KJEP 16:1 (2019), pp. 3-20

Analysis of bachelor's degree curricula through Item Response Theory and Association Rules

Hayette Khaled
Université de Bejaia, Algeria
Pablo Gregori
Universitat Jaume I de Castellón, Spain
Raphaël Couturier
Université Bourgogne Franche-Comté (UBFC), France

Abstract

The objective of this paper is to provide a methodology for statistical analysis of curricula, whose results give insights, on the one hand, on the relation between fluency in each subject of a bachelor's degree and the overall ability of learners in that degree. On the other hand, an ordering relation among the fluency of the different subjects, expected or not, emerges from the observed data. We illustrate it with the analysis of a bachelor's of computer science, to which we have applied the Graded Response Model of Item Response Theory, along with the implicative graph of Association Rules.

Keywords : bachelor's degree curriculum, student marks, Graded Response Model, Item Response Theory, Association Rules

Introduction

Item Response Theory (IRT) is a framework which emerged from psychometrics in the 1950s (Embretson & Reise, 2000; Van der Linden & Hambleton, 1997) and whose aim is to provide efficient tools for measuring *overall abilities* (or other features of individuals, called *latent traits*) through questionnaires. The total score is not simply the sum of each item score. Items are assumed to have different degrees of difficulty. Then the estimation of the ability of individuals is entangled in the estimation of the difficulty level of the items, in the same procedure.

IRT has also been applied to fields such as medicine (Thomas et al., 2013), and topics as smoking among adolescents (Hedeker, Mermelstein, & Flay, 2006). Furthermore, it is widely used in education to calibrate and evaluate the items of questionnaires as well as students' abilities. In recent decades, educational evaluation has been increasingly using techniques based on IRT in order to develop new tests, teaching strategies and skill selection (Bodin, 2010; Hamdare, 2014; Johns, Mahadevan, & Woolf, 2006; Rowan et al., 2001).

The goal of this paper is to provide curriculum designers, of a bachelor's degree, with a methodology for the analysis of their planned curriculum. The application of models of IRT to student marks in each subject, allows to describe features of the subjects, such as difficulty or discrimination, and give more perspectives for their comparison. It also provides with the estimation of the overall ability of students (under the model), and allows comparison with their grade point average. We propose that the new pieces of information are useful in the process of revision of curricula, in the intensive dialogue between educational developers and academic staff, as proposed by O'Neill (2010).

As a complement to IRT, association rules will be used on the same dataset in order to understand the mutual relationship among the skills of students in each subject. This topic emerged from the market basket analysis, in the search of patterns of items sold frequently together. Statistical Implicative Analysis (SIA) allows to detect important rules that can be represented as an oriented graph, called *implicative graph*, and describe the set of relations found among the variables, very helpful for the interpretation of eventually underlying causality relations.

Khaled, Ghanem, and Couturier (2014) used the implicative graph in order to discover the relationship among the success of students in the different subjects of a bachelor's of computer science. This paper improves that methodology by adding the use of an IRT model, providing insightful information on how each and every subject contributes to the overall ability in the bachelor's degree.

In Section 2, the Graded Response Model of IRT is introduced and Section 3 reviews the basics of quality measures in association rules. Section 4 presents our illustrative example, our questions and the answers provided by GRM model and rule mining. Finally, our findings are compared with previous research in Section 5.

The Graded Response Model

Test theories try to measure a target *latent trait,* which is present in the population at different levels (a numeric value denoted by letter θ). The level is estimated through questionnaires (tests). In the Classical Test Theory, the global score of the test is the addition of the scores of single items, hence items must be designed carefully in order to measure with equal intensity the latent trait. In the IRT setting, it is accepted that the items might measure the latent trait with different relevance. Then, the estimation of the latent trait of individuals involves a double process, where the relevance of items is estimated in the same procedure.

In our application, the population is the set of students of a given bachelor's degree, and the latent trait θ is their *overall ability within the given discipline*. Then each degree's subject is considered to be an item, and the mark of a student at the given subject is the score attained in that item.

The simplest models proposed by IRT are for items scoring *pass/fail* (i.e., binary or dichotomous). In general, the probability that a student with ability level θ will pass a subject *i* is an increasing function given by the logistic function:

- Rasch *(or 1P, one-parameter) model:* $\frac{\exp(\theta - b_i)}{1+\exp(\theta - b_i)}$, where b_i is the *difficulty* parameter.
- *2P model:* $\frac{\exp\ a_i(\theta - b_i)}{1+\exp\ a_i(\theta - b_i)}$, where the new *discrimination* parameter a_i determines the speed of success with regard to θ.
- *3P model:* $C_i + (1 - C_i)\frac{\exp\ a_i(\theta - b_i)}{1+\exp\ a_i(\theta - b_i)}$ where C_i is the probability of passing just by chance (as in multiple choice test without penalty).

Figure 1 (left) shows an example of curve (called, in general, *Item Characteristic Curves*, ICC) for each model. They help to interpret the difficulty and the discrimination capacity of each item.

A more realistic situation is to consider items that can score several ordered levels (for instance, *success*, *partial success* and *fail*). Ordinal items, Likert scales, scale assessment responses, graded test responses, etc., lead to this type of ordered polytomous data. Several models have been proposed for these data: the Partial Credit Model (Masters, 1982), the Generalized Partial Credit Model (Muraki, 1992), and the Graded Response Model (GRM) (Samejima, 1969), to which we shall attach:

- *GRM.* Item i with possible scores 0, 1, ⋯, k, is characterized by a discrimination parameter $a\,i$, and then a set of increasing difficulty parameters (b_{io} = -∞, b_{i1}, ⋯, b_{ik}) such that the probability of scoring "level k or more" is given

by $\frac{\exp\ a_i(\theta - b_{i,k})}{1+\exp\ a_i(\theta - b_{i,k})}$. Figure 1 (right) shows an example of the Item Category Characteristic Curves (ICCC) for an item with 4 levels under the GRM.

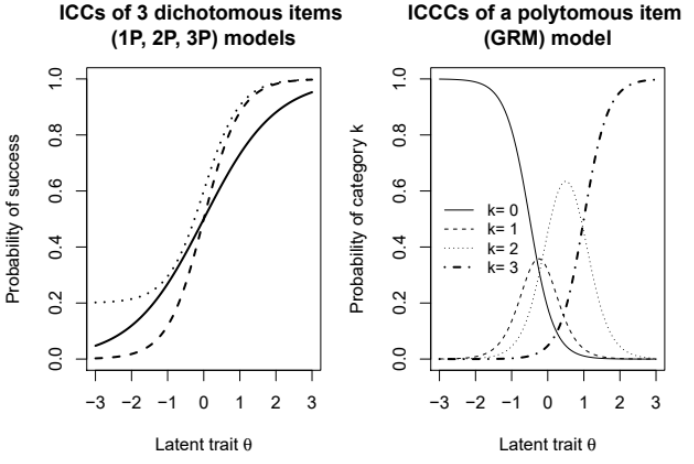

Figure 1. (Left:) Item Characteristic Curve (ICC) for 3 dichotomous items. Item 1 (continuous line) has only difficulty parameter b_1= 0. Item 2 (dotted line) has the same difficulty parameter, and discrimination parameter α_2= 2 (more discriminatory than Item 1 for individuals with latent trait values around 0). Item 3 (dashed line) has the same difficulty and discrimination parameters, and a guessing parameter C_3= 0.2(a non negligible chance of success just by guessing). (Right:) Item Category Characteristic Curve (ICCC) for a unique polytomous item with categories 0, 1, 2 and 3, modelled with GRM under α_i = 3, b_{i1} = -0.5, b_{i2}= 0, b_{i3}= 1. For instance, individuals with low level of latent trait ($\theta \leftarrow 1$) are more likely to get categories 0 or 1; individuals with an average level (0 < θ < 0.5) are more likely to get categories 2 or 1; and individuals with high level trait (θ > 1) are more likely to get categories 3 or 2.

In the sequel, the GRM shall be applied to our dataset of students' marks —which needs to be transformed into three ordered categories labeled "weak", "average" and "strong"— in order to estimate: (1) the global level of proficiency of students in their discipline, (2) the difficulty and discrimination of every subject, and (3) the distribution of overall students' ability, in comparison with the distribution of their given marks. We discard using a model for continuous data, because it requires a larger amount of data for the model fit, and the interpretability of its results becomes more complex, hence less practical.

Association Rules and statistical implicative analysis

Educational research is interested in finding patterns underlying the educational processes: the way students learn, how teaching strategies lead to results, and so on. Patterns can take the form of segmentation (cluster analysis), explanation (factor analysis), prediction or classification (regression models), etc. An interesting and fruitful approach is the search for association rules (i.e., potential causes and effects), hidden in the complexity of the processes. Statistics turns a set of raw data into a set of rules. Experts have the role of turning some of those rules into new pieces of knowledge, disclosing the underlying rational arguments and discarding the other ones because of mere chance or lack of explanation.

Rules usually link two properties that individuals in the sample may hold or not. Association rules emerged from the market basket analysis, linking goods bought jointly in market transactions and were efficiently found with the Apriori algorithm (Agrawal & Srikant, 1994). Let us denote by a and b a couple of properties —that individuals can hold or not, and also the binary variables corresponding to the properties (value 1 for individuals holding the property and 0 otherwise). If we ponder the rule $a \rightarrow b$, we can measure the following aspects:

- The *support* is the proportion of individuals holding both properties a and b. A high support means a strong association, but it is not a clue of causality, since it is a symmetric definition.
- The *confidence* is the proportion of individuals satisfying b within the group of individuals satisfying a. A high confidence means that b is likely to occur in individuals holding property a. It is the most popular and intuitive measure of the strength of a rule. However, if property b is too frequent in the sample, a large confidence does not necessarily imply that a induces a positive effect on b, and the rule should not be highlighted.
- When one is interested in detecting whether property a raises the chances of property b, other measures must be used, such as the *lift* (the ratio between the confidence of the rule and the frequency of the consequent b), or *Loevinger's* index H (Loevinger, 1947), used in Psychometrics.

A value of lift larger than 1 (or $H > 0$) means that an individual holding property a has more chances of holding also property b, than other individuals from whom we do not know anything. For instance, a lift of 1.5 improves the chances of finding property b by 50% within individuals already holding property a. This value can occur in rules of either low or high confidence, so both quality measures should always be reported together.

The *implication intensity* (denoted by φ) is the result of testing whether the value of lift is *significantly* larger than 1 in the whole population. It was introduced in Lerman, Gras, and Rostam (1981a, 1981b), and was mainly applied to Mathematics Education and Social Sciences (Gras & Totohasina, 1995; Gras et al., 1996; Batanero , Navarro-Pelayo, & Godino,

1997; Pantziara, Gagatsis, & Elia, 2009; Fazio, Battaglia, & Di Paola, 2013), but it has been growing in methodology and application fields (Gras & Couturier, 2011; Gras, Kuntz, & Briand, 2003; Gras, Régnier, Marinica, & Guillet, 2013; Gras, Suzuki, Guillet, & Spagnolo, 2008). *Statistical Implicative Analysis* was the given name to a set of tools for cluster analysis of variables, association rules, and cluster analysis of association rules (Gras & Kuntz, 2006). Another approach for the statistical implication that we are not adopting in this work, but which is nevertheless worth mentioning, was developed in the line of (Bernard, 2002; Bernard & Charron, 1996; Bernard & Poitrenaud, 1999).

The computation φ is available through software CHIC (Couturier, 2008; Gras, Ag Almouloud, Ratsimba-Rajohn, & Couturier, 2017) as well as through the R Statistical Software (R Core Team, 2013) by a freely available package called *rchic* (Couturier, 2017). The set of significant rules (i.e., exceeding the given threshold) is depicted as a directed graph, showing a neat network of variables (properties) linked by arrows (where intensities are color coded, and the value of confidence can be added, see the results of Section 4-3). This structure is an opportunity for researchers: some rules can be meaningful and found to be the basis of a new empirical hypothesis.

Analysis of a bachelor's degree curriculum by students marks questions and data description

The proposed methodology is the application of IRT and Association Rules to students' marks of subjects conforming a bachelor degree. It shall be illustrated with a particular bachelor's degree, but it can be replicated in any level of studies and discipline, working with the corresponding team of experts in the field. The main concerns can be summarized by the following questions:

- If we assume an *overall ability* of students of the bachelor's degree, how can it be estimated from the students' marks? Does every subject measure that ability in the same way?
- How is the ability of students in each particular subject related to the ability at other subjects?

In order to answer these two questions, the focus was put on 2nd-year students. First year students are starting their degree, and it could be said that their data may contain noise, in the sense that students have not very homogeneous backgrounds or prerequisites, and some of them can even find out that the particular degree is not what they expected it to be. Third year students have defenses, which affects marks in the other regular subjects.

The survey was conducted with 2nd-year students of a bachelor's of Computer Science of *University A / Mira Bejaia* (Algeria). Data were collected from academic courses 2010-2011, 2011-2012 and 2012-2013. Since all of them have shown very similar results, we illustrate our findings with the ones of the last year.

These students are trained and assessed in the following 15 subjects: Architecture (Arch), Data Structures (DStr), Information Systems (IS), Numerical Analysis (NumA), Probability and Statistics (PS), Mathematical Logic (MatLog), Signal Processing (SP), English 1 (EngFall), Algorithm and Data Structures (AlDStr), DataBases (DB), Operating Systems (OS), Languages Theory (LTh), Linear Programming (LP), Software Engineering (SwE) and English 2 (EngSpring).

Our raw data is a table with the students' final marks in each subject (students are rows and subjects are columns, see Tab. 1). Marks are in the usual scale from 0 to 20 as inherited from the French Education System.

Table 1

First rows and columns of original data (marks of students per module in the usual range [0,20]) and their transformation into ordinal data (weak, average and strong).

	Arch	DStr	IS	NumA	PS	MatLog	SP	EngFall
12MI0051	8.83	13.00	9.33	2.83	5.47	7.67	8.50	15.58
10MI083	3.50	6.75	8.67	1.33	5.67	7.00	1.83	12.92
10MI094	11.00	8.12	7.33	11.50	12.00	12.00	13.67	10.25
12MI0086	10.00	9.12	6.50	6.33	16.00	13.00	11.67	15.17
12MI0060	10.33	6.25	7.00	7.83	16.00	9.50	10.67	15.25
11MI207	9.50	8.81	7.67	7.25	14.70	12.67	14.50	16.25
12MI0051	wea	ave	wea	wea	wea	wea	wea	str
10MI083	wea	wea	wea	wea	wea	wea	wea	ave
10MI094	ave	wea	wea	ave	ave	ave	ave	ave
12MI0086	ave	wea	wea	wea	str	ave	ave	str
12MI0060	ave	wea	wea	wea	str	wea	ave	str
11MI207	wea	wea	wea	wea	ave	ave	ave	str

In order to fit our data to the GRM and obtain conclusions about the connection among the overall ability and the different subjects, numeric data were transformed into three level ordinal data: *weak* (marks below 10; fail), *average* (from 10 and below 15; pass with a lower grade), and *strong* (marks 15 and above; pass with a higher grade). Table 1 presents the first rows and columns before and after the transformation.

The distribution of students' grades per subject is shown in Figure 2. It is easy to see that Numerical Analysis and Information Systems are the most difficult ones.

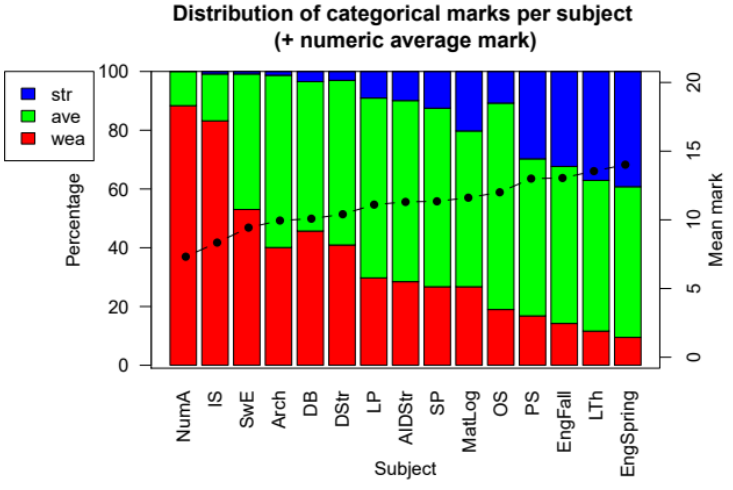

Figure 2. Distribution of categorical marks per subject (left axis). Subjects are ordered by numeric average mark (right axis). One can appreciate the least successful subjects NumA and IS, and the most successful ones (EngSpring, LTh, EngFall,PS,...).

The usual way of measuring a global score of students is the Grade Point Average, i.e., the averaging of marks of all subjects in the academic course (under the coding weak=0, average=0.5 and strong=1 in our case). That is what classical test theory would obtain as estimation of the overall ability of students. In the following section we study the distribution of these global results, along with the distribution of their general ability in Computer Science (estimated under the GRM), and the joint distribution of both measurements.

The overall ability of computer science students and the role of subjects according to the graded response model

Under the paradigm of IRT, the GRM model had a poor goodness of fit under the assumption of a difficulty parameter shared by all subjects. However, the fit was adequate when that restriction was relaxed.

The Item Category Characteristic Curve (ICCC) of each subject is shown in Figure 3. We can see that, even the best students in overall ability ($\theta \approx 6$) have a small probability of getting a strong mark in *IS*; NumA has no student with strong mark, and EngFall and EngSpring have flatter curves, meaning that marks in those subjects do not differentiate very well among weak or stronger overall ability students.

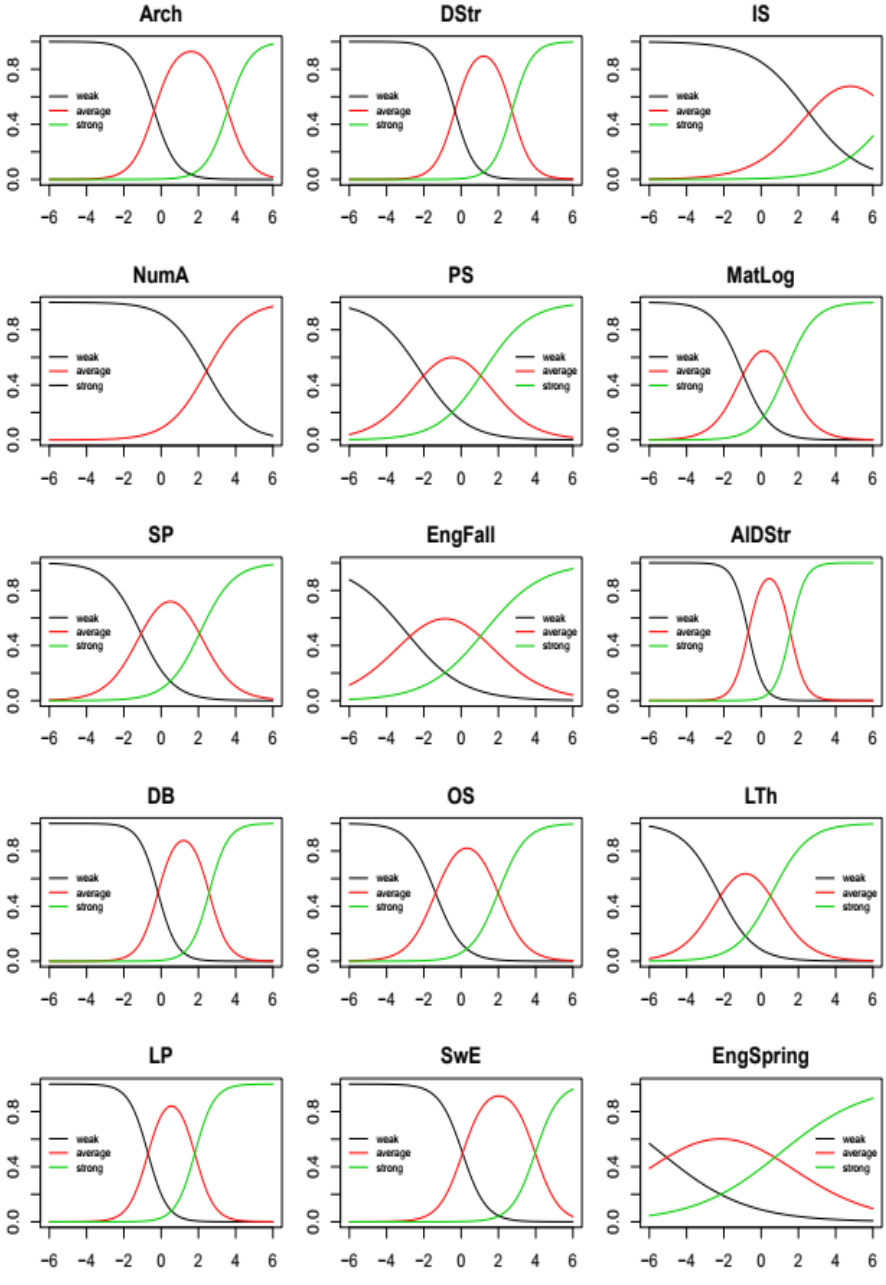

Figure 3. Plots of Item Category Characteristic Curve (ICCC) for each response category, for all the subjects. We can see for every level of overall ability θ, the chances of getting each category at each subject.

The Item Information Curve (IIC) explains, for each value of ability θ, the precision that each item (subject in our case) provides for the estimation of the particular level θ. Students with different, but very low levels of θ usually get the same grade in an item. Then, that item cannot distinguish between those low levels, and the precision for the estimation is very poor. The same situation holds at the high level θ side. In the mean values, items provide different precision depending on their difficulty and discrimination parameters. That is the reason why these curves are bell shaped. Figure 4 shows the information curves for all subjects, allowing comparison to how each shows more or less *information* (the inverse of variance or uncertainty) at the estimation of each ability level θ. Items provide more information (precision) at the estimation of levels around the intervals [-2, 0] and [1, 3].

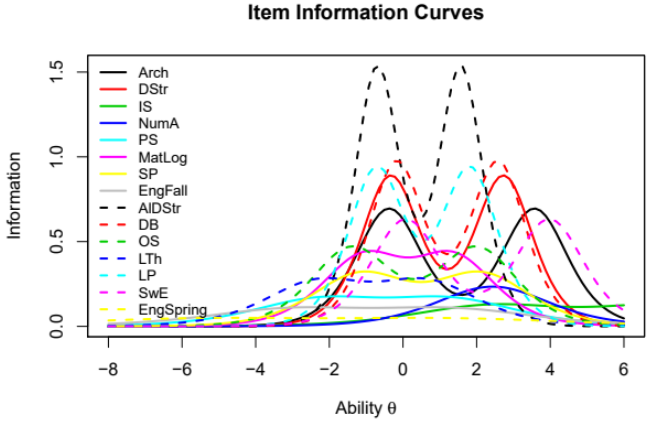

Figure 4. Item Information Curve (IIC) of each subject. It shows how precise (or informative) each subject is concerning the estimation of the ability of the student, at every level θ. Subjects such as EngSpring, EngFall, IS and PS provide very low precision, at any level θ. On the other hand, AIDStr shows the largest precision for θ in the central range [-2, 2], and modules such as DB, LP and DStr are also among the ones providing more precision at the estimation of θ in the same range.

The Test Information Function (TIF) is the sum of the IICs for all items and it is shown in Figure 5. We can see that the totality of subjects (all 2nd-year subjects' marks) gives a more precise estimation of students' ability level θ, when its true value lies between -1 and 3, and less precise estimation otherwise. It means that, for students with low and high overall ability, the estimation suffers from more sample variance and it can be considered as "less certain".

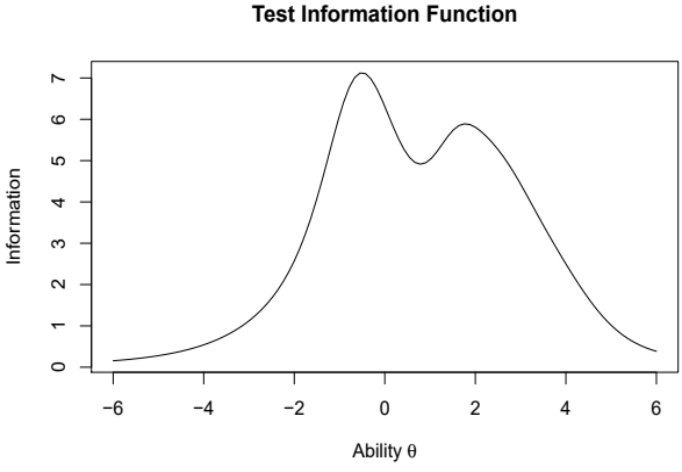

Figure 5. The Test Information Function shows that the estimation of the overall ability θ of students, from their marks in all subjects is more precise (i.e., has less variance) when the real value of θ is between -1 and 3.

Finally, the comparison among subjects is shown in Figure 6, where the plots of ICCCs are split into categories, reflecting how likely a given result is, for a generic student of ability level θ.

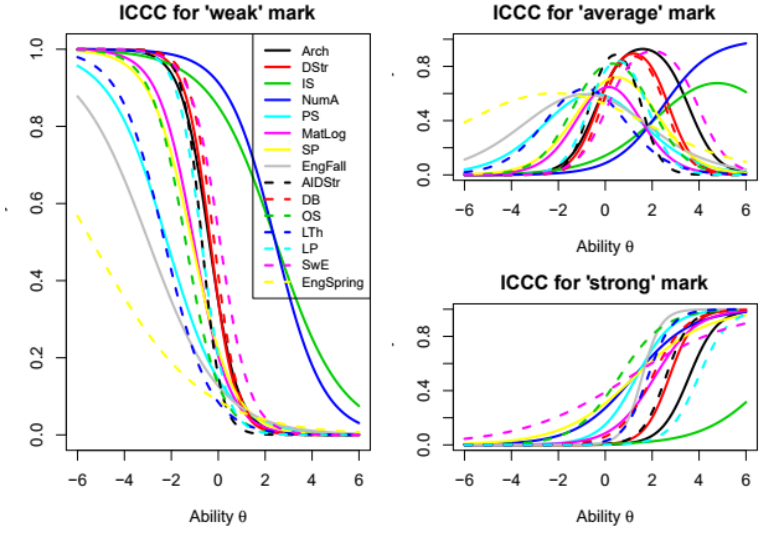

Figure 6. Plots of ICCCs split by category. At every plot (category) we can easily compare the probability of getting that category (mark) for all different subjects, given the particular ability θ of a generic the student.

Figure 7 shows the distribution of the overall results of students, seen both in the classical way as the GPA of the levels ("weak"=0, "average"=0.5, "strong"=1), and in the overall estimated ability proposed by our IRT model.

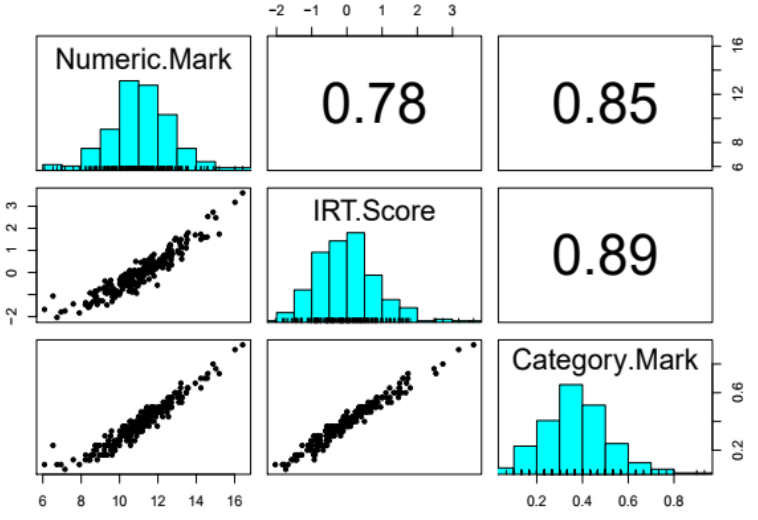

Figure 7. Marginal and pairwise distributions of three ways of measuring the global score of students. Numeric.Mark represents the average of original numeric marks of each student over all the subjects. IRT.Score is the estimated overall ability θ of each student under the GRM. Category.Mark corresponds to the global point average, considering the categorisation (wea=0, ave=0.5, str=1), i.e., what the classical test theory attributes to each student. Correlation is surprisingly high and scatterplots are narrow, suggesting an agreement between the three choices.

The IRT fitted model and classical test theory scores have a high concordance: the Kendall correlation coefficient is 0.89, and the scatterplot is really narrow, showing that, for instance, students with similar global marks correspond to very similar estimated overall ability, all along the range of marks (and conversely, all along the range of abilities). When comparing IRT scores to original numeric marks, correlation is lower and the scatterplot is more spread, but it is natural, since the fit has been conducted with the categorical transformed marks, not with the numeric original ones.

Association and eventual causality among the results of different subjects

As a complement to the findings of the preceding section, the aim would be to relate the students' marks among the different subjects by using the association rules of data mining. It would be interesting to detect rules with a high level of accomplishment (i.e., confidence), but also rules showing premises with significant effect over conclusions. Whether there is causality or just chance is an ontological question that researchers in education must try to discern using epistemological reasoning.

Our target rules are of the type "*students with (strong/average/weak) marks in subject A do generally have (strong/average/weak) marks in subject B*", and also of the kind "*getting (strong/average/weak) marks in subject A improves significantly the chances of getting (strong/average/weak) marks in subject B*". Even if the antecedent raises the chances from "minute" to "small", those pieces of information are very relevant.

According to our sample of moderate size, we have used the classical implication intensity under the Poisson model, as in Khaled et al. (2014). In order to obtain the rules, we have defined binary variables such as Arch.wea, meaning a student gets a weak mark in subject Arch (for Architecture), and so on. The R package *rchic* (Couturier, 2017) was used in order to produce the relations shown in Fig. 8, by choosing a rather conservative threshold level of 0.99 (it is equivalent to a 1% significance level of a test for rejection of independence).

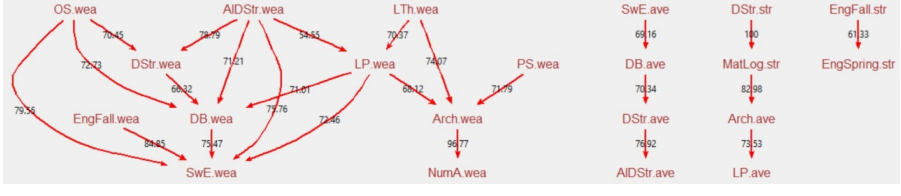

Figure 8. Implicative graph of rules exceeding intensity 99%. Confidence is marked in every arrow. Weak level produces more rules than average or strong level. This information helps specialists to understand relations among success in subjects, and assist them in decision taking.

As an example of interpretation, the displayed rule $EngFall.wea \rightarrow SwE.wea$ suggests some statistically significant effect of "having weak result in EngFall" over "having weak result in SwE", which is very far from being pure chance and that should be analyzed further. The high confidence of the rule can also be easily interpreted (84.85% of students with weak results in EngFall have been found to have also weak results in SwE).

A priori, some competencies can only be gained after some other ones have been seized. The dataset may reflect some of these "expected" ordering relations, or some other, unexpected. The researcher has the role of either finding sense or discarding the unexpected relationship that has arisen from the data analysis.

Almost every significant rule involves subjects at the same level of results (weak, average or strong). It is somewhat expected, as it is related to an existing correlation among marks. Only the rule *MatLog.str* $\rightarrow$ *Arch.ave* describes that having good results in Mathematical Logic raises significantly the chances of having *only* average results in Architecture.

Another interesting remark is that the structure of the rules involving subjects with weak results is far richer than the one corresponding to subjects with average or good results. Threshold can be reduced to level 0.95 in order to get the implicative graph of the 5%-significant rules. On the one hand, it displays many more relations among subjects and levels —which might be of interest to specialists—, but on the other hand, the risk of displaying a rule which is the product of pure chance is higher. The following implications have been found to be relevant to our colleagues:

- $EngFall.str \rightarrow EngSpring.str$: even when the confidence around 60% is not high, the intensity ensures that a strong result in English 1 (Fall) significantly raises the chances of getting strong results in English 2 *(Spring)*.
- $DStr.ave \rightarrow AlDStr.ave$ and $AlDStr.wea \rightarrow DStr.wea$: Algorithm and Data Structures is the continuation of Data Structure. Both rules have a confidence around 75%. For students with average results, the implication follows the ordering among subjects. However, students with weak results produce the rule in the reverse way.
- $SwE.ave \rightarrow DB.ave$ and $DB.wea \rightarrow SwE.wea$: here we have a similar situation among the subjects of DataBases and Software Engineering, which are based on conception — the first one on project conception and the second one on software conception.
- $AlDStr.wea \rightarrow DB.wea$: we can interpret that being weak in Algorithm and Data Structures induces to be weak in DataBases, because this subject involves the use of SQL language, which is very close to the algorithmic of Algorithm and Data Structures.

Conclusion

We have analyzed the marks obtained by 2nd-year students of a bachelor's of Computer Science in order to show a methodology that combines IRT and Association Rules in order to provide with two related but clearly different pieces of information concerning an educational problem. First, the fit of the GRM model of IRT has provided:

- How student marks make up the abstract concept of *overall ability* in this particular discipline, by an acceptable fit to the Grade Response Model of IRT (see Section 4.2).
- How difficult each subject is, as a function of every possible latent overall ability a student may have. This difficulty is not as easy as a single parameter, since there are three levels of result (weak, average and strong), and we can interpret it from the plots of the Item Characteristic Category Curves in Figure 3. In our example, Numerical Analysis and Information Systems are the most difficult ones, but the

difficulty of Software Engineering (with ICCCs shifted to right with respect to most of other subjects) can also be appreciated. On the other hand, English 2 is clearly the easiest subject, followed by English 1, Signal Processing and Language Theory.

- How accurate every subject is at estimating the latent overall ability of the student. In our example, English 1 and English 2 show a low discrimination power, since in a wide range of ability levels two of the three possible outcomes are very likely to happen. This is found on the ICCC plots when curves are very smooth. This phenomenon also occurs, but in a less pronounced form, to Probability and Statistics, Signal Processing and Language Theory. The rest of the modules discriminate much better the latent overall ability of students, as one can see in Figure 3.
- How accurate the complete course - seen as the list of all subjects - is at the task of estimating the latent overall ability of the student (see Figure 5).
- How student results are distributed, according to the latent overall ability, and how it is related to the global point average (see Figure 7). In this case, there is an important agreement in the sense that marks and estimated ability level do correlate very well.

Second, the use of association rules has produced Figure 8, which allows us to say that:

- Most highlighted rules connect subjects mastered at the same level: weak results in some subjects produce an effect of weak results in other subjects, and similarly for average, and for strong results.
- Indeed, there is a more complex structure of rules involving subjects with weak results. This ordering should be considered by teachers and curriculum designers in order to analyze contents relations among the subjects.
- Data Structures is the only subject that produces rules in the three groups of student results: (1) students with strong results have larger chances to have strong results at Mathematics and Logics; (2) students with average results raise their chances to have average results in Algorithms and Data Structures, and (3) students with weak results are more likely to have weak results at Software Engineering.

It is important to mention that a very strict level of 0.99 was used, that could have been reduced to 0.95, producing a larger number of implications. However, by the nature of the implication intensity — as a hypothesis test — researchers should use it, either as a way of testing assumptions made a priori or as a means of finding surprising facts that should be analyzed further, questioned, and tested on a larger sample.

Teachers are strongly advised to analyze their students' marks by these methods in order to determine the difficulty of the subjects and to review their contents if necessary, as well as to improve courses or examination subjects, for example. Furthermore, these two methods can be used to refine the subjects of a specialty, or to assist in choosing among optional training subjects. We claim this methodology can be exported to any type of study degree (college, bachelor's, masters) and domain (Sciences, Engineering, Medicine, Social Sciences, Law, etc.) in order to analyze the relationship among subjects and the knowledge domain, and provide specific valuable information that teachers, curriculum designers and administrators can handle for better assisted decision making.

Address for correspondence

Pablo Gregori
Instituto Universitario de Matemáticas y Aplicaciones de Castellón,
Departamento de Matemáticas,
Universitat Jaume I de Castellón, Campus Riu Sec, E-12071, Spain
Email: gregori@uji.es

Biodata

- The first author: Hayette Khaled is working at the Laboratoire d'Informatique MEDicale (LIMED) as Faculté des Sciences Exactes at Université de Bejaia in Bejaia, Algeria.
- The corresponding author: Pablo Gregori is working at Universitat Jaume I de Castellón, Spain, Instituto Universitario de Matemáticas y Aplicaciones de Castellón, Departamento de Matemáticas.
- The third author: Raphaël Couturier is working at Université Bourgogne Franche-Comté (UBFC), CNRS in Belfort, France.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Proceedings of the 20th International Conference on Very LargeData Bases, VLDB'94* (Vol. 1215, pp. 487-499). San Francisco: Morgan Kaufmann Publishers Inc.

Batanero, C., Navarro-Pelayo, V., & Godino, J. D. (1997). Effect of the implicit combinatorial model on combinatorial reasoning in secondary school pupils. *Educational Studies in Mathematics, 32*(2), 181-199.

Bernard, J. M. (2002). Implicative analysis for multivariate binary data using an imprecise Dirichlet model. *Journal of Statistical Planning and Inference*, *105*(1), 83-103.

Bernard, J. M., & Charron, C. (1996). Bayesian implicative analysis, a method for the study of oriented dependences 1: Binary data. *Mathematics and Social Sciences, 134*, 5-38. [In French]

Bernard, J. M., & Poitrenaud, S. (1999). Multivariate Bayesian implicative analysis of a binary survey: Quasi-implications and simplified Galois lattices. *Mathematics and Social Sciences, 147*, 25-46. [In French]

Bodin, A. (2010). Towards an adaptive-driven test, combining the use of the IRT and the ASI, for the evaluation of the common base of knowledge and skills. *Research in Education (Mathematics), 20*, 383-409. [In French]

Couturier, R. (2017, December). Rchic [Web log post]. Retrieved from http://members.femto-st.fr/raphael-couturier/en/rchic

Couturier, R. (2008). CHIC: Cohesive hierarchical implicative classification. In R. Gras, E. Suzuki, F. Guillet, & F. Spagnolo (Eds.), *Statistical implicative analysis: Theory and applications* (pp. 41-53). Berlin, Heidelberg: Springer.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. London: L. Erlbaum Associates.

Fazio, C., Battaglia, O. R., & Di Paola, B. (2013). Investigating the quality of mental models deployed by undergraduate engineering students in creating explanations: The case of thermally activated phenomena. *Physical Review Special Topics-Physics Education Research*, *9*(020101), 1-21.

Gras, R., & Couturier, R. (2011). Specificities of Statistical Implicative Analysis (ASI) in relation to other quality measures in association rules. *Research in Education-GRIM, 20*(1), 19-57. [In French]

Gras, R., & Kuntz, P. (2006). Discovering R-rules with a directed hierarchy. *Soft Computing*, *10*(5), 453-460.

Gras, R., & Totohasina, A. (1995). Students' conceptions on conditional probability revealed by a data analysis method: Implication- similarity - correlation. *Educational Studies in Mathematics, 28*(4), 337-363. [In French]

Gras, R., Almouloud, S. A., Bailleul, M., Larher, A., Polo, M., Ratsimba-Rajohn, H., & Totohasina, A. (1996). *Statistical implication: A new exploratory data analysis method: Applications to didactics*. Grenoble: Ed. La Pensée Sauvage. [In French]

Gras, R., Almouloud, S., Ratsimba-Rajohn, H., & Couturier, R. (2017, December). Data analysis software C.H.I.C [Web log post]. Retrieved from http://www.ardm.eu/contenu/logiciel-d-analyse-de-donnees-chic [In French]

Gras, R., Kuntz, P., & Briand, H. (2003). Oriented hierarchy of generalised rules in implicative analysis. *Review of Artificial Intelligence, 17*(3), 145-157. [In French]

Gras, R., Régnier, J. C., Marinica, C., & Guillet, F. (2013). *Statistical implicative analysis exploratory and confirmatory method for the research of causality*. Toulouse: Cépaduès Editions. [In French]

Gras, R., Suzuki, E., Guillet, F., & Spagnolo, F. (Eds.). (2008). *Statistical implicative analysis: Theory and applications*. Berlin: Springer.

Hamdare, S. (2014). An adaptive evaluation system to test student caliber using item response theory. *International Journal of Modern Trends in Engineering and Research*, *1*(5), 329-333.

Hedeker, D., Mermelstein, R. J., & Flay, B. R. (2006). Application of item response theory models for intensive longitudinal data. In T. H. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 84-108). New York: Oxford University Press.

Johns, J., Mahadevan, S., & Woolf, B. (2006). Estimating student proficiency using an item response theory model. In M. Ikeda, K. D. Ashley, & T. W. Chan (Eds.), *International conference on intelligent tutoring systems* (pp. 473-480). Heidelberg: Springer.

Khaled, H., Ghanem, S., & Couturier, R. (2014, November). Analysis of Bejaia University Computer Science students' marks through the CHIC software and Statistical Implicative Analysis. In S. Sidhom Chairperson, *2014 4th International Symposium ISKO-Maghreb: Concepts and Tools for knowledge Management* (pp. 1-8). IEEE.

Lerman, I. C., Gras, R., & Rostam, H. (1981a). Elaboration and assessment of an implication

index for binary data. 1. *Mathematics and Social Sciences, 74*, 5-35. [In French]

Lerman, I. C., Gras, R., & Rostam, H. (1981b). Elaboration and assessment of an implication index for binary data. 2. *Mathematics and Social Sciences, 75*, 5-47. [In French]

Loevinger, J. E. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, *61*(4), i.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

O'Neill, G. (2010). Initiating curriculum revision: Exploring the practices of educational developers. *International Journal for Academic Development*, *15*(1), 61-71.

Pantziara, M., Gagatsis, A., & Elia, I. (2009). Using diagrams as tools for the solution of non-routine mathematical problems. *Educational Studies in Mathematics*, *72*(1), 39-60.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/.

Rowan, B., Schilling, S. G., Ball, D. L., Miller, R., Atkins-Burnett, S., & Camburn, E. (2001). Measuring teachers' pedagogical content knowledge in surveys: An exploratory study. *Ann Arbor: Consortium for Policy Research in Education,* PA: University of Pennsylvania, 1-20.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*, 100-114.

Thomas, M. L., Brown, G. G., Thompson, W. K., Voyvodic, J., Greve, D. N., Turner, J. A., & Potkin, S. G. (2013). An application of item response theory to fMRI data: Prospects and pitfalls. *Psychiatry Research: Neuroimaging*, *212*(3), 167-174.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.