_i,\, \alpha_j,\, \delta_j) = \frac{1}{1 - e^{-\alpha_j(\theta_i - \delta_j)}}, \tag{2}$$

where $X_{i,j}$ is the dichotomous response of student i to CO j. $\boldsymbol{X}$ is the potentially sparse *CO response matrix* capturing all responses. The IRT model defined by Equation 2 can be fitted using maximum likelihood estimation. If we optimize only the difficulty parameters δ_j and fix all $\alpha_j = 1$, we refer to it as *Rasch* or 1-parameter logistic model (1PL) [14]. If all α_j are free, we call it *Birnbaum* or as 2-parameter logistic model (2PL) [14]. Generalizing the *Birnbaum* model, the multidimensional IRT model (MIRT) [12] characterizes CO discrimination and student traits using multidimensional parameter vectors. MIRT explains observational data via multiple latent variables, which in our context can be interpreted as distinct latent traits that describe a student's ability to complete COs successfully. We refer to the 2-dimensional IRT model as *2PL-2DIM* and the 3-dimensional IRT model as *2PL-3DIM*.

IRT is designed to explain student performance data a posteriori–i.e., it explains past data by fitting course difficulty and student trait parameters. While one could use these parameters to make predictions about the future (i.e., how difficult will a course be next year), we emphasize that this paper only focuses on explaining past performance data to derive learning analytical insights. Because of this, we rely on information criteria (described in Section 3.3.3) that trade-off model fit with model complexity for different IRT models with *in-sample* data. We *do not* try to predict future student performance, which would require other means of model evaluation (e.g., cross-validation).

3.3 Verification of Model Assumptions and Model Selection

To verify the suitability of IRT for CA, we assess if multi-year course response data meets IRT's theoretical assumptions.

3.3.1 Dimensionality. To assess the suitability of one- and multidimensional IRT models, we study the number of latent dimensions required to explain variance in student performance data. We do so by performing principal component analysis (PCA) on the grade point CO response matrix $\boldsymbol{X}^{[0,100]}$ [21]. Because PCA demands a complete CO response matrix, we need to address the sparsity common in course examination data. We assume that skills associated with individual courses are content-based and do not vary from CO to CO (e.g., the content of the CS1 course is time-invariant). This assumption allows us to aggregate data from different COs of the same course to form a denser *course response matrix*. The remaining missing values (e.g., due to drop-out students) are filled via multiple iterative PCA imputation (MIPCA) [19], leaving us with a dense aggregated response matrix $agg(\boldsymbol{X}^{[0,100]})$ with 19 courses for CS and 17 courses for ME. MIPCA allows us to perform PCA on a complete matrix and estimates imputation-induced uncertainty in the recovered principal components (PCs). We use a scree plot visualizing the eigenvalues of the covariance matrix $\boldsymbol{C}_{agg(X^{[0,100]})}$ of the aggregated matrix as a complementary criterion for assessing latent dimensionality [21].

3.3.2 Local Independence. In the CA context, IRT's local independence (LI) assumption states that a student's probability of passing a CO is independent of their performance in other COs, given their latent trait. To determine the degree to which course performance data meets the LI assumption, we use the Q3 criterion [14]. Q3 studies residual correlations and quantifies pairwise dependencies. If the Q3 score of a pair deviates by more than a threshold value of 0.2 from the average Q3 score of all pairs, it is commonly suspected that LI is at risk [13, 14].

Because of the multi-year horizon, we cannot compute residuals for each individual CO pair (e.g., there is no CS student that took Statistics in 2013 and Databases in 2021). As in the dimensionality assessment, we assume that relationships between individual courses are content-based and thus time-invariant. We calculate Q3 values for the 19 courses in CS and 17 courses in ME using a *Rasch* model fitted on their aggregated dichotomous matrices ($agg(\boldsymbol{X})$).

3.3.3 Model Selection. After determining an upper bound on the number of latent dimensions, we fit corresponding *Rasch*, *Birnbaum*, and multidimensional IRT models. We select the final model using common information criteria–i.e., Akaike information criterion (AIC), Bayesian information criterion (BIC), and sample size adjusted Bayesian information criterion (SABIC) [14]. These criteria quantify the trade-off between model fit (log-likelihood) and potential overfitting (number of model parameters) and are form of *in-sample* validation.

3.4 Validity and Reliability Assessment

We evaluate validity and reliability of IRT-based student and course difficulty estimates before deriving CA insights.

3.4.1 Concurrent Validity and Regression Validation. We assess validity to verify that the fitted student and CO measures capture the constructs of interest (i.e., student ability and CO difficulty). We utilize concurrent and regression validity methods for this purpose. Concurrent validity compares IRT's trait parameters with variables designed to measure the same attribute (e.g., GPA measuring student performance). Regression validation studies whether our measures have higher predictive power than comparable variables.

We study *concurrent validity* by considering the correlations between fitted IRT trait and difficulty values and student GPA and CO pass rate (PR) statistics. In line with GPA adjustment research (e.g.,

[17]), we expect a positive correlation between student trait parameters and GPAs and a negative correlation between CO difficulty parameters and CO PRs.

For *regression validation*, analog to IRT, which uses a logistic regression model that explains the data via student trait and CO difficulty values, we fit an alternative logistic model that uses student GPA scores and CO PRs as inputs. This allows us to evaluate the IRT model's a posteriori fit by contrasting it with a model that employs GPA and CO PRs, assessing the in-sample predictive capabilities of both sets of features.

3.4.2 Internal Consistency Reliability and Simulation Study. We assess reliability to verify consistency of measures recovered by IRT. We employ two approaches: Internal consistency using split-half testing and a simulation study. Split-half testing partitions the dataset into two disjoint sets as basis for two measurements. Then, internal consistency in measurements is determined by examining whether the model produces comparable results on the two sets. The simulation study examines how much data under a certain missing value rate is required to ensure a robust model fit.

For *internal consistency reliability*, we split the dataset into two disjoint subsets. We fit two independent Rasch models on the subsets to assess the consistency of recovered model parameters. We quantify consistency by computing the Pearson correlation between the two models fitted on disjoint data. The dataset is split in two ways: random and time-dependent. First, we estimate two latent trait values for each student by splitting their CO responses into two disjoint subsets at *random*. Second, we estimate two latent trait values for each student for the *earlier* and *later* half of their CO responses. The time-invariant split addresses IRT's constant time-invariant trait assumption. While we address potential changes in course difficulty by fitting different parameters for different semesters, we need to warrant consistency in student trait values. We focus on students with at least 12 CO responses to obtain a temporal separation over multiple semesters. For both splitting approaches, we assess internal consistency by computing the Pearson correlation coefficient between the trait value pairs.

Using a *simulation study*, we evaluate the reliability of the IRT difficulty parameter estimation. Following common methodology (e.g., [21, 26]), we generate a ground truth IRT model by sampling student trait and CO difficulty values from a standard Gaussian and simulate student responses for different expected CO sizes ($\{50, 75, 100, 150, 200, 250, 300\}$). To mimic missing responses, we randomly mask individual response matrix entries with a probability equal to the missing value ratio of our real data (29%). The number of simulated students is chosen to meet the expected CO size. Following [25], we generate data for 1,000 seeds. We report root mean square error (RMSE) and Pearson correlation metrics of the learned difficulty parameters using the ground truth values.

4 DATA AND PREPROCESSING

The CS and ME datasets used for this study capture exam scores from two Bachelor's programs at a German university.

The CS dataset covers a time period of nine years (2013-2021), and includes exam data from 1098 students. It consists of 19 compulsory courses, each graded on a scale from 0 to 100 points. To *pass* an exam, a minimum score of 50 points is required, otherwise the exam is considered *failed*. Before obtaining the data, anonymization was performed by removing all demographic information and by adding uniform stochastic noise ranging from -5 to 5 to each grade. Preprocessing steps were implemented to warrant data quality and validity. These steps involved focusing on students' first exam attempts, excluding reattempts, and only considering students with at least 5 observed non-zero grades and COs with at least 20 students. As a result of preprocessing, the CS dataset was refined to include 664 students and 127 COs.

The ME dataset spans a period of ten years (2012-2021) and consists of data from 3059 students. It encompasses 18 compulsory courses with exams graded on a discrete scale ranging from 5.0 (worst) to 1.0 (best). A grade of 5.0 indicates a *failed* exam. All other grades indicate a *pass*. We transformed the original 5.0 to 1.0 scale to an international 0 to 100 point grade scale by referring to university guidelines. Similar to the CS dataset, anonymization was performed by removing demographic information. The preprocessing steps were identical to those performed for the CS dataset. After preprocessing, the ME dataset contains data from 1651 students and 177 COs.

Except for the project-based software engineering COs in the CS dataset, each CO grade was determined via a single examination at the semester's end, highlighting their importance. It is worth noting that the CS and ME datasets represent separate degree programs with no course overlap. Finally, before analyzing the data using the dichotomous IRT models, we converted the grade point data in each dataset to pass/fail data.

5 RESULTS

5.1 Testing Time-Invariance in Course Characteristics

Following the methodology (Section 3.1), we first perform the likelihood ratio test (LRT) for the two datasets. The LRT statistic for the CS dataset is 998.83 with a p-value $p < 0.001$ and 110 degrees of freedom. The LRT statistic for the ME dataset is 653.71 with a p-value of $p < 0.001$ and 113 degrees of freedom. Thus, in both tests, the null hypothesis (Eq. 1) of a simpler model being sufficient was rejected with high significance. This emphasizes the importance of accounting for temporal variations in course properties, such as course difficulty. These findings highlight the need for CA methodologies that can account for temporal course variations and advocate the following IRT-based analyses.

5.2 Verification of Model Assumptions and Model Selection

5.2.1 Dimensionality. To inform the model selection, we investigate how many latent dimensions are required to explain the variance captured in the course response matrix. After aggregating responses from different COs (see Section 3.3), the missing value ratios of individual courses vary between 7% and 44%. The ME courses' ratios are lower ($< 29\%$) than the CS courses' ratios. In both programs, we observe more missing values in courses recommended for later semesters. We generate 200 dense response matrices for CS and ME with different MIPCA imputations.

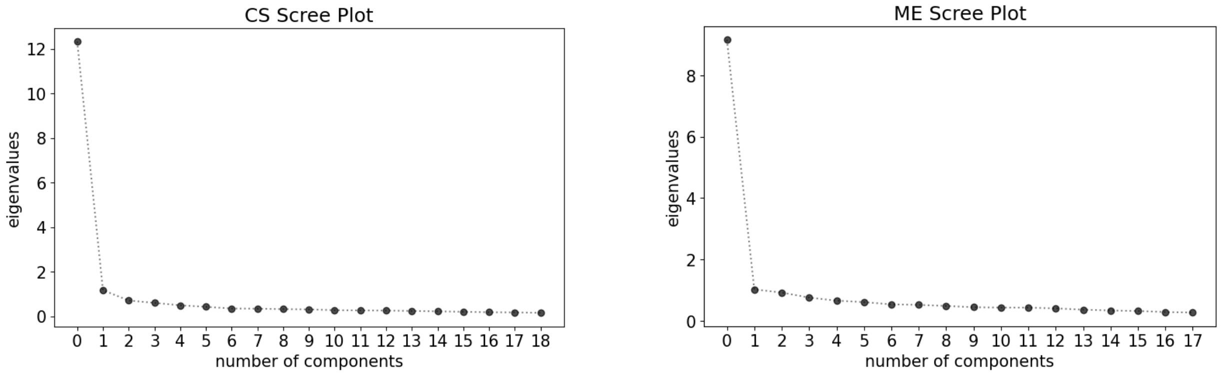

Figure 1: Scree plots visualizing the eigenvalues of the student course grade covariance matrix for a single MIPCA imputation. [Left] Computer Science (CS) program. [Right] Mechanical Engineering (ME) program.

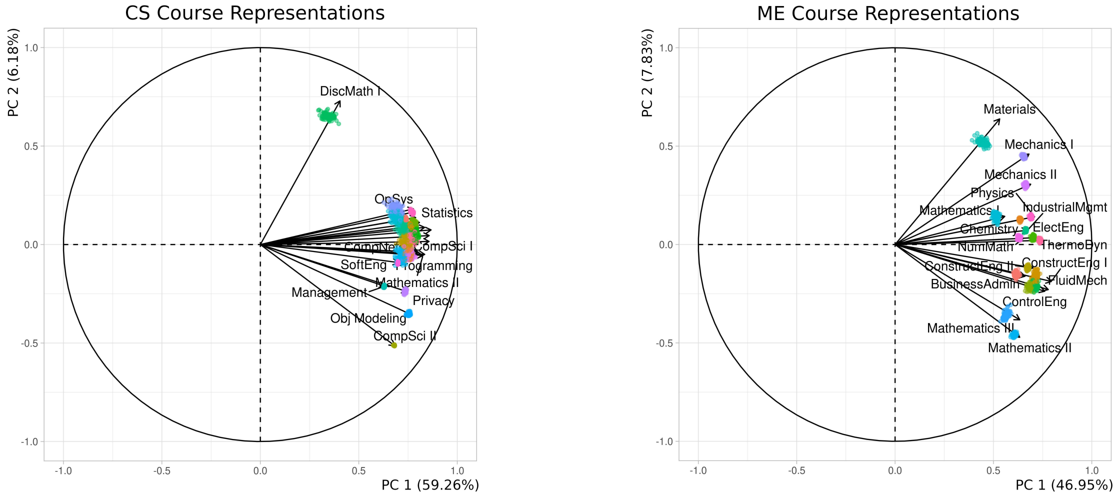

Figure 2: Scatter plots visualizing the variance in course representations using the first 2 principal components (PCs) recovered by different MIPCA imputations. [Left] Computer Science (CS) program. [Right] Mechanical Engineering (ME) program.

Focusing on *one* of the imputed matrices from each program, we visualize the eigenvalues of their corresponding covariance matrices in two Scree plots (CS: Figure 1 left, ME: Figure 1 right). In both Scree plots, we see one large eigenvalue (above 12 for CS and above 8 for ME). All other eigenvalues are significantly smaller and do not vary much in magnitude, which suggests at most one or two relevant latent dimensions represented by the first and second PC.

While the Scree plots focused on a *single* imputation, we now study the amount of uncertainty induced by *multiple* MIPCA imputations for each program. Figure 2 visualizes the individual CS (left subplot) and ME (right subplot) courses in the latent space defined by the first (x-axis) and second (y-axis) PC. The spread in the course representations indicates the degree of uncertainty induced by the imputations. For CS, we observe that representations tend to vary more for courses with more missing values (e.g., DiscMath I). Overall, however, the amount of induced uncertainty in the course representations is small, indicating that the recovered PCs are robust towards the exact imputation that is performed. We observe that most course representations align with the first PC and exhibit less variation in the second PC. Further, we see that PC 1 captures 59.26% and PC 2 captures 6.18% of the variance (axes in Figure 2 left). This behavior aligns with the eigenvalue relationships we observed in the Scree plot (Figure 1 left). For ME, the picture is similar. The representations are less spread out than those of CS, which is plausible, considering the smaller missing value ratios. However, the scatter of the course representations along the second PC is larger, and the courses are less close to each other. We see that PC 1 captures 46.95% of the variance and PC 2 captures 7.83% of the variance (axes in Figure 2 right). The first PC, therefore, explains less variance than the first PC in CS. This aligns with the largest eigenvalue we observed in the Scree plot (Figure 1 right), which is smaller than for the CS dataset. Thus, we consider one and two latent dimensions adequate for model selection and further consider the possibility of a third dimension.

5.2.2 Local Independence. For CS, a low average residual correlation of −0.06 provides evidence that the course performance data

Table 1: IRT model comparison. [Left] CS program. [Right] ME program.

CS Models	AIC	BIC	SABIC
Rasch	8439.4	**9015.1**	**8608.7**
Birnbaum	8445.1	9587.7	8781.2
2PL-2DIM	8372.8	10082.1	8875.6
2PL-3DIM	**8363.9**	10635.5	9032.1

ME Models	AIC	BIC	SABIC
Rasch	12123.9	**12806.5**	**12390.3**
Birnbaum	12153.2	13508.0	12682.1
2PL-2DIM	12049.6	14076.6	12840.9
2PL-3DIM	**12017.1**	14711.1	13068.8

meets IRT's local independence assumption. Out of the 171 course pairs, only three course pairs exhibited Q3 values outside the relative 0.2 threshold [13]. These three pairs are OpSys/Databases (Q3=0.149), Programming/Obj Modeling (Q3=0.216), and Mathematics I/Mathematics II (Q3=0.297). The correlations between these course pairs might be due to overlapping learning objectives. While we could address these higher Q3 values by modeling these three pairs via combined course grades, we decided to model them separately to have more fine-grained CO difficulty estimates.

For ME, we observe a similarly low average residual correlation of −0.05. In contrast to CS, none of the 136 possible course pairs exceeds the 0.2 threshold for Q3. Again, this provides evidence that the ME course data meets IRT's local independence assumption.

5.2.3 Model Selection. For each program, we train *Rasch*, *Birnbaum*, and *2PL-2DIM* IRT models and compare their fits using the information criteria AIC, BIC, and SABIC (CS: Table 1 left, ME: Table 1 right). In both programs, the relative ranking of criteria scores is the same: While the lower AIC score indicates that the *2PL-3DIM* model is preferred, the lower BIC and SABIC scores, which are more conservative regarding the number of model parameters, indicate that the *Rasch* model is more suitable. In addition, the *Rasch* performs better than the *Birnbaum* model in all three criteria. Thus, we focus on the *Rasch* model in the following analysis steps.

5.3 Validity and Reliability Assessment

5.3.1 Concurrent Validity and Regression Validation. For *concurrent validity*, we relate *Rasch* student trait estimates to student GPAs (CS: Figure 3 left, ME: Figure 3 right) and CO difficulty estimates to CO pass rates (PRs) (CS: Figure 4 left, ME: Figure 4 right). For both CS and ME, we see a strong positive correlation between student trait and GPA with a Pearson coefficient of $r = 0.931$ ($p < 0.001$) for CS and $r = 0.810$ ($p < 0.001$) for ME. We see a strong negative correlation between CO difficulty and CO PR with Pearson coefficients of $r = -0.908$ ($p < 0.001$) and $r = -0.842$ ($p < 0.001$) for CS and ME, respectively. This is consistent with our intuition that a higher student trait value relates to a higher GPA and a higher CO difficulty value relates to a lower PR. However, we observe high PRs for most ME COs, resulting in noisy student trait estimates (trait > 0.5) and difficulty values below zero for all COs. For CS, in Figure 4 left, we observe that COs with very high PRs (> 95%) stand out visually from the rest of the distribution. We examined the individual COs more closely and marked COs that fall into the period 2020-2022 as COVID-19 pandemic COs in red. An accumulation of pandemic COs among the COs with PRs > 95% is visible.

Next, we show the results for the *regression validation.* Table 2 shows that the student trait and CO difficulty measures yield better model fit indicators than the GPA and PR features when predicting CO outcomes. Although IRT only uses dichotomous (pass/fail) data to estimate the student trait, the models of both CS (Table 2 left) and ME (Table 2 right) perform better for all metrics compared to the student GPA model, which has access to more detailed point grade data.

5.3.2 Internal Consistency Reliability and Simulation Study. Following Section 3.4 we assess the *internal consistency reliability* by computing split half tests in two ways. First, when using a *random* CO response partitioning, we observe Pearson correlations of 0.801 (p<0.001) for CS and 0.675 (p<0.001) for ME, which indicate good reliability. Second, when using temporal partitioning, we also observe high correlations of 0.797 (p<0.001) for CS and 0.685 (p<0.001) for ME, which supports IRT's constant latent trait assumption.

We further conduct a *simulation study* to test how much data is required to ensure a reliable *Rasch* model parameter fit. Figure 5 shows average RMSE and Pearson correlation values and corresponding 90% confidence intervals by comparing CO difficulty values learned using different amounts of student data to ground truth difficulty parameters. We observe RMSE values < 0.33 (when training on ≥ 75 students per CO) and correlation values > 0.7 (in all cases), indicating that we can achieve a satisfactory model fit using small-scale data [26].

5.4 Investigating Model Parameters

To reveal variations in course difficulty over time, we visualize the CO difficulty values estimated by the *Rasch* models. We show the difficulty of compulsory courses for different semesters (CS: Figure 6, ME: Figure 7). Using bootstrapping, we provide confidence intervals for the individual difficulty parameters by first generating 100 datasets of equal size to our original dataset using sampling with replacement (student-wise) and then fitting IRT parameters for each.

5.4.1 Computer Science. For CS, again, we marked COs falling into the period 2020-2022 in red as COVID-19 pandemic COs. First, it can be seen that the difficulty of individual COs can vary over time. Looking at trends in difficulty, we observe that some courses became less difficult (e.g., CompSci II), some became more difficult (e.g., Mathematics I), some had low fluctuations (e.g., Privacy), and others had high fluctuations (e.g., Programming and Statistics). Focusing on the pandemic COs, we see a systematic downward trend in CO difficulty. Only CompArch and Privacy maintained their difficulty level during the pandemic. We employ a t-test on the overall pass rate (PR) of CS COs before and during the pandemic to show the statistical significance of this downward trend. The test indicates the significance of the PR differences (CS: t-statistic: 7.471,

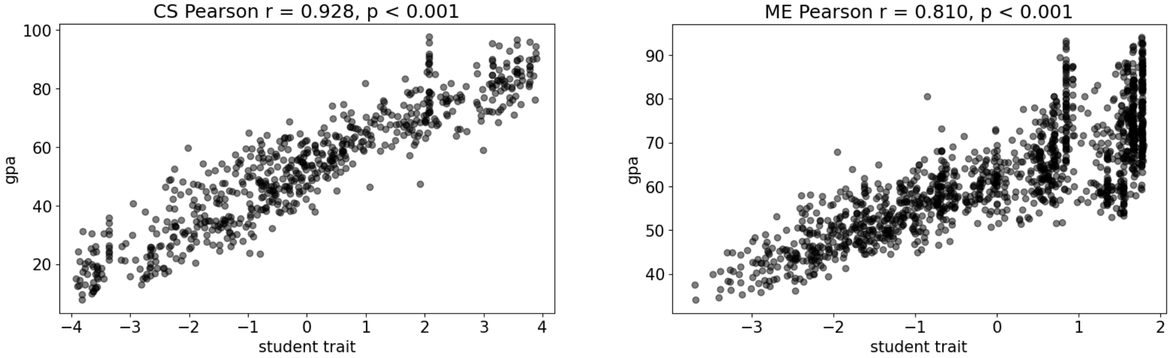

Figure 3: Scatter plots indicating correlation between student trait estimates based on *Rasch* model and student GPAs. [Left] Computer Science (CS) program. [Right] Mechanical Engineering (ME) program. program.

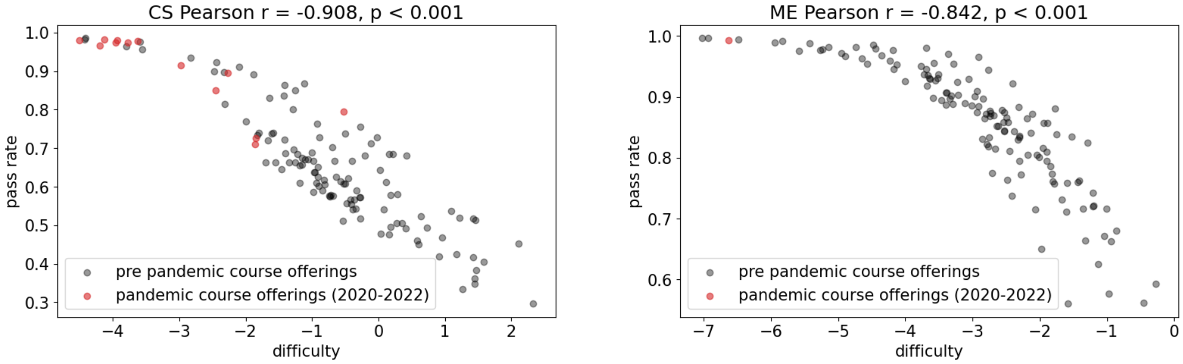

Figure 4: Scatter plots indicating the correlation between course offering (CO) difficulty estimates based on *Rasch* model and CO PR. COVID-19 COs (marked in red) accumulate at low difficulty. [Left] CS program. [Right] ME program.

Table 2: Model fit indicators for logistic regression models fitted using Rasch parameters and student GPA + CO PR. On all four metrics, the Rasch model performs better. [Left] CS program. [Right] ME program.

CS Models	ACC	AUC	NLL	RMSE
Rasch	**0.840**	**0.918**	**0.346**	**0.332**
GPA + PR	0.813	0.893	0.390	0.356

ME Models	ACC	AUC	NLL	RMSE
Rasch	**0.892**	**0.913**	**0.243**	**0.276**
GPA + PR	0.859	0.721	0.366	0.333

p-value: < 0.001, mean PR COVID COs: 0.875, mean PR 'remaining' COs: 0.558). Lastly, it is noticeable that COs with very low difficulty (< -3) (discussed in Section 5.3) have wider confidence intervals indicating uncertainty in the estimation process.

We observe a strong correlation between CO difficulty estimates and PRs (Figure 3 right). Remarkably, the *Rasch* model enables us to determine trait-*adjusted* PRs that allow us to compare COs frequented by different student cohorts (*unadjusted* PRs are inherently confounded by the trait level of their cohort). The IRT-adjusted PR of a course is the average probability that a student with an average trait value of (for CS -0.007) will pass the course, calculated across all its corresponding COs. In detail, we estimate Rasch model pass probabilities for the respective COs by taking the average student trait value. We then average these pass probabilities, weighting them according to the number of students examined in the COs. Table 3 shows IRT-adjusted and unadjusted PRs for all courses. We observe that IRT-adjusted PRs often do not vary much from unadjusted PRs (this might differ for individual COs). SoftEng and OpSys show particularly small differences. In contrast, Databases and Management show particularly large differences. In the first semester, we observe a general upward correction in the adjustment PRs and a downward correction from the second semester.

5.4.2 *Mechanical Engineering.* Inside the ME program we also detect different trends in temporal CO difficulty variations (Figure 7). Due to the generally higher pass rates, difficulty values fall mostly into the negative range. This is consistent with the width of the confidence intervals, which varies more within a level of difficulty for ME (e.g., BusinessAdmin).

Regarding the COVID-19 pandemic COs, we do not highlight them in red because there is only one (Figure 4 right), which also

Table 3: Mean pass rates (PRs) of compulsory CS courses over all semesters and mean PRs adjusted using mean *Rasch* student ability and course difficulties parameters. We see upward/downward adjustments during earlier/later semesters.

Sem.	Course Name	Avg Size	PR	Adj. PR
I	Mathematics I	82	0.641	0.690
	Statistics	64	0.673	0.684
	CompSci I	79	0.650	0.687
	Programming	69	0.622	0.581
	Economics	79	0.688	0.763
II	Mathematics II	69	0.650	0.655
	CompNets	74	0.678	0.634
	CompSci II	76	0.581	0.615
	Obj Modeling	76	0.621	0.566
	Management	59	0.620	0.374
III	DiscMath	69	0.620	0.555
	CompArch	76	0.648	0.616
	CompSci III	58	0.840	0.784
IV	Data Struct.	53	0.750	0.688
	SoftEng	67	0.823	0.824
	WebEng	74	0.771	0.825
	OpSys	70	0.632	0.633
V	Privacy	72	0.661	0.643
	Databases	83	0.585	0.488

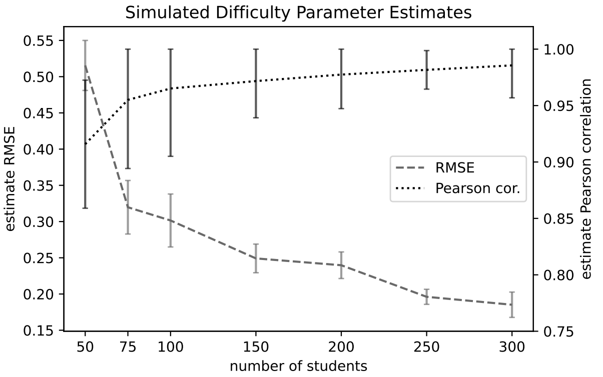

Figure 5: Simulation study across 1,000 simulated datasets. We provide average root-mean-square-error (RMSE) and Pearson correlation values compared to ground truth parameter sets and indicate 90% confidence intervals.

has a very low difficulty value. However, we noticed that most pandemic COs were discarded during preprocessing, as their PRs were 100 percent after the initial steps. On the unfiltered data, the t-test again revealed a statistically significant difference between the PRs of pandemic COs and the remaining COs (ME: t-statistic: 7.091, p-value < 0.001, mean PR pandemic COs: 0.965, mean PR 'remaining' COs: 0.721). Since PRs are strongly correlated with difficulty (Figure 4 right), this provides evidence that ME pandemic COs have significantly higher PRs.

6 DISCUSSION

Our analysis illustrates how item response theory (IRT) can serve as a methodology for curriculum analytics (CA). Importantly, IRT allows us to address the open problem of gaining insight into variations in course difficulty that can occur inside educational programs over time. This enables policymakers to monitor the effects of *conscious* curriculum changes, implemented to alter the properties of individual courses to support the achievement of their intended outcomes. Furthermore, the methodology can be used to detect *unintended* variations in course difficulty and prompt stakeholders to investigate underlying causal factors, to ensure fairness in course experience between cohorts [33].

The findings confirm that the difficulty level of individual courses can exhibit different trends over time (Figure 6 and 7). While the difficulty of some courses stays constant, for other courses, difficulty values increase, decrease, or show other types of fluctuations. Existing CA approaches do not capture such temporal variations because they assume time-invariant course characteristics. This is reflected in the concept drift issues of process mining and simulation techniques [10] and the stationary data generation assumption underlying prediction-based approaches. The data efficiency of IRT models makes them attractive in settings where datasets are too small for more complex Markov network- or deep learning-based approaches. This is particularly important when we want to recover temporal variations in course offerings (CO), as achieving a temporal resolution necessitates the use of even more data. Thus, IRT-based approaches are suitable for extending existing CA methods (e.g., process mining, simulation, and prediction) that do not account for time-dependent variation.

Beyond these algorithmic improvements, our study of temporal course difficulty variations yields valuable insights into shock phenomena of learning like the recent COVID-19 pandemic and is instrumental in assessing the effects of policy changes. Our analysis revealed a significant decline in the difficulty level of pandemic COs in both Computer Science (CS) and Mechanical Engineering (ME) programs (CS: Figure 6, ME: Figure 7). We discussed our findings with the faculty board and reflected on curriculum changes implemented in response to the pandemic. There was a systematic and conscious shift from traditional classroom lectures to an online-based teaching paradigm that puts a greater emphasis on problem-based learning, which might have enabled a larger number of students to achieve the learning objectives. Further investigations are needed to understand the exact effects of these curriculum changes on student learning outcomes and CO difficulty. Improved learning outcomes seem plausible, as problem-based learning activities can promote increased student engagement compared to conventional classroom lectures [2].

Course pass rates (PR) are a common statistic used to monitor courses inside educational degree programs (e.g., [11, 15, 35]). While course PRs can be informative, our analysis highlights that they need to be interpreted with care (Table 3). Course PRs, as a measure

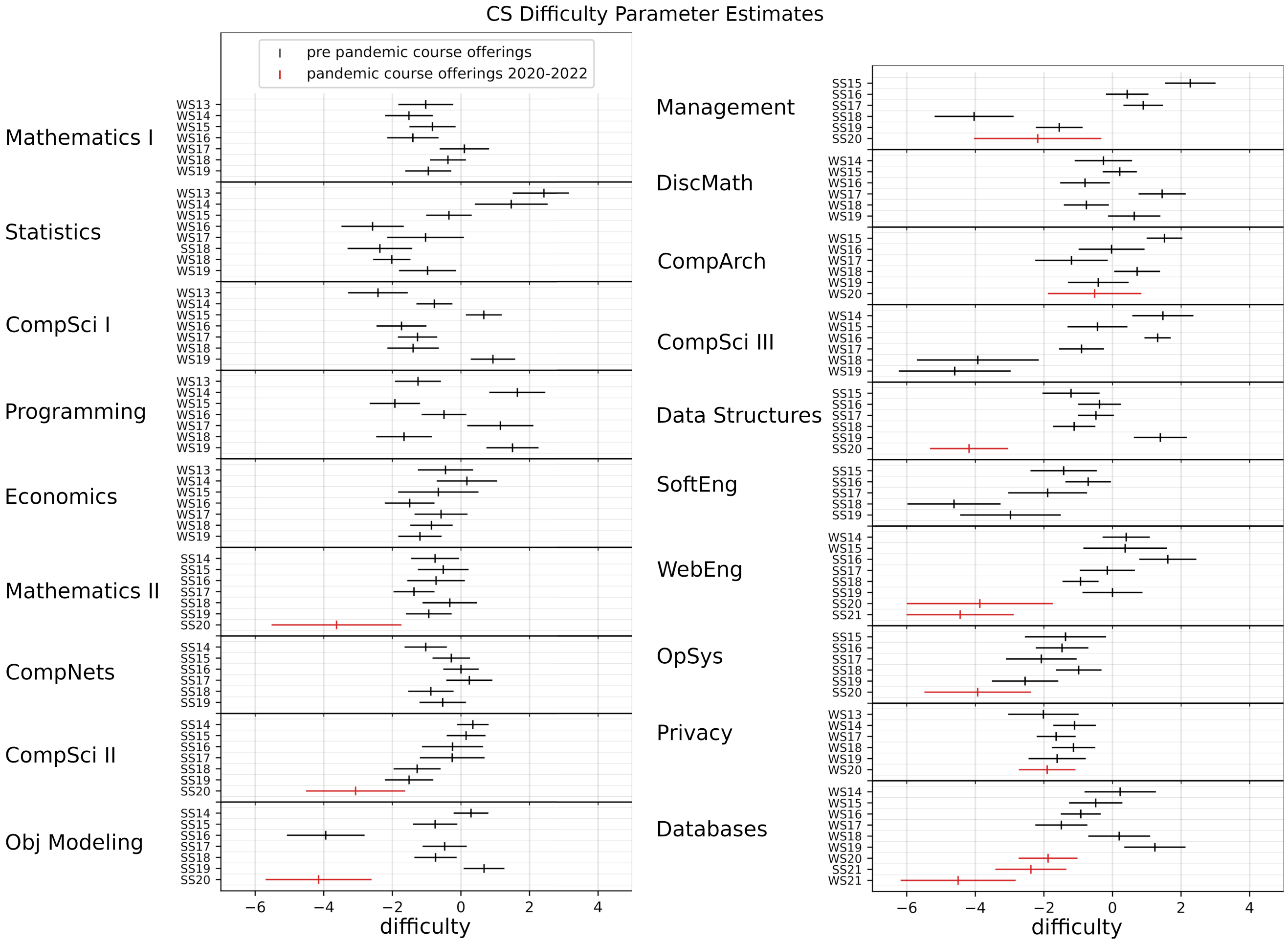

Figure 6: Plot visualizing variations in Computer Science (CS) course offering (CO) difficulty (as captured by *Rasch* IRT model) over time together with 95% confidence intervals (as determined by bootstrapping). We observe different patterns in difficulty trends (e.g., growth, oscillation, . . .). COVID-19 COs (highlighted in red) show systematically lower difficulty.

of course difficulty, can be confounded by the trait level of their respective student cohort. These factors can become even more pronounced during policy changes, such as those caused by the COVID-19 pandemic. Therefore, relying on course PRs can yield misleading insights into the true quality of the curriculum [7] and thus may affect the robustness of program (re-)accreditation processes that do not adjust for related biases in historic data [20]. The IRT framework allows us to define IRT-adjusted course PRs that quantify how well a student of average trait would have performed in each course. The results suggest that the unadjusted course PRs are too high in the later semesters, presumably due to dropouts in earlier semesters. Thus, a semester-by-semester curriculum and course quality assessment approach involving IRT-adjusted course PRs can offer a more dynamic and context-sensitive evaluation. Incorporating these refinements, institutions can create resilient and data-driven curriculum quality assurance, which is important in times of policy shifts with significant impacts, e.g., COVID-19 pandemic COs. This can make re-accreditation processes more robust and sensitive to internal and external changes in the learning environment. Accreditation bodies can include temporal course difficulty variations as part of their holistic evaluation criteria.

IRT as a CA methodology has implications for various stakeholders and problems in the higher education domain. Multidimensional IRT models can calibrate multiple skill traits at student and course levels, contributing to more holistic student profiles (e.g. [1, 34]). This allows academic departments to calibrate their major programs in conjunction with minor courses. Such calibration can help create a more coherent educational experience, bridging gaps in skills and knowledge that might exist between different college majors. Student advisors can employ skill data to guide students not only based on their academic performance but also their latent skill attributes. For instance, students contemplating a change in major or minor can be directed towards programs that are likely to align more closely with their current skill sets, thus potentially improving overall retention rates. We are currently working with

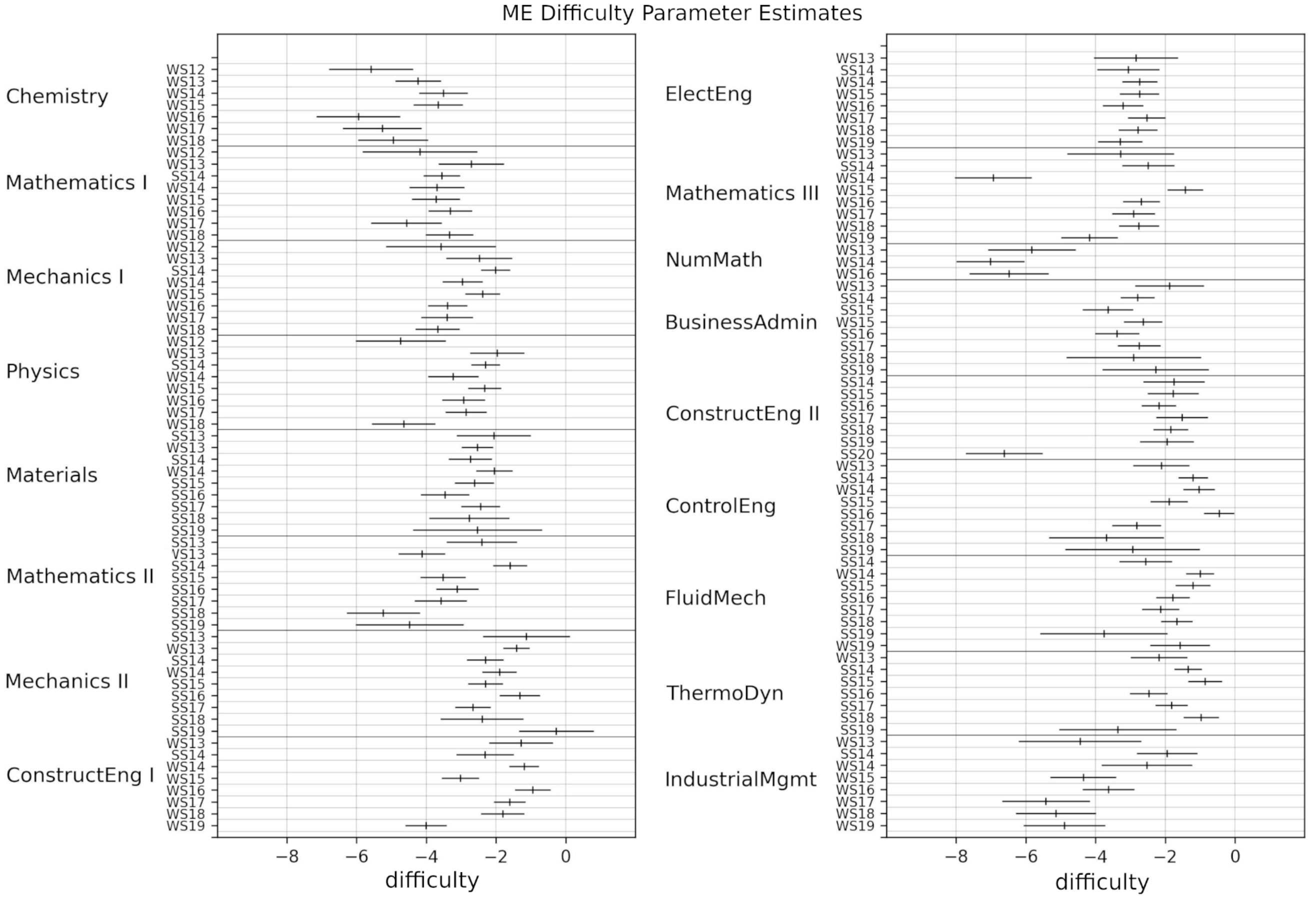

Figure 7: Plot visualizing variations in Mechnical Engineering (ME) course offering (CO) difficulty (as captured by *Rasch* IRT model) over time together with 95% confidence intervals (as determined by bootstrapping). We observe different patterns in difficulty trends (e.g., growth, oscillation, . . .). CO difficulties values are negative due to higher course pass rates.

student advisors to launch an advisor-facing dashboard to make these insights actionable [8].

Finally, IRT can inform course articulation decisions, which remain challenging for various stakeholders [24]. Different courses may cover similar learning material but vary significantly in difficulty levels. IRT models that monitor difficulty represent a data-driven approach that can inform articulation officers about the comparability of courses across individual institutions or departments. Such modeling approaches can be applied when intersections in data from different institutions exist, (i.e., students who already transferred between the two). IRT can model this intersection to calibrate the difficulty levels of potential course articulation pairs. This can improve the efficiency of credit transfer processes and ensure that students receive an equitable education regardless of their institutional pathway.

7 LIMITATIONS AND FUTURE WORK

IRT is predicated on two assumptions: (i) local independence (LI) and (ii) time-invariant latent student trait. In our context, the LI assumption posits that a student's probability of passing a particular course offering (CO) is independent of their performance in other COs, given their latent trait. Considering this assumption, this study focuses on first-attempts examination data to avoid dependencies on reattempts. To assess the degree to which our dataset meets the LI assumption, we further have employed the Q3 criterion [14], which has shown a low average residual correlation value of −0.06 for CS and −0.05 for ME. Only three of 171 CS course pairs exhibited a Q3 score outside the 0.2 threshold. While we could have addressed this by modeling these pairs as combined courses with a single grade, we decided in favor of more fine-grained difficulty estimates.

Our validity and reliability analysis suggest that a single-dimensional time-invariant student trait can be adequate to describe students' ability to pass COs in the two considered degree programs. This could be due to the specifics of the German system, where students choose their major before starting their studies, and the courses of each considered major are close in terms of technical content. In addition, all included courses are offered in a single department overseeing the major. Differences with other countries, which may offer more diverse majors in several departments, could lead to multidimensional student and course traits. Future work may investigate whether the dimensionality of other degree programs at different institutions (academic or professional) might differ and how latent traits between multiple overlapping majors can be calibrated using IRT, e.g., using shared COs.

One should be careful when interpreting student trait values as "ability to pass courses in a CS/ME program on the first attempt" as they might be more constant than specific aspects of student knowledge. The primary aim of this study was to *monitor variations in course difficulty*. Thus, the trait values should be considered in this context when interpreting the results. This limitation becomes evident when examining the ME program, which exhibits considerably higher pass rates compared to CS. This results in lower variations in the data, limiting the informational value for distinguishing between good and outstanding students, as shown during the validity assessment. One solution that holds promise is to employ polytomous IRT models, including Rating Scale Models and Partial Credit Models [21]. These models provide a nuanced capture of grading criteria and student performance. This makes it possible to distinguish not just between failing and passing students but also between good and outstanding students.

We do not explain the causal factors behind the observed difficulty variations. Empirically significant stakeholder interviews need to be conducted to narrow down potential causal factors. Three potential candidates are highlighted based on the discussion held with the faculty board. Firstly, policy changes, teacher turnover, or changing the department in which the course is offered can affect teaching style. Second, over time, the method of evaluation may shift, for example, from written to oral exams. Furthermore, external environment factors such as a pandemic can profoundly impact the entire educational process, resulting in a range of downstream effects, as we have witnessed.

8 SUMMARY AND CONCLUSION

We have shown how item response theory (IRT) can serve as foundation for a novel type of curriculum analytics (CA) methodology that enables us to monitor variations in course difficulty inside educational degree programs over time, which is essential for ensuring fairness in the treatment of individual student cohorts and consistency in GPA scores. The findings of the model selection process, as well as validity and reliability analyses, highlight the robustness of IRT course difficulty measures and their suitability for deriving CA insights. Our methodology revealed significant variations in course difficulty that were previously unknown to stakeholders. In particular, we detected a systematic downward shift in course difficulty levels during the COVID-19 pandemic. Furthermore, we found that conventional course pass rate measures need to be interpreted with care as they are confounded by temporal variations in course difficulty and cohort performance. We introduced *IRT-adjusted* pass rates as an alternative measure that addresses these effects. Overall, our findings prompt policymakers to monitor course difficulty variations over time to verify that implemented policy changes achieve intended effects and to revise policies as necessary.

Acknowledgements. This work is supported by the Ministry of Science (NRW, GER) as part of the project KI.edu:NRW.

REFERENCES

[1] Terry A Ackerman. 1994. Using multidimensional item response theory to understand what items and tests are measuring. *Applied measurement in education* 7, 4 (1994), 255–278.

[2] Stephanie Ahlfeldt*, Sudhir Mehta, and Timothy Sellnow. 2005. Measurement and analysis of student engagement in university classes where varying levels of PBL methods of instruction are in use. *Higher Education Research & Development* 24, 1 (2005), 5–20.

[3] Silvia Bacci, Francesco Bartolucci, Leonardo Grilli, and Carla Rampichini. 2017. Evaluation of student performance through a multidimensional finite mixture IRT model. *Multivariate Behavioral Research* 52, 6 (2017), 732–746.

[4] Silvia Bacci and Michela Gnaldi. 2015. A classification of university courses based on students' satisfaction: An application of a two-level mixture item response model. *Quality & Quantity* 49, 3 (2015), 927–940.

[5] Michael Backenköhler and Felix Scherzinger et al. 2018. Data-Driven Approach towards a Personalized Curriculum. In *Proceedings of the 11th International Conference on Educational Data Mining*. International Educational Data Mining Society, Raleigh, NC, 246–251.

[6] Frederik Baucks and Laurenz Wiskott. 2022. Simulating Policy Changes In Prerequisite-Free Curricula: A Supervised Data-Driven Approach. In *Proceedings of the 15th International Conference on Educational Data Mining*. Int. EDM Society, Durham, UK, 470–476.

[7] Frederik Baucks and Laurenz Wiskott. 2023. Mitigating Biases using an Additive Grade Point Model: Towards Trustworthy Curriculum Analytics Measures. In *Proceedings of the 21th Fachtagung Bildungstechnologien (DELFI)*. Gesellschaft fuer Informatik e.V., Aachen, Germany, 41–52.

[8] Frederik Baucks and Laurenz Wiskott. 2024. *Empowering Advisors: Designing a Dashboard for University Student Guidance*. Springer VS, Wiesbaden, GER. In press.

[9] Peter J Bickel and Kjell A Doksum. 2015. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. CRC, Boca Raton, USA.

[10] Alejandro Bogarín, Rebeca Cerezo, and Cristóbal Romero. 2018. A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining & Knowledge Discovery* 8, 1 (2018), 12–30.

[11] Malcolm Brown, Mark McCormack, Jamie Reeves, D Christopher Brook, Susan Grajek, Bryan Alexander, Maha Bali, Stephanie Bulger, Shawna Dark, Nicole Engelbert, et al. 2020. *2020 educause horizon report teaching and learning edition*. Technical Report. Educause.

[12] Philip Chalmers. 2012. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software* 48 (2012), 1–29.

[13] Karl Bang Christensen, Guido Makransky, and Mike Horton. 2017. Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied psychological measurement* 41, 3 (2017), 178–194.

[14] Rafael Jaime De Ayala. 2013. *The theory and practice of item response theory*. Guilford, New York, NY, USA.

[15] Nick Deschacht and Katie Goeman. 2015. The effect of blended learning on course persistence and performance of adult learners: A difference-in-differences analysis. *Computers & Education* 87 (2015), 83–89.

[16] Valentina Di Stasio. 2014. Education as a signal of trainability: Results from a vignette study with Italian employers. *European Sociological Review* 30, 6 (2014), 796–809.

[17] John Hansen, Philip Sadler, and Gerhard Sonnert. 2019. Estimating High School GPA Weighting Parameters With a Graded Response Model. *Educational Measurement: Issues and Practice* 38, 1 (2019), 16–24.

[18] Weijie Jiang, Zachary A. Pardos, and Qiang Wei. 2019. Goal-Based Course Recommendation. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (Tempe, AZ, USA) *(LAK19)*. ACM, New York, NY, USA, 36–45.

[19] Julie Josse, François Husson, et al. 2011. Multiple imputation in principal component analysis. *Advances in data analysis and classification* 5, 3 (2011), 231–246.

[20] René F Kizilcec and Hansol Lee. 2022. *Algorithmic fairness in education*. Routledge, Abingdon, UK, 174–202.

[21] Patrick Mair. 2018. *Modern psychometrics with R*. Springer, Cham, CH.

[22] Gonzalo Mendez, Xavier Ochoa, Katherine Chiluiza, and Bram de Wever. 2014. Curricular Design Analysis: A Data-Driven Perspective. *Journal of Learning Analytics* 1, 3 (Nov. 2014), 84–119.

[23] Roland Molontay, Noémi Horváth, Júlia Bergmann, Dóra Szekrényes, and Mihály Szabó. 2020. Characterizing curriculum prerequisite networks by a student flow approach. *IEEE Transactions on Learning Technologies* 13, 3 (2020), 491–501.

[24] Zachary A. Pardos, Hung Chau, and Haocheng Zhao. 2019. Data-Assistive Course-to-Course Articulation Using Machine Translation. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale* (Chicago, IL, USA). Association for Computing Machinery, New York, NY, USA, 1–10.

[25] Fulya Baris Pekmezci and Asiye ŞENGÜL Avşar. 2021. A guide for more accurate and precise estimations in Simulative Unidimensional IRT Models. *International Journal of Assessment Tools in Education* 8, 2 (2021), 423–453.

[26] Alper Sahin and Duygu Anil. 2017. The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sci.: Theory & Practice* 17, 1 (2017), 321–335.

[27] Ahmad Slim, Gregory L Heileman, Jarred Kozlick, and Chaouki T Abdallah. 2014. Employing markov networks on curriculum graphs to predict student performance. In *13th International Conference on Machine Learning & Applications*. IEEE, IEEE, Detroit, MI, USA, 415–418.

[28] Daniel Spurk and Andrea E Abele. 2011. Who earns more and why? A multiple mediation model from personality to salary. *Journal of Business and Psychology* 26, 1 (2011), 87–103.

[29] Isabella Sulis, Mariano Porcu, and Vincenza Capursi. 2019. On the use of student evaluation of teaching: a longitudinal analysis combining measurement issues and implications of the exercise. *Social Indicators Research* 142, 3 (2019), 1305–1331.

[30] Isabella Sulis, Mariano Porcu, and Nicola Tedesco. 2011. Evaluating Lecturer's Capability Over Time. Some Evidence from Surveys on University Course Quality. In *New Perspectives in Statistical Modeling and Data Analysis*. Springer Berlin Heidelberg, Berlin, Heidelberg, 13–20.

[31] Michael L. Thomas. 2011. The Value of Item Response Theory in Clinical Assessment: A Review. *Assessment* 18, 3 (2011), 291–307.

[32] Nikola Trcka, Mykola Pechenizkiy, and Wil van der Aalst. 2010. *Process mining from educational data*. CRC, Boca Raton, USA, 123–142.

[33] Suraj Uttamchandani and Joshua Quick. 2022. *An introduction to fairness, absence of bias, and equity in learning analytics*. Solar, NYC, USA, 205–212.

[34] Wim J van der Linden and Ronald K Hambleton. 2013. *Handbook of Modern Item Response Theory*. Springer, New York, NY, USA.

[35] Wai Yee Wong and Marcel Lavrencic. 2016. Using a Risk Management Approach in Analytics for Curriculum and Program Quality Improvement.. In *PCLA@ LAK*. SOLAR, Edinburgh, UK, 10–14.