Redefining Students Success

Samuel Berestizhevsky, Tanya Kolosova
YieldWise Inc.

Abstract

Creating successful learners is a challenging task for educational institutions. As a rule, the measurement of student learning performance is carried out using tests and exams. However, test and exam data is often subjected to inadequate analysis, which leads to incorrect conclusions about the progress of student learning and therefore misleading recommendations on how to improve the learning process.

Traditional assessment of students' proficiency is done by summing up or averaging raw scores of an exam. However, this approach does not consider differences in the difficulty of the exam questions, interdependencies of the questions as well as differences in the ability of the students. In this paper we show that the traditional assessment approach produces misleading results.

In this paper, we show how to analyze exam data using the Polytomous Rasch Measurement Model combined with the Relational Bayesian Networks methodology. We demonstrate that assessment of students' proficiency using these methods is realistic, accurate and reliable. Such assessment is instrumental in creating the Student Success Profile for each course. These profiles help to create actionable recommendations for addressing gaps in student education, and eventually help educational institutions to develop better and more successful learners, identify and handle issues in the educational process before they become problems, and, ultimately, significantly reduce students' attrition.

Keywords: Polytomous Rasch Measurement Model, Relational Bayesian Networks, Student Success Profile, Item Characteristic Curve.

Introduction

A series of knowledge exams are commonly used to assess a student's qualifications. However, the raw scores of these exams are often incorrectly analyzed. This leads to incorrect conclusions about students' strengths and shortcomings, and leads to misleading recommendations on how to eliminate these shortcomings. The reasons for incorrect analysis often originate from the misuse of raw exam scores. When students are evaluated through a series of knowledge exams, it is tempting to manipulate the raw exam data using simple mathematics. However, researchers agree that the use of raw scores to assess and compare students' achievement is erroneous.

The Difficulty of Items (Questions)

Do students exert an equal amount of effort to answer each item (question) in a knowledge exam? The answer is "No," since it is highly unlikely that all items are of similar or equal complexity. In some cases, educators try to solve this problem by assigning a different number of points (weights) to items, thereby reflecting the educator's perception of varying complexity of the items. However, it is the students' ability that determines the degree of difficulty of the exam item. Measuring students' proficiency (or qualification) using sum or average of raw exam scores, while ignoring the difficulty of various items and different abilities of students, creates misleading results. To solve the problem of correctly measuring student achievement, we use the Polytomous Rasch Measurement Model (PRMM), which correctly analyzes the raw exam scores, while simultaneously assessing the difficulty of the items as well as the students' abilities.

Foundational Items

In any course, knowledge of various topics is interdependent. Thus, we cannot assume that the knowledge required to answer one exam item does not depend on the knowledge necessary to answer other items. Suppose one exam question (item) tests the knowledge of a specific mathematical technique, and two other questions evaluate the use of this technique for solving problems. A lack of knowledge needed to answer the first question leads to failure to provide correct answer to the other two questions. On the other hand, sufficient knowledge of the first question increases chances of success in the two related questions – thus, the first question is considered foundational. Such relations among exam items are not always simple and obvious; they may include dependencies on more than one item and therefore are not easy to detect. Identification of the foundational items is essential for the continued success of students. We use the Relational Bayesian Networks (RBN) to solve this problem successfully.

Overview of the Approach

In this paper, we propose an approach of extracting actionable insights from knowledge exams. This approach has five features that do not exist in traditional methods of measuring the effectiveness of learning processes.

Ability of Students and Difficulty of Items: The proprietary algorithm of the PRMM can process incomplete data (for example, missing values) and provide a reliable estimate of the difficulties of items and abilities of students. This provides educators with not only accurate information about true achievements of the students but also identifies malfunctioning (or faulty) items. Eliminating such items improves the quality of exams.

Causal Relationships Among Items: We identify cause-effect relationships among exam items and identify foundational items. This functionality is implemented using the RBN methodology. The identification of foundational items, the causal relationships among the items, and their dependence on the abilities of the students play an important role in the development of the Student Success Profile.

Student Success Profile: Using the results of the PRMM and the RBN, we create Student Success Profiles for each course. The most important outcome is the quantitative values for the components of the Student Success Profile. Using Student Success Profiles, universities and colleges can determine the threshold of their students' ability to ensure the students' success in their studies.

Student Proficiency Cards: We also create Student Proficiency Cards that contain individual students' proficiency in each item and serve as a basis for determining their overall proficiency in the course. Based on these Cards, we compose recommendations for personalized training programs that improve students' abilities, eliminate existing gaps, and increase students' chances of success in the course.

Students Class Strengths and Gaps: We aggregate data from individual Student Proficiency Cards to evaluate class competency in the course. Educators can use this information to identify possible gaps in the course and ways to address them.

Case Study

The final exam of the course "Structure and Interpretation of Computer Programs" was held for 2nd-year students of the Faculty of Electrical Engineering and Computer Science. The exam was given to forty students. It contained 16 questions, each graded as either 1 (F), 2 (D), 3 (C), 4 (B), or 5 (A). The traditional grading system based on the averaging of raw exam scores presented only the following distribution of grades in the course: 2 out of 40 (5%) students received a B, 30 (75%) students received a C, 8 (20%) students received a D. The traditional grading method couldn't explain the reasons for the many low grades (C and D). We applied the proposed approach to analyze the exam data and to make informative and actionable conclusions.

Difficulty of Items

The difficulty of the exam items, assessed by the PRMM, reflects how easy or hard it was for students to answer each question. Items with lower difficulty are easier to answer for students, and items with higher difficulty are harder to answer for students. Figure 1 demonstrates a substantial difference in the difficulty level of each item. Ignoring this important data in the analysis creates false conclusions.

The PRMM not only estimates the difficulty of the items but also determines the OutFit value, an outlier-sensitive fit. The OutFit is a mean-square residual summary statistic, which has expectation 1.0, and a range from 0 to infinity. The OutFit value greater than 1.0 indicates underfit to the Rasch model, meaning the data is less predictable than the model expects. An OutFit value of 1.3 (see Table 1) indicates that there is 30% more randomness in the data than modeled, and that the level of difficulty of the item does not always correspond to the ability of students. There are 3 items shaded gray in Table 1, for which the OutFit value is greater than 1.3, so they appear to be malfunctioning. It is possible that the questions were not clearly worded, may contain errors, or may have other causes leading to a misunderstanding.

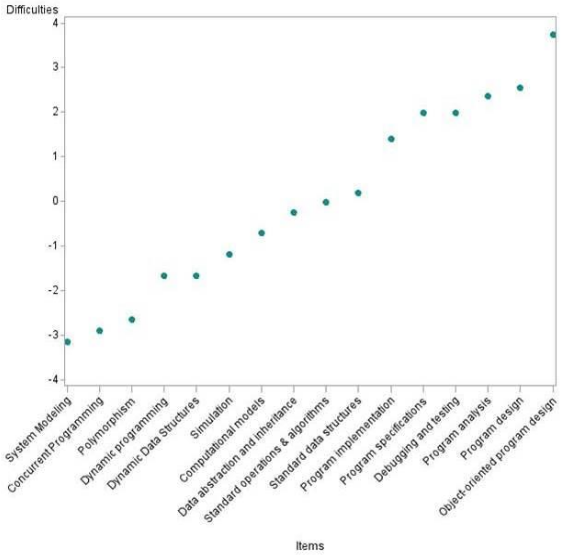


Figure 1. Difficulty of Exam Items

Table 1. Items Difficulty

#	Item	Difficulty	OutFit
1	System Modeling	-3.14	3.04
2	Concurrent Programming	-2.90	0.90
3	Polymorphism	-2.65	0.87
4	Dynamic Programming	-1.67	0.65
5	Dynamic Data Structures	-1.67	0.46
6	Simulation	-1.19	1.34
7	Computational Models	-0.71	0.91
8	Data Abstraction and Inheritance	-0.25	0.70
9	Standard Operations & Algorithms	-0.02	0.60
10	Standard Data Structures	0.19	0.58
11	Program Implementation	1.41	0.50
12	Program Specifications	1.98	0.44
13	Debugging and Testing	1.98	0.58
14	Program Analysis	2.36	0.75
15	Program Design	2.55	5.56
16	Object-Oriented Program Design	3.73	1.13

Ability of Students

While making conclusions about the students' performance, it is very important to evaluate their abilities taking into account the difficulties of the exam questions. Such conclusions reflect the real proficiency of the students, as opposed to the raw exam scores. The ability of students is estimated by the PRMM, which is conditional on the difficulty of the items: small numbers mean lower student abilities and large numbers indicate higher abilities. Figure 2 shows that students have different abilities, and this must be taken into account when analyzing.

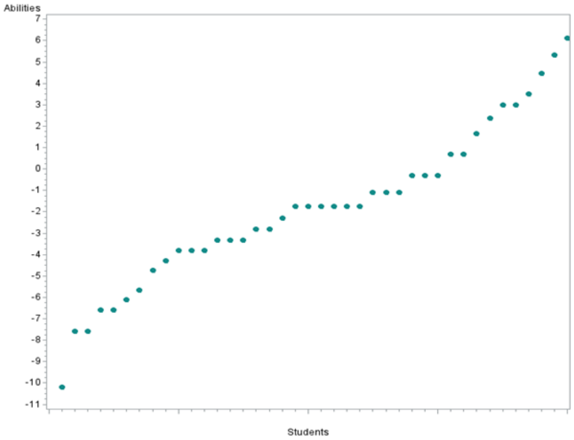


Figure 2. Ability of Students

Item Characteristic Curve

The PRMM creates Item Characteristic Curves (ICC) that describe the relationship between student ability and the likelihood (probability) that students will receive a specific score (Grade Category). Each item in the exam has its ICC, and for each item the probability that each student will answer a particular question correctly is estimated. For example, the ICC in Figure 3 is created for the "Dynamic Data Structures" item. Each ICC curve represents the probability that students will receive a specific score (Grade Category), depending on their ability. Thresholds (solid vertical lines) determine the ability for which the probabilities of adjacent scores (Grade Categories) are equal. The red dots on the curves denote the actual students' scores (Grade Categories).

Point A is located on the green curve (Grade Category 2) and represents student SID038 with the ability of -3.81. This student received a score of 2, while the probability of obtaining this score with their ability is only 0.11. The broken blue vertical line that passes through point A meets the yellow curve (Grade Category 3) at the point corresponding to the probability of 0.89. This indicates that the student SID038 has a probability of 0.89 to get a score of 3 instead of 2.

Point B is located on the pink curve (Grade Category 4) and represents student SID005 with an ability of 2.99. This student received a score of 4, while for their ability, the probability of getting this score is only 0.30.

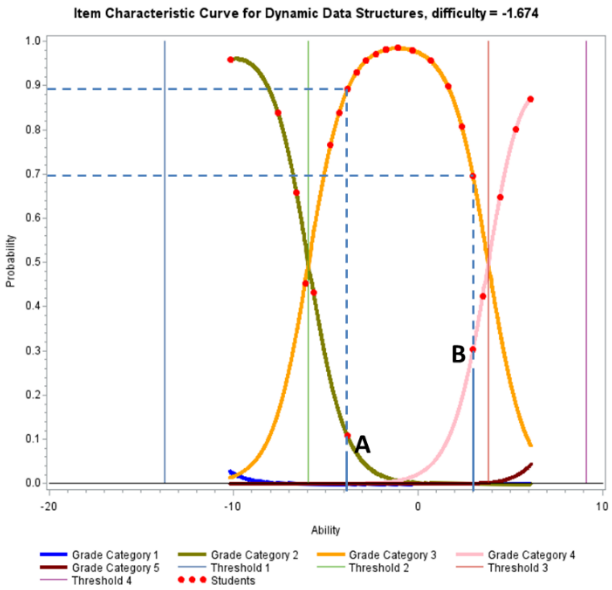


Figure 3. ICC for "Dynamic Data Structures" Item

The broken blue vertical line reaches the yellow curve (Grade Category 3) at the point where the corresponding probability of getting a score of 3 is 0.70. Can a score 4 for this student be a lucky guess?

Causal Relationships Among Items

Identifying the causal relationships among the exam items allows us to determine which foundational knowledge contributes to success in the exam. We use RBN methodology to identify probabilistic causal relationships among the exam items and the ability of the students. The RBN visualizes the dependence of one item on another in the form of a graph (see Figure 4). The arrows in the graph reflect how the students' knowledge of one item affects competency in another item. According to the created RBN, the ability of students is influenced directly (arrows originated from green-colored items) by their knowledge of four items: Dynamic Programming, Computational Models, Standard Operations and Algorithms, Standard Data Structures.

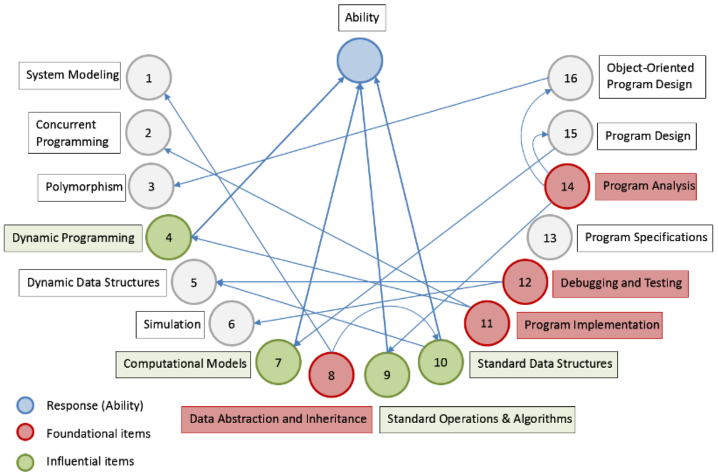


Figure 4. Relational Bayesian Networks

In the educational process, it is important to determine which items are foundational. We identify foundational items as those that are critical for improving students' abilities in the course. The following four items are identified as foundational: Data Abstraction and Inheritance, Program Implementation, Debugging and Testing, and Program Analysis.

Student Success Profile

The results produced by the RBN and the PRMM – a set of foundational and influential items, as well as the probabilities of obtaining specific scores on these items – form the basis of the Student Success Profile for the course. In this case study, we create a Student Success Profile for the course "Structure and Interpretation of Computer Programs." Components of the Student Success Profile for the course are exam items that are associated with the lowest scores necessary for students to master the course successfully.

The Student Success Profile states:

- Which items should be considered for success, and which should be excluded (items highlighted in gray were identified as faulty and were excluded from the Success Profile).
- What is the lowest score a student should get for each item to be considered proficient.
- Which of the items tested during the exam are foundational (red bordered) for the success of a student in the course.

Table 2. Student Success Profile for "The Structure and Interpretation of Computer Programs"

#	Item	Difficulty	Score	Item Importance
1	System Modeling	-3.14	4	Excluded
2	Concurrent Programming	-2.90	4	
3	Polymorphism	-2.65	4	
4	Dynamic Data Structures	-1.67	3	
5	Dynamic Programming	-1.67	3	
6	Simulation	-1.19	3	Excluded
7	Computational Models	-0.71	3	
8	Data Abstraction and Inheritance	-0.25	3	Foundational
9	Standard Operations & Algorithms	-0.02	3	
10	Standard Data Structures	0.19	3	
11	Program Implementation	1.41	3	Foundational
12	Debugging and Testing	1.98	3	Foundational
13	Program Specifications	1.98	3	
14	Program Analysis	2.36	3	Foundational
15	Program Design	2.55	3	Excluded
16	Object-Oriented Program Design	3.73	3	

Student Proficiency Cards

In this case study, 75% of students were given a grade of C by the traditional approach. Is a C a good enough grade to be successful in this course? Do these 75% of students have the same level of proficiency in the course? Are these students on the road to success? We provide answers to these questions using Student Proficiency Cards, the Student Success Profile for the course, and ICCs.

Item Level Proficiency: the Student Proficiency Card determines the level of competency of the student in each exam item and associates it with the importance of the item and the required level of knowledge:

- Strength – the student exceeds the score requirement for the item in the Success Profile,
- Fit – the student meets the score requirement for the item in the Success Profile,
- Opportunity to Fit – the student has a high probability of meeting the score requirement for the item in the Success Profile,

- Gap – the student’s actual and expected scores are lower than the score required for the item in the Success Profile.

Student Proficiency Cards contain actual and expected (assessed by the PRMM) scores for each item.

Exam Level Proficiency: Student Proficiency Cards are the basis for determining the proficiency of each student in the exam:

- Exceeds Proficiency (A) – the student demonstrates Strength in all items of the Success Profile,
- Proficient (B) – the student shows Strength or Fit in all items of the Success Profile,
- Foundational Proficiency (C) – the student shows Strength or Fit in all foundational items of the Success Profile,
- Partially Proficient (D) – the student shows Strength, Fit or Opportunity to Fit in all foundational items of the Success Profile,
- Insufficient Proficiency (F) – the student was not classified in any of the four groups mentioned above.

Student Proficiency Cards presented in Table 3 below contain the following data:

- Gray-shaded items are excluded from consideration being identified as malfunctioning.
- Red-bordered items were identified as foundational.
- Light-green cells determine the highest probability of scores for each item.
- The “Actual Score” column contains the score obtained by the student in a particular item.
- The “Most Likely Score” column contains the score that, according to the PRMM, is most probable for the student to obtain (see the probabilities in the light-green shaded cell).

Let’s look at two examples. Student SID018 with the ability -1.73 is Partially Proficient (D), as they exhibit Opportunity to Fit in the foundational item “Debugging and Testing” (see Table 3).

Table 3. Proficiency Card for Student SID018, Ability -1.73, Partially Proficient (D)

Item	Actual Score	Prob. Score 1	Prob. Score 2	Prob. Score 3	Prob. Score 4	Prob. Score 5	Most Likely Score	Success Profile	Status
System Modeling	3	0.00	0.00	0.98	0.02	0.00	3	4	Gap
Concurrent Programming	3	0.00	0.00	0.98	0.01	0.00	3	4	Gap
Polymorphism	3	0.00	0.01	0.98	0.01	0.00	3	4	Gap
Dynamic Data Structures	3	0.00	0.02	0.98	0.00	0.00	3	3	Fit
Dynamic Programming	3	0.00	0.02	0.98	0.00	0.00	3	3	Fit
Simulation	3	0.00	0.02	0.97	0.00	0.00	3	3	Fit
Computational Models	3	0.00	0.04	0.96	0.00	0.00	3	3	Fit
Data abstraction and Inheritance	3	0.00	0.06	0.94	0.00	0.00	3	3	**Fit**
Standard Operations & Algorithms	3	0.00	0.07	0.93	0.00	0.00	3	3	Fit
Standard Data Structures	3	0.00	0.09	0.91	0.00	0.00	3	3	Fit
Program Implementation	3	0.00	0.25	0.75	0.00	0.00	3	3	**Fit**
Debugging and Testing	2	0.00	0.37	0.63	0.00	0.00	3	3	**Opportunity**
Program Specifications	3	0.00	0.37	0.63	0.00	0.00	3	3	Fit
Program Analysis	3	0.00	0.46	0.54	0.00	0.00	3	3	**Fit**
Program Design	2	0.00	0.51	0.49	0.00	0.00	2	3	Gap
Object-Oriented Program Design	2	0.00	0.77	0.23	0.00	0.00	2	3	Gap

Another student, SID029, with the same ability as student SID018, has Foundational Proficiency (C) as they demonstrated Fit for all foundational items (see Table 4). This student also shows Gap in proficiency: “Concurrent Programming,” “Polymorphism,” and “Object-Oriented Program Design.”

Table 4. Proficiency Card for Student SID029, Ability -1.73, Foundational Proficiency (C)

Item	Actual Score	Prob. Score 1	Prob. Score 2	Prob. Score 3	Prob. Score 4	Prob. Score 5	Most Likely Score	Success Profile	Status
System Modeling	3	0.00	0.00	0.98	0.02	0.00	3	4	Gap
Concurrent Programming	3	0.00	0.00	0.98	0.01	0.00	3	4	Gap
Polymorphism	3	0.00	0.01	0.98	0.01	0.00	3	4	Gap
Dynamic Data Structures	3	0.00	0.02	0.98	0.00	0.00	3	3	Fit
Dynamic Programming	3	0.00	0.02	0.98	0.00	0.00	3	3	Fit
Simulation	3	0.00	0.02	0.97	0.00	0.00	3	3	Fit
Computational Models	3	0.00	0.04	0.96	0.00	0.00	3	3	Fit
Data abstraction and Inheritance	3	0.00	0.06	0.94	0.00	0.00	3	3	**Fit**
Standard Operations & Algorithms	3	0.00	0.07	0.93	0.00	0.00	3	3	Fit
Standard Data Structures	3	0.00	0.09	0.91	0.00	0.00	3	3	Fit
Program Implementation	3	0.00	0.25	0.75	0.00	0.00	3	3	**Fit**
Debugging and Testing	3	0.00	0.37	0.63	0.00	0.00	3	3	**Fit**
Program Specifications	2	0.00	0.37	0.63	0.00	0.00	3	3	Opportunity
Program Analysis	3	0.00	0.46	0.54	0.00	0.00	3	3	**Fit**
Program Design	2	0.00	0.51	0.49	0.00	0.00	2	3	Gap
Object-Oriented Program Design	2	0.00	0.77	0.23	0.00	0.00	2	3	Gap

Based on the Students Proficiency Cards, we can create recommendations on how to improve students’ abilities, eliminate existing gaps, and increase the chances of success in the course.

Comparison of Students

We can distinguish among students who cannot be differentiated using traditional averaging (or summation) of scores. Let’s look at the examples of two Student Proficiency Cards mentioned above in Table 3 and Table 4: using the traditional method, students SID018 and SID029 received the same average value of their scores: 2.81. That is, they have the same grade C in the course (grade C is associated with score 3, which is the closest to the average value of 2.81). However, we revealed that these students are different. Proficiency Cards demonstrate that the student SID029 is in Fit with all the foundational items, and thus demonstrates Foundational Proficiency (C) in the course. The student SID026 is in Fit with only three out of four foundational items, and therefore only Partially Proficient (D). The traditional approach failed to reveal this critical difference.

Students Class Strengths and Gaps

Students’ proficiency in the exam is based on the scores obtained for the foundational items, and not on all items, where some of them may have low importance or just be derived from the foundational items: 4 out of 40 (10%) students are Proficient (B), 11 (27.5%) students have Foundational Proficiency (C), 8 (20%) students are Partially Proficient (D), 17 (42.5%) students have Insufficient Proficiency (F). The exam grades calculated by the proposed approach are significantly different from the exam grades calculated using the traditional averaging method. The traditional approach assigned the same grade C to thirty students (75%), while the proposed approach allowed differentiating of these students. We will show that the traditional method (Figure 5) failed to identify students with Insufficient Proficiency (F). We identified 17 students who have insufficient knowledge of the foundational items (see Figure 6). Although these students have passed the exam per the traditional approach, in the future they are at risk of attrition or failure.

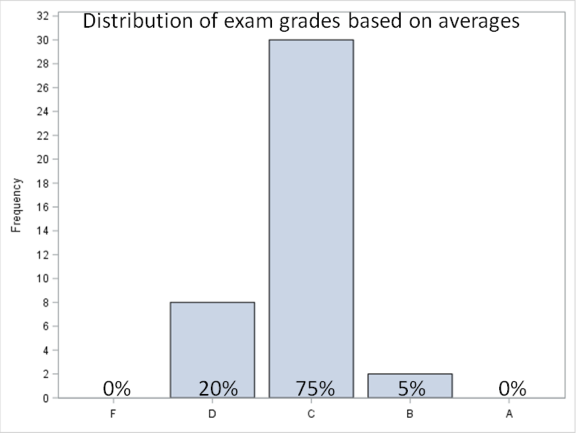


Figure 5. Distribution of Grades Based on Averages

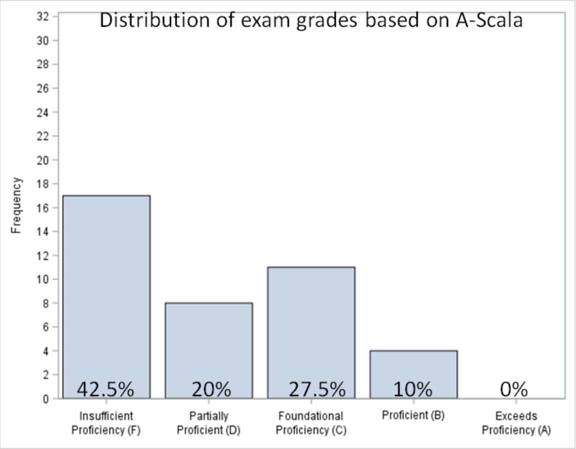
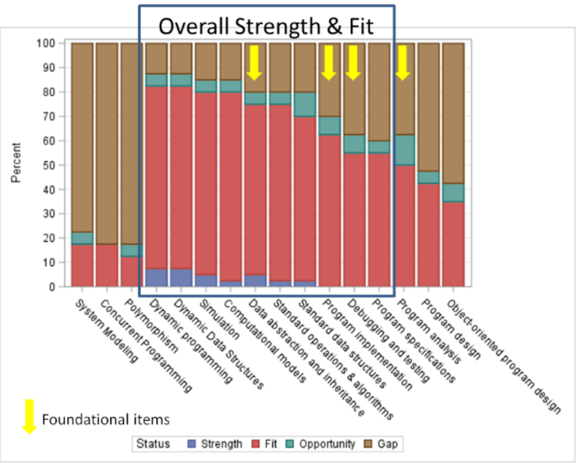


Figure 6. Distribution of Grades Based on the Proposed Approach

Strengths and Gaps in the Course

We identify areas of common strengths and gaps in the course, thus providing effective and actionable feedback to the teacher. The bar chart in Figure 7 displays the following insights: (i) the first three exam items from the left, although being the easiest, present Gaps regarding the Success Profile; (ii) the three most difficult items (from the right) have 50% or less Fit to the Success Profile; (iii) foundational items determine the most important topics of the course, and in one of them, "Program Analysis" item, less than 50% of students showed Fit; (iv) ten items in the blue square outline topics in which students mostly demonstrate Strength and Fit.



Items are ordered from the lowest difficulty to the highest

Figure 7. Students' Strength & Fit to the Course

Conclusion

According to the traditional measurement of proficiency, students may receive grades that allow them to pass exams. However, the reality can be different. Our approach helps to identify the real situation with the students' proficiency in the course. It provides a correct and accurate measurement of student proficiency and provides informative and actionable recommendations for addressing gaps in student education. The approach helps educational institutions to develop successful learners, to identify and handle issues in the educational process before they become problems, and, ultimately, significantly reduce students' attrition.

References

Fischer G.H., Molenaar I.W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer-Verlag.

George E.Jr. (1992). The Measurement of Writing Ability With a Many-Faceted Rasch Model. *Applied Measurement in Education.* Vol 5(3) 171-191

Introduction to Statistical Relational Learning (2007). Getoor L., Taskar B., Francis Bach F. (Eds.), Cambridge: MIT Press.

Kersting K., De Raedt L. (2001) Adaptive Bayesian Logic Programs. Inductive Logic Programming. *ILP 2001. Lecture Notes in Computer Science*, vol 2157. Berlin: Springer.

Linacre J.M. (2004). Rasch Model Estimation: Further Topics. *Journal of Applied Measurement,* 5(l),95-110. University of the Sunshine Coast, Australia.

Preisach C., Schmidt-Thieme L. (2006) Relational Ensemble Classification. *IEEE Sixth International Conference on Data Mining (ICDM'06)*

Wright B.D., Stone M.H. (1979). *Best Test Design*. Chicago: MESA Press

Wu M., Adams R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach.* Melbourne: Educational Measurement Solutions.