Tracing Changes in University Course Difficulty Using Item Response Theory

Frederik Baucks[*1], **Robin Schmucker**[*2], **Laurenz Wiskott**[1]

[1] Institute for Neural Computation, Computer Science Department, Ruhr University Bochum
[2] Machine Learning Department, Carnegie Mellon University
frederik.baucks@ini.rub.de, rschmuck@cs.cmu.edu, laurenz.wiskott@ini.rub.de

Abstract

Curriculum analytics (CA) studies educational program structure and student data to ensure the quality of courses inside a curriculum. Ensuring low variation in course difficulty over time is crucial to warrant equal treatment of individual student cohorts and consistent degree outcomes. Still, existing CA techniques (e.g., process mining/simulation and curriculum-based prediction) are unable to capture such temporal variations due to their central assumption of time-invariant course behavior. In this paper, we introduce item response theory (IRT) as a new methodology to the CA domain to address the open problem of tracing changes in course difficulty over time. We show the suitability of IRT to capture variance in course performance data and assess the validity and reliability of IRT-based difficulty estimates. Using data from 664 CS Bachelor students, we show how IRT can yield valuable insights by revealing variations in course difficulty over multiple years. Furthermore, we observe a systematic shift in course difficulty during the COVID-19 pandemic.

1 Introduction

Maintaining low temporal variation in course difficulty in academic and professional degree programs is an important task to ensure equal treatment of individual student cohorts and to ensure consistent and informative grade point average (GPA) scores. GPA scores are a central measure used in decision processes by employers and academic institutions and are known to be correlated with students' future career success (e.g., (Spurk and Abele 2011; Di Stasio 2014)).

The field of Curriculum Analytics (CA) studies educational program structure and student data to assess the quality of individual courses inside a curriculum and their relationships to each other. Existing CA approaches that rely on process mining and simulation techniques to monitor student activities inside a curriculum, are known to suffer from concept drift issues and are unable to capture differences between *individual offerings* of the *same course* (e.g., CS1 in winter 2018 and CS1 in winter 2019) (Bogarín, Cerezo, and Romero 2018). Similarly, CA approaches that make curriculum structure-based predictions employ the IID assumption and are unable to quantify the effects of distribution shift.

[*] These authors contributed equally.

This paper addresses the open question of tracing changes in course difficulty inside educational degree programs over time. We introduce item response theory (IRT)–originally proposed for standardized testing (van der Linden and Hambleton 2013)–as a promising new methodology for CA. We assess the suitability of IRT for analyzing students' multi-year course performance data and show how IRT can yield valuable insights regarding course difficulty variations using data from a Computer Science (CS) Bachelor's program. We hope that IRT-based approaches can play an important role in ensuring the consistency and fairness of educational degree programs. The key contributions of this paper include:

- **IRT for tracing course difficulty over time**: We assess the suitability of IRT methodology for CA by studying variance in course performance data and by evaluating the validity and reliability of resulting model parameters. IRT explains performance data via parameters that capture latent course difficulty and student trait allowing us to quantify variations in course difficulty over time.
- **Case study**: Evaluation of IRT methodology using 9 years of course grade data from a CS Bachelor's program. We estimate difficulty values for individual offerings revealing substantial variations in course difficulty over time. Furthermore, we observe a systematic change in course difficulty during the COVID-19 pandemic.

2 Related Work

Curriculum Analytics (CA) is a subfield of Learning Analytics and Educational Data Mining 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 degree and professional certification programs (Bogarín, Cerezo, and Romero 2018).

Different metrics such as curriculum coherence (Mendez et al. 2014) and student retention (Wong and Lavrencic 2016) have been proposed to monitor curriculum quality. Other existing CA approaches can be classified into three main categories based on underlying methodology: (i) process mining, (ii) process simulation, and (iii) curriculum structure-based prediction. Process mining techniques have been proposed to create visualizations of the educational process focusing on the order of interactions with individual curriculum elements (e.g., (Trcka, Pechenizkiy, and van der Aalst 2010; Bogarín, Cerezo, and Romero 2018)). As an extension to process mining, simulation approaches have been explored to estimate effects of potential curriculum changes (e.g., (Molontay et al. 2020; Baucks and Wiskott 2022)). Lastly, different prediction techniques have been developed to predict future student performance (Slim et al. 2014) and to make personalized curriculum recommendations (Backenköhler et al. 2018; Jiang, Pardos, and Wei 2019).

In this paper, we address the open question of how to trace changes in course difficulty inside a curriculum over time which is crucial for ensuring equal treatment of individual student cohorts and consistent 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 (Bogarín, Cerezo, and Romero 2018). Similarly, prior prediction studies build on the IID assumption and are unable to quantify the effects of distribution shift (i.e., varying course difficulty). While descriptive statistics such as course *pass rates* (PR) and student retention 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?

IRT has been proposed in the context of standardized testing to address fundamental limitations of classical test theory (i.e., (i) the inability to compare student scores obtained from different tests and (ii) the dependence of item parameters on the test taker cohort) (van der Linden and Hambleton 2013). Outside the domain of standardized testing IRT based approaches have for example been used for adjusting high school GPAs based on subject difficulty (Hansen, Sadler, and Sonnert 2019) and for health assessments (Thomas 2011). Related to CA multiple IRT-based approaches have been proposed to model students' university course satisfaction in a single year (e.g., (Bacci and Gnaldi 2015)) and over multiple years (e.g., (Sulis, Porcu, and Tedesco 2011; Sulis, Porcu, and Capursi 2019)) based on students' teaching evaluation (SET) surveys. While student satisfaction is an important metric, concerns have been raised about the low correlation between SET evaluations and learning outcomes (Uttl, White, and Gonzalez 2017).

Closest to the spirit of this paper is a work by Bacci et al. (2017) which proposed a multidimensional latent class IRT (LC-IRT) model to classify 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 into four groups by their last name and each group was taught courses by different lecturers. As part of their work Bacci et al. (2017) pointed out variations in course difficulty between individual groups. In contrast, our work focuses on accurately tracing changes in course difficulty over *multiple years* using data from a Computer Science Bachelor's program consisting of 19 courses over nine years. We show that IRT can yield valuable insights from students' multi-year performance data. Bacci et al. (2017) trained a comparatively more complex IRT model, but reported difficulties with fitting course discrimination parameters even when working with a small number of courses. In our work, we employed the simpler Rasch model (Rasch 1960) as it yielded the highest confidence regarding difficulty parameter fit.

3 Methodology

Focusing on the CA domain, we introduce the IRT framework. We then define a multi-step IRT-based methodology (i.e., (i) dimensionality assessment, (ii) model selection, and (iii) validity/reliability assessment) for tracing course difficulty changes over time which we later evaluate on data from a CS bachelor's program. Importantly, IRT assumes that a student's latent trait does not vary over time (detailed discussion in Section 5). With this assumption, IRT's suitability for describing data from a multi-year program is not obvious. Therefore, we also evaluate the suitability of the IRT framework for CA using reliability and validity methods.

3.1 Item Response Theory

In the following, we assume a curriculum consisting of a number of courses offered repeatedly in different semesters with dichotomous grades ("pass"/"fail"). We use the term *course offerings* (CO) to refer to one course in one semester. Focusing on CA, the idea underlying IRT is to relate each student's average course PR to the overall probabilities with which students pass individual COs. The relationship between student and CO PRs can be modeled by fitting a sigmoid function for each CO known as item response function (IRF). The inverse image (x-axis) of the IRF consists of student trait values, which can be thought of as a form of student PRs. The image (y-axis) of the IRF is the probability of passing a certain CO.

For CO j the position of its IRF on the x-axis (i.e., the trait value at which the IRF has the largest 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

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

where $X_{i,j}$ is the dichotomous response of student i to CO j. X is the potentially sparse *CO response matrix* capturing all responses. The IRT model defined by Equation 1 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) (Rasch 1960). If all α_j are free, we call it *Birnbaum* or as 2-parameter logistic model (2PL) (Birnbaum 1968).

The *Birnbaum* model has been generalized to a multidimensional IRT model (Chalmers 2012) which characterizes CO discrimination and student traits using multidimensional parameter vectors. This multidimensional model explains observational data via multiple latent variables, which can be interpreted as distinct student skills. In the following, we refer to the 2-dimensional IRT model as *2PL-2DIM*.

3.2 Dimensionality Assessment

In our setting, IRT explains student course performance data via student trait and CO difficulty parameters. The number of latent dimensions required to explain the data relates to the number of distinct traits that describe a student's ability to complete COs successfully. To assess the number of dimensions we perform principal component analysis (PCA) on the grade point CO response matrix $X^{[0,100]}$ (Mair 2018). Because the PCA algorithm 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 change from offering to offering (e.g., the content of the CS1 course is time-invariant). This assumption allows us to aggregate the data from different offerings of the same course to form a denser *course response matrix*. The remaining missing values (e.g., due to drop-out students) are filled using multiple iterative PCA imputation (MIPCA) (Josse, Husson et al. 2011), leaving us with a dense aggregated course response matrix $agg(X^{[0,100]})$ with 19 courses. MIPCA allows us to perform PCA on a complete matrix and can estimate imputation-induced uncertainty in the recovered principal components (PCs). Finally, we use a Scree plot visualizing the eigenvalues of the covariance matrix $C_{agg(X^{[0,100]})}$ of the aggregated course response matrix as a complementary criterion for assessing latent dimensionality (Mair 2018).

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) (Akaike 1998), Bayesian information criterion (BIC) (Schwarz 1978) and sample size adjusted Bayesian information criterion (SABIC). These criteria quantify the trade-off between model fit (log-likelihood) and potential overfitting (number of model parameters).

3.4 Validity and Reliability Assessment

One core assumption underlying IRT is that the latent student trait stays constant over time which is natural in the standardized testing domain. In the CA context, it is not obvious that IRT is suitable to model data from a multi-year degree program. We, therefore, need to ensure the validity and reliability of the parameters recovered by IRT for CA.

We study concurrent validity by considering correlations between IRT parameters and student GPAs and CO PRs. In line with GPA adjustment research (e.g., (Hansen, Sadler, and Sonnert 2019)) we expect a positive correlation between student trait parameters and GPAs, and a negative correlation between CO difficulty parameters and PRs.

We evaluate the reliability of the difficulty parameter estimation via a simulation study. Following common methodology (e.g., (Sahin and Anil 2017; Mair 2018)) 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 recommendations by Pekmezci and Avşar (2021), we generate data for 1,000 seeds. We report root mean square error (RMSE) and Pearson correlation metrics of the learned difficulty parameters using ground truth.

4 Experiments

4.1 Dataset Description

The dataset used for our study provides exam scores from a CS Bachelor's program at an anonymous university in Germany. Between 2013 and 2022, exam data from 1098 students was collected for 19 compulsory courses including data from graduated, enrolled, and dropout students. The grading scale of each exams is $[0, 100]$. An exam is considered *passed* if at least 50 percent is achieved and *failed* otherwise. Except for the project-based software engineering course, each course grade was determined via a *single* written examination at the end of the semester which emphasizes the importance of these individual assessments.

Before obtaining the data, anonymization was performed by removing all demographic information and by adding a uniform stochastic noise between $[-5, 5]$ to each grade. We performed the following preprocessing steps: Considering IRT's local independence assumption, we focused on students' first exam attempts and omitted reattempts. Further, students with < 5 observed grades > 0 were omitted, and we omitted COs with less than 20 students to promote a stable difficulty parameter fit. This resulted in a dataset with 664 students and 127 COs. Since we use dichotomous IRT models, we converted the grade point to 'pass'/'fail' data.

4.2 Dimensionality Assessment

To inform the model selection, we investigate how many latent dimensions are required to explain variance captured in the course response matrix. After aggregating responses from different COs (see Subsection 3.2), the missing value ratios of individual courses vary between 7% and 44%. We observe more missing values in courses recommended for later semesters. We generate 1,000 dense response matrices by filling missing values with different MIPCA imputations.

Focusing on *one* of the imputed matrices, we visualize the eigenvalues of its corresponding covariance matrix in a Scree plot (Figure 1). We see one large eigenvalue above 12. All other eigenvalues are significantly smaller and do not not vary much in magnitude which suggests one or two latent dimensions represented by the first and second PC.

While the Scree plot focused on a *single* imputation, we now study the amount of uncertainty induced by *multiple* MIPCA imputations. Figure 2 visualizes the individual courses in the latent space defined by the first (x-axis) and second (y-axis) PC. The spread in the individual course representations shows the degree of uncertainty induced by the MIPCA imputations. We observe that representations tend to vary more for courses with more missing values. 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. For the dimensionality assessment, we observe that most

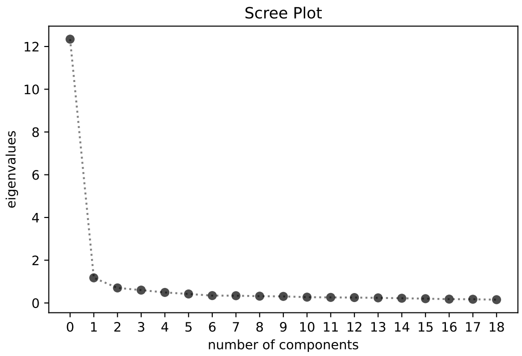

Figure 1: Scree plot visualizing the eigenvalues of the student course grade covariance matrix for a single imputation.

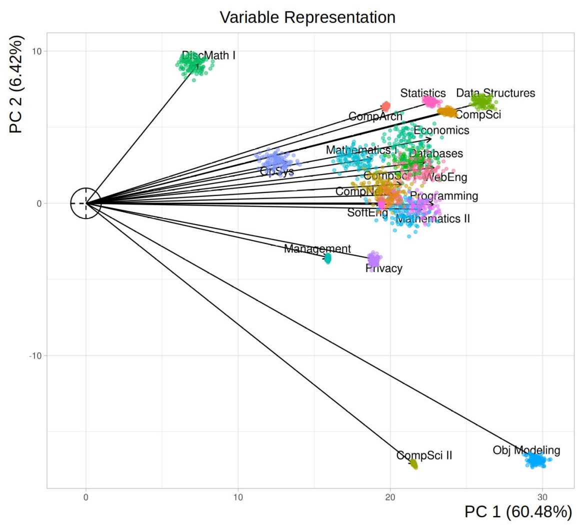

Figure 2: Scatter plot showing variance in course representations using first 2 PCs for different MIPCA imputations.

course representations are aligned with the first PC and exhibit less variation in the second PC. Further, we see that PC 1 captures 60.48% and PC 2 captures 6.42% of the variance (Figure 2 axis). This also aligns with the eigenvalues relationships we observed in Figure 1. We thus consider one and two latent dimensions in following model selection.

4.3 Model Selection

We train *Rasch*, *Birnbaum*, and *2PL-2DIM* IRT models and compare their fits using the information criteria AIC, BIC, and SABIC (Table 1). While the lower AIC score indicates that the *2PL-2DIM* 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* model performs better than the *Birnbaum* model in all three criteria. Thus, we focus on the *Rasch* model in the following experiments.

Model	AIC	BIC	SABIC
Rasch	8439.35	**9015.13**	**8608.73**
Birnbaum	8445.11	9587.67	8781.21
2PL-2DIM	**8372.75**	10082.10	8875.58

Table 1: Information criteria for different IRT models.

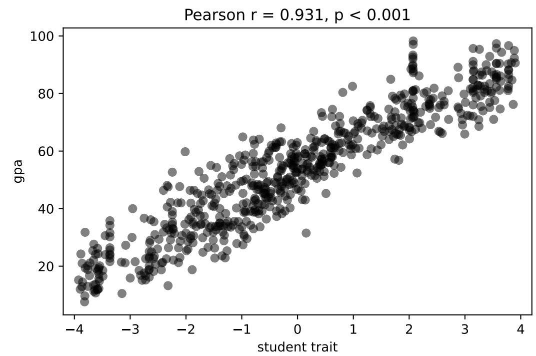

Figure 3: Scatter plot indicating correlation between student trait estimates based on *Rasch* model and student GPAs.

4.4 Validity and Reliability Assessment

We examine the student trait and course difficulty parameters learned by the *Rasch* model. To assess concurrent validity, we relate student trait estimates to student GPAs (Figure 3) and CO difficulty estimates to CO PRs (Figure 4). We see a strong positive correlation between student trait and GPA with a Pearson coefficient of $r = 0.931$ ($p < 0.001$). We see a strong negative correlation between CO difficulty and PR with a Pearson coefficient of $r = -0.908$ ($p < 0.001$). This meets our intuition that a higher student trait value relates to a higher GPA and a higher CO difficulty value relates to a lower PR. In Figure 4, we observe that COs with very high PRs ($> 95\%$) visually stand out from the rest of the distribution. We examined the individual COs more closely and marked COs that fall into the period 2020-2022 as pandemic COs in red. A strong accumulation of pandemic COs among the COs with PRs $> 95\%$ is visible.

Simulation Study Following Subsection 3.4 we conduct a simulation study to test how much data is required to ensure a reliable *Rasch* model fit. Figure 5 shows average RMSE and Pearson correlation values and corresponding 90% confidence intervals by comparing CO difficulty values learned from 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 (Sahin and Anil 2017).

4.5 Investigating Model Parameters

As additional comparison of IRT's student trait and CO difficulty parameters to student GPAs and CO PRs, Table 2 shows that student trait and CO difficulty lead to a better model fit than GPA and PR when predicting CO outcomes.

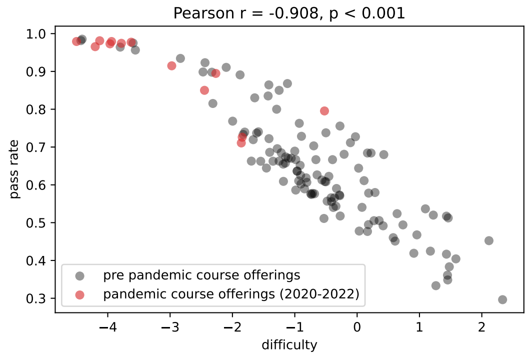

Figure 4: Scatter plot indicating correlation between CO difficulty estimates based on *Rasch* model and CO PR.

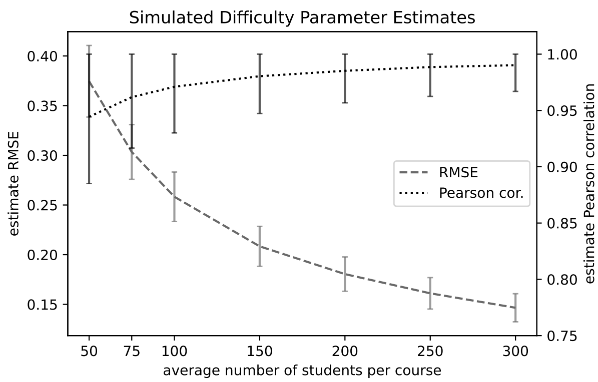

Figure 5: Simulation study across 1,000 simulated *Rasch* datasets. We provide average RMSE and Pearson correlation values by comparing learned difficulty to ground truth difficulty parameters and plot 90% confidence intervals.

Here, analog to the IRT model which is a logistic regression model that explains the data using student trait and CO difficulty values, we fitted a logistic regression that explains the data using student GPA scores and CO PRs. This is informative as it allows us to compare the predictive power of the two variable pairs. Although the IRT model only uses dichotomous (pass/fail) data to determine the student trait, it performs better for all metrics compared to the GPA model which has access to detailed point grade data.

To trace changes in course difficulty over time, we visualize the estimated CO difficulty values for each of the 19 compulsory courses for different semesters (Figure 6). We quantify the reliability of the model fitting process, by providing confidence intervals derived from the Fisher information matrix used in the Wald test (Agresti 2003). Again, we marked COs falling into the period 2020-2022 in red as 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 oth-

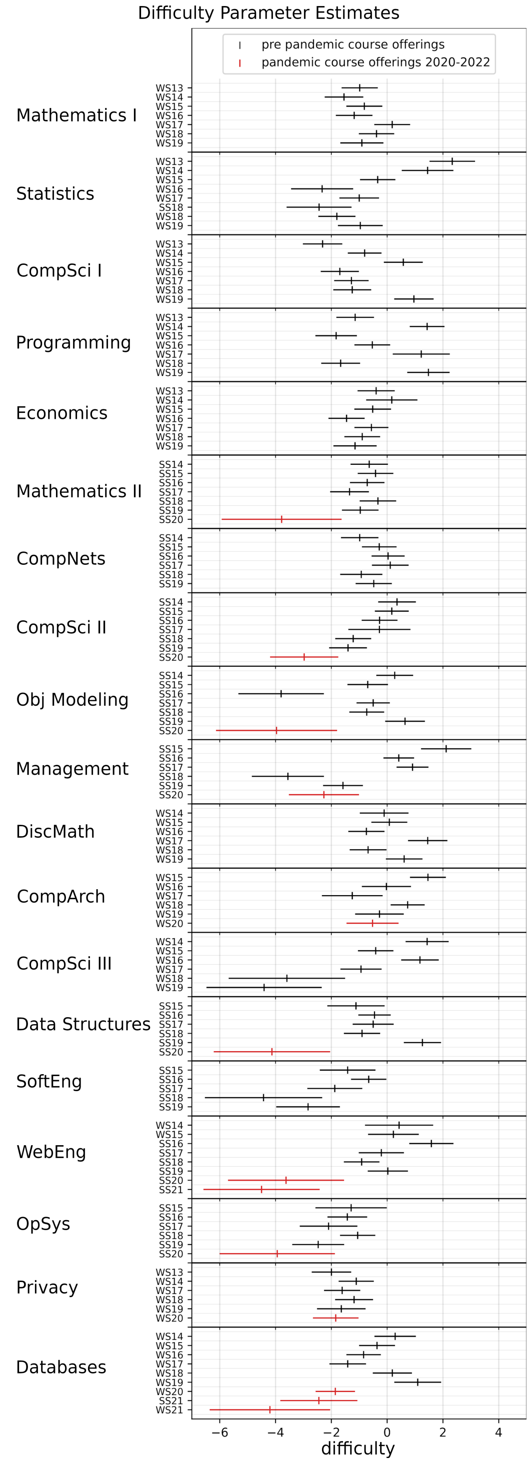

Figure 6: Scatter plot visualizing changes in CO difficulty (as captured by *Rasch* IRT model parameters) over time together with 95% confidence intervals (as determined by Wald test). We observe different patterns in CO difficulty trends (stationary, increasing, decreasing, oscillation). Marked in red are pandemic COs (conducted after WS2019) which exhibit substantially lower difficulty values compared to their non-pandemic versions.

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

Table 2: Model fit indicators for logistic regression models using Rasch parameters and student GPA + CO PR.

Course Name	Mean Size	PR	Adjusted PR
Mathematics I	82	0.641	0.719
Statistics	64	0.673	0.653
CompSci I	79	0.650	0.704
Programming	69	0.622	0.586
Economics	79	0.688	0.759
Mathematics II	69	0.650	0.641
CompNets	74	0.678	0.659
CompSci II	76	0.581	0.612
Obj Modeling	76	0.621	0.570
Management	59	0.620	0.424
DiscMath	69	0.620	0.573
CompArch	76	0.648	0.629
CompSci III	58	0.840	0.788
Data Structures	53	0.750	0.658
SoftEng	67	0.823	0.828
WebEng	74	0.771	0.842
OpSys	70	0.632	0.624
Privacy	72	0.661	0.611
Databases	83	0.585	0.490

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

ers 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. Lastly, it is noticeable that the COs with very low difficulty (< -3) (discussed in Subsection 4.4) have wider confidence intervals indicating uncertainty in the estimation process.

We observed a strong correlation between CO difficulty estimates and PRs (Figure 4). Remarkably, the *Rasch* model enables us to determine trait *adjusted* PRs that allow us to compare COs taken by different student cohorts (*unadjusted* PRs are confounded by the traits of their respective cohort). Here, the adjusted PR of a course is the mean probability for a student of average trait value (-0.007) of passing that course computed over all respective COs as determined by the *Rasch* model. Table 3 shows adjusted and un-adjusted PRs for all courses. We observe that 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 upwards correction in the adjustment PRs, and from the second semester on a downward correction.

5 Discussion and Future Work

Our experiments showed that item response theory (IRT) based methodology can provide valuable insights in the curriculum analytics (CA) domain. Particularly, IRT allows us to address the open problem of tracing changes in course difficulty over time. This includes instances where the institution consciously decides to alter the difficulty of a specific course, as well as situations where unintended difficulty changes occur. Our methodology can quantify the effects of policy changes and can in case of unintended variations raise a flag to start the search for underlying causal factors.

We observed that course difficulty values can exhibit different trends over time. Difficulty values can increase, decrease, or can show other types of fluctuations. Existing CA approaches cannot capture such temporal effects because they assume constant course properties. This is reflected in the concept drift issues of process mining and simulation techniques (Bogarín, Cerezo, and Romero 2018) and the IID assumption underlying prediction-based approaches. IRT-based techniques could be used to improve such methodological shortcomings in future work by accounting for course changes over time. In particular, this is useful when datasets are too small for temporal resolution with Markov/Bayes networks or deep learning techniques.

IRT is predicated on two key assumptions: (i) local independence and (ii) constant latent trait. In our context, the local independence 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 focused on first-attempt examination data. Future work will employ the Q3 criterion (Yen 1993) to quantify to what degree course performance data meets this assumption. The constant latent trait assumption posits that a student's latent trait stays constant across examination items. The construct has been addressed by limiting exams to first attempts. Future work will use split half reliability for further validation. However, the resulting meaning of the trait as "ability to pass courses in a CS program on the first attempt" should be interpreted with care as it might be more constant than certain specific aspects of student knowledge. The primary aim of this study was to quantify changes in course difficulty, thus the trait values should be considered in this context when interpreting the results. One limitation of our study is that we only considered data from a single degree program at a single university. Applying this methodology to other types of degree programs (academic or professional) will be important to assess the generalizability of the proposed methodology.

The pass rates (PR) of individual course offerings are confounded by the trait level of their respective student cohort. The IRT framework allows us to define trait adjusted course PRs, that quantify how well a student of average trait would have performed in each course. The results suggest that in the later semesters, the unadjusted PRs are too high presumably due to dropouts in earlier semesters. Adjusting PRs via the application of polytomous IRT models (e.g., rating scale and partial credit models (Mair 2018)), that can capture more information about grading criteria, is an interesting direction for future work.

We saw a systematic drop in the difficulty of most compulsory courses during the COVID-19 pandemic (Figure 6). This systematic shift raises the question of underlying causal factors. Two potentials explanations are: (i) A lowered course niveau. (ii) A more beneficial learning environment (e.g., online teaching, communication of learning objectives). First would lead in the long run, to knowledge gaps and could harm student's academic and professional advancement. The second would lead to opposite results.

Lastly, we hope that similar IRT-based approaches will become standard CA tools to quantify and control variations in course difficulty over time to ensure the equal treatment of different student cohorts and consistent GPA scores.

References

Agresti, A. 2003. *Categorical data analysis*. Wiley & Sons.

Akaike, H. 1998. *Information Theory and an Extension of the Maximum Likelihood Principle*, 199–213. New York, NY: Springer New York.

Bacci, S.; Bartolucci, F.; Grilli, L.; and Rampichini, C. 2017. Evaluation of student performance through a multidimensional finite mixture IRT model. *Multivariate Behavioral Research*, 52(6): 732–746.

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

Backenköhler, M.; Scherzinger, F.; Singla, A.; and Wolf, V. 2018. Data-Driven Approach towards a Personalized Curriculum. *International Educational Data Mining Society*.

Baucks, F.; and Wiskott, L. 2022. Simulating Policy Changes In Prerequisite-Free Curricula: A Supervised Data-Driven Approach. In *Proceedings of the 15th International Conference on Educational Data Mining*, 470.

Birnbaum, A. L. 1968. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bogarín, A.; Cerezo, R.; and Romero, C. 2018. A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(1): e1230.

Chalmers, R. P. 2012. mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, 48: 1–29.

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

Hansen, J.; Sadler, P.; and Sonnert, G. 2019. Estimating High School GPA Weighting Parameters With a Graded Response Model. *Educational Measurement: Issues and Practice*, 38(1): 16–24.

Jiang, W.; Pardos, Z. A.; and Wei, Q. 2019. Goal-Based Course Recommendation. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK19, 36–45. New York, NY, USA: ACM.

Josse, J.; Husson, F.; et al. 2011. Multiple imputation in principal component analysis. *Advances in data analysis and classification*, 5(3): 231–246.

Mair, P. 2018. *Modern psychometrics with R*. Springer.

Mendez, G.; Ochoa, X.; Chiluiza, K.; and de Wever, B. 2014. Curricular Design Analysis: A Data-Driven Perspective. *Journal of Learning Analytics*, 1(3): 84–119.

Molontay, R.; Horváth, N.; Bergmann, J.; Szekrényes, D.; and Szabó, M. 2020. Characterizing curriculum prerequisite networks by a student flow approach. *IEEE Transactions on Learning Technologies*, 13(3): 491–501.

Pekmezci, F. B.; and Avşar, A. Ş. 2021. A guide for more accurate and precise estimations in Simulative Unidimensional IRT Models. *International Journal of Assessment Tools in Education*, 8(2): 423–453.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in mathematical psychology. Aarhus, Denmark: Danmarks Paedagogiske Institut.

Sahin, A.; and Anil, D. 2017. The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory & Practice*, 17(1n): 321–335.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*, 461–464.

Slim, A.; Heileman, G. L.; Kozlick, J.; and Abdallah, C. T. 2014. Employing markov networks on curriculum graphs to predict student performance. In *13th International Conference on Machine Learning & Applications*, 415–418. IEEE.

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

Sulis, I.; Porcu, M.; and Capursi, V. 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): 1305–1331.

Sulis, I.; Porcu, M.; and Tedesco, N. 2011. Evaluating lecturer's capability over time. Some evidence from surveys on university course quality. In *New Perspectives in Statistical Modeling and Data Analysis*, 13–20. Springer.

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

Trcka, N.; Pechenizkiy, M.; and van der Aalst, W. 2010. Process mining from educational data. *Handbook of educational data mining*, 123–142.

Uttl, B.; White, C. A.; and Gonzalez, D. W. 2017. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54: 22–42.

van der Linden, W. J.; and Hambleton, R. K. 2013. *Handbook of Modern Item Response Theory*. New York, NY, USA: Springer.

Wong, W. Y.; and Lavrencic, M. 2016. Using a Risk Management Approach in Analytics for Curriculum and Program Quality Improvement. In *PCLA@ LAK*, 10–14.

Yen, W. M. 1993. Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, 30(3): 187–213.