UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

OSCAR YAIR ORTEGON

ANALYSIS AND ADAPTATION OF QUESTIONNAIRES BASED ON ITEM RESPONSE THEORY

Thesis presented in partial fulfillment of the requirements for the degree of Master of Computer Science at Programa de Pós-Graduação em Computação, Instituto de Informática Aplicada da Universidade Federal do Rio Grande do Sul.

Advisor: Prof. Dr. Leandro Krug Wives

Porto Alegre
June, 2019

CIP - CATALOGING-IN-PUBLICATION

Ortegon Romero, Oscar
ANALYSIS AND ADAPTATION OF QUESTIONNAIRES BASED ON ITEM RESPONSE THEORY / Oscar Ortegon Romero. -- 2019.
104 f.
Orientador: Leandro Krug Wives.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Informática, Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2019.

1. IRT. 2. Item Response Theory. 3. Online Learning Systems. 4. Questionnaires. I. Krug Wives, Leandro, orient. II. Título.

ACKNOWLEDGMENT

I want to thank YHWH because day by day he has been opening doors that nobody else has been able to.

YHWH is faithful
Deuteronomy 7:9

This work is dedicated to all those who collaborated in the development of this research.

To Professor Leandro Krug Wives among his many contributions for his guidance and patience that allowed me to remember the passion for research. To my colleagues in the laboratory 213, especially to Guilherme and Ana who, in addition to their knowledge, allowed me to share their joys and dreams and can be called them friends.

To my family, especially my mother and my sister, always present in my life, sometimes geographically far away but in my heart forever.

Finally, Elena my wife, because for her today I am here, full and happy, my source of support and joy.

Emma: I hope this effort will serve as an example, my daughter. The time does not matter, the place or your age. Find the way of your dreams, trusting in those who love you and the hand of the almighty.

ABSTRACT

This work presents a model for the design and creation of virtual adaptive evaluations for e-learning environments, combining Item Response Theory (IRT) along with log analysis of previous questionnaires. The proposed model allows the definition of a methodology for the ranking and categorization of questions. Such ranking provides valuable feedback to the teacher or tutor who can refine and adapt the questionnaire. The results of the experiments showed the existence of questions that concentrate the greatest amount of knowledge acquired by the students leaving the other questions with the possibility of being improved to increase the quality of the questionnaire used. We believe that this approach should become an essential tool for the creation of questionnaires that are more concise and effective in the context of virtual courses.

Keywords: IRT, Item Response Theory, Online Learning Systems, Questionnaires.

ANÁLISE E ADAPTAÇÃO DE QUESTIONÁRIOS COM BASE NA TEORIA DA RESPOSTA DO ITEM

RESUMO

Este trabalho apresenta um modelo para o planejamento e a criação de questionários adaptativos utilizados em ambientes virtuais de aprendizagem. O modelo apresentado combina o uso da Teoria de Resposta ao Item (TRI) com a análise histórica de questionários. Com base nisso, propõe-se uma metodologia para a categorização e ranqueamento das questões pertencentes aos questionários. Tal ranking provê um feedback valioso para o professor ou tutor, que pode então refinar e adaptar o questionário. Os resultados dos experimentos evidenciaram a existência de questões que concentram a maior quantidade de conhecimento adquirido pelos alunos deixando as demais questões com a possibilidade de serem aprimoradas para aumentar a qualidade do questionário utilizado. Espera-se que o uso da metodologia auxilie o docente na elaboração de questionários mais concisos e eficazes.

Palavras-chave: TRI; Teoria da Resposta ao Item, Sistemas de aprendizagem online, Questionários.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CAT	Computer Adaptive Test
CTT	Classical Test Theory
DL	Distance Learning
FI	Information Function
ICC	Item Characteristic Curve
IRT	Item Response Theory
LTM	Latent Trait Models under IRT
MIRT	Multidimensional Item Response Theory
Se	Typical Estimation Error
UFRGS	Universidade Federal do Rio Grande do Sul
VLE	Virtual Learning Environments

SUMMARY

1. INTRODUCTION

The rise of Distance Learning (DL) has brought a new paradigm to teaching strategies. The DL field relies on computer-aided tools for the exchange of learning objects, enabling interclass communication, and sometimes even to assess learners' knowledge through online tests. One of the most common tools used to support DL are Virtual Learning Environments (VLE), such as Moodle[1].

One category of VLE that has been proven very effective in supporting learning is adaptive learning systems. An adaptive system, as described by (BRUSILOVSKY, 2001), is one that treats the problem of "one-size-fits-all", when users with different preferences and backgrounds receive the same standardized content. Learning environments are one of the most successful applications of adaptive strategies (BRUSILOVSKY, 2001). One of the reasons is because the user educational profile, such as her previous knowledge should be taken into consideration when presenting new content.

Computer Adaptive Tests (CAT) are one important kind of examination used inside an adaptive learning environment, and they are used to deal with the multitude of learners with different backgrounds that exist in an online course. In CAT an algorithm manages the test presentation, the questions selection, and also decides dynamically when the test should be finished. In the end, the test verifies the student's answers and estimate each student level of knowledge (CHALHOUB;DEVILLE, 1999)

Item Response Theory (IRT) is one of the methodologies that can be used to implement a CAT and to analyze the Bank of questions. IRT is popular in the educational field because it has been successfully used in qualitative processes of psychological and educational evaluation. It is used to measure and evaluate students' acquired knowledge and the development of necessary skills in some subject (VENDRAMINI, 2002). IRT is a framework for modeling student responses on a set of assessments. It is used to describe the relationship between the proficiency of a student and the likelihood of correctly answering a test item. There are two assumptions IRT models make about the tests to be adapted. First, namely, the unidimensional criteria where each question of the test measures one, and only one, skill; second, the local independence criteria, where each question of the test is independent of the other, i.e., any answer provided in another question should not influence the answer of a question.

Besides these assumptions, IRT models can differ from one another by the number of parameters they use to describe each question in a test. In this sense, there are three possible parameters to be implemented in an IRT model, and they are the following: Item discrimination index or "a", which indicates how each learner, with different abilities, differ from others relating the probability of choosing an answer over another; item difficulty index or "b" that considers the learner skill scale and verifies the required skill for choosing a correct answer;

[1] http://moodle.org/

and random chance of success or “c”, which corresponds to the probability of giving a correct answer to a question even though not fulfilling such questions knowledge requirements.

According to Pasquali (2003), the models can be generated with one, two or three parameters. The model with one parameter or Rasch model is one that analyzes the item difficulty “b” and the discrimination Index with a constant value (a = 1). The logistic model with two parameters or (2PL), analyses the discrimination index “a” and the item difficulty “b”. Finally, the logistic model that uses the three parameters (3PL) considers the discrimination index “a”, the item difficulty “b” and the random chance of success “c”.

Based on these parameters using statistical and mathematical tools, IRT seeks to find a theoretical description to explain the behavior of empirical data generated from the application of the psychometric instrument over the questionnaires. Such theoretical description helps in evaluating the technical quality of each question and also estimates the level of knowledge each student has on a specific topic. Analyses presented in the work of Araujo and Bortoli (2003) conclude that some advantages of using IRT are i) the generation of precise metrics to assess the questions and the user level, and ii) the generation of dynamic questionnaires that adapt themselves to the student level of knowledge.

In this context, this work structures and presents a methodology to apply IRT over a set of non-adaptive questionnaires. The main goal of such methodology is by using previous answers given to one questionnaire, perform a selection of the most important questions of such questionnaires. Such most important questions (optimal model) are then ranked in order of the ones provoking less error by the students. In this sense, we believe that delivering such ranking of questions to a teacher can help in the improvement of the other questions. Such refinement of questions is a way of adapting the questionnaires to the learners’ knowledge and decreasing the mean error rate.

In this sense, the following research hypotheses are presented.

1.1 Hypothesis

- H0: Questions applied by teachers to students do not have differences in terms of contribution to a questionnaire. The variability lack of questions in the questionnaires is not a problem for the students.

- H1: The item response theory can identify variability in the amount of information from the questions applied to the students, this variability is a problem for the students.

In order to validate the above hypotheses, the following questions are formulated.

1.2 Research Questions

- **Q1**: Are all the questions necessary for a questionnaire? Are some of them more important than others?
- **Q2**: Can it be established a ranking of importance (or contribution) of the questions of a questionnaire?
- **Q3**: Can the position of the question in the ranking indicate badly formulated questions?
- **Q4**: Some questions can be easier or more difficult for one group of students?
- **Q5**: Can the concentration made by IRT analysis be evaluated for identifying if a question is badly formulated, i.e., if a great population of students failed, or if it is a difficult question, i.e., if some students hit and other students fail.

1.3 Document Structure

This work including the Introduction is structured in eight chapters. Chapter 2 presents the general concepts that support this work, beginning with the Classical Test Theory explaining the main limitations generating the Item response theory such an alternative to resolving these limitations. The principal mathematical models are explained in this two theories and some concepts that allow generating a general idea that importance of the test for the evaluates process of the distance learning; chapter 3 presents investigations proposed in the literature with a different approach that combines the item response theory for evaluating the learning process. Then chapter 4 proposes the methodology for the realization of this approach, in chapter 5 the methodology is tested presenting two experiments and explain them in detail. The analysis and discussion of the results of the experiments are described in chapter 6, and finally, chapter 8 describes the conclusions and future work proposed for this work.

2. BACKGROUND

As the learning process evolved and as a consequence of the success achieved by tests in the evaluation area, there was introduced a need to develop a theoretical framework to allow the validation of the interpretations and inferences made from tests and allow estimation of measurement errors inherent in any process of this type. This general theoretical framework called Classical Test Theory allowed establishing a functional relationship between the observable variables based on the empirical scores obtained by the subjects in the tests or in the elements that compose them and the unobservable variables.

In this context, Item Response Theory (IRT) was born as an alternative solution to the problems generated by the relationship between the results obtained by the subject and the error resulting from the measurement process (HERNANDEZ;HAMBLETON, 1992). It should be understood that the study proposed defines the item as each question that belongs to a test and that its objective is to measure a single skill that according to Dreyfus et al. (1980) A skill is the ability to carry out a task with determined results often within a given amount of time, energy, or both.

To better understand the why of IRT, it is necessary to know beforehand the Classical Test Theory, in this sense, in this chapter, we address this theory together with the concepts of error, observed error, and real error.

2.1 Classical Test Theory

The Classical Test Theory (CTT) (SPEARMAN, 1904; NOVICK, 1966), also known as Classical Theory of Testing, is defined around three basic concepts:

- empirical score (X);
- true scores (V);
- scores due to error (e).

The central objective was to find a statistical model that adequately supported the test scores found and allowed the estimation of measurement errors associated with any measurement process.

In this sense, Spearman’s linear model (SPEARMAN, 1904) is an additive model in which the observed (dependent variable) score of a subject in a test (X) is the result of the sum of two components, i.e., the true score (independent variable) in the test (V) and the error (e). The score is then given by the following equation:

$$\mathbf{X = V + e}$$

According to Fernandez et al. (1992), from this model, CTT will develop a whole set of deductions aimed at estimating the amount of error that affects test scores. To work, the following assumptions must be taken into account:

- The score (V) is the mathematical expectation of the empirical score (X): V = E (X);
- The correlation between the true "n" scores in a test and the measurement errors is equal to zero, i.e., R+V+E = 0;

With the previous assumptions of the CTT model, the following deductions are established:

- The measurement error (e) is the difference between the empirical (X) and the true (V) measurement, i.e., e = X-V;
- The mathematical expectation of measurement errors is 0 (zero), then they are unbiased, i.e., E (e) = 0;
- The mean of the empirical score is equal to the average of the true ones;
- True scores would not cope with mistakes, i.e., Cov (V, e) = 0;
- The covariance between empirical and true scores is equal to the variance of true ones: cov (X, V) = S2 (V);
- The covariance between the empirical scores of two tests is equal to the covariance between the true ones: cov (Xj, Xk) = cov (Vj, Vk);
- The variance of the empirical score is equal to the variance of the true plus the errors: S2 (X) = S2 (V) + S2 (e);
- The correlation between the empirical score and the error is equal to the quotient between the standard deviation of the errors and that of the empirical ones, i.e., rxe = Se/S.

Among the main limitations of CTT we have, according to Fernandez et al. (1992), that the characteristics of the test and the scores of the people can not be separated; in addition, the score of a person is defined as the number of questions that are correct and the difficulty of an item as the proportion of people who answer it correctly in a certain group.

This has a series of negative consequences:

- The characteristics of the items depend on the group of people in whom they have been applied;
- The score of a person depends on the particular set of items administered;
- The score that a person obtains will be different if we apply two tests that measure the same characteristic but whose level of difficulty is different.

This makes it very difficult to compare these scores, which can only be interpreted in relation to the test in which they were obtained. It can be stated that CTT presents two fundamental problems in test analysis, one related to the sources of error and how they should be operated relative to their different sources, and another more specific one regarding the invariance of measurements and properties of the measuring instruments. That is, if two different tests measure the same variable for different individuals, we cannot know which of those variables is better, because the results are not on the same scale.

The solution adopted was to express the scores relatively in terms of a normative or standard group, acceptable solution but, that can not follow the rigor of a scientific measurement which makes it difficult to verify that this measurement can be reliable if they are in function of the instrument used as presents (LORD, 1953) in his considerations on the nature of the metric provided by scores on a mental test. That is if we consider for example that the length of an object depends on the type of rule used to measure it.

The requirement to give an adequate solution to all the problems generated grew as the use of tests became widespread. The two problems connected with invariance will find an adequate solution within the framework of Item Response theory.

2.2 Item Response Theory

According to Hambleton (1991). IRT is a methodology that estimates the ability (s) of an individual in an area of knowledge and the characteristics of the items considered relevant for an evaluation, that is, that may interfere with the response given by a particular examinee to an item.

In this context, skill is a latent variable, that is, a variable that cannot be measured directly, differently from variables such as weight, height, temperature, etc. Therefore, variables such as anxiety, satisfaction, intelligence, knowledge, which are not directly measured, are classified as latent; this type of variable is measured from observable secondary variables related to it, in the case of competence, the secondary variable observed is the given by the respondent to an item. IRT proposes models for latent variables and is currently applied in several areas such as education, psychiatry, psychology, and several others. IRT has as its basic unit the item, that is, each question of a test, the test being then a set of items.

2.3 Test dimensionality

Considering the set of all abilities that affect the response of the examinee(s) in at least one item of a *J* item test, we will have the vector:

$$\boldsymbol{\emptyset = (\emptyset_1,\emptyset_2,\emptyset_3,...,\emptyset_m)^t}$$

In his work, Lord (1970) calls this “complete latent space vector”, where the ability of each examinee is a point of this latent space. The size of the test will then be m, ie the full latent space dimension. A test is called one-dimensional if m = 1, so there is a unique ability affecting the respondent's response. In this case:

$$\mathbf{Ø = (Ø_{1})}$$

2.4 Item parameters

In addition to the ability according to Baker (2001), IRT also estimates the parameters of items that are characteristic of the item, such parameters are:

1) $b_{(j)}$ is the difficulty parameter of item j.
2) $a_{(j)}$ is the discrimination parameter of item j.
3) $c_{(j)}$ is the parameter of the random hit probability of item j.

Associated to each item j, we will have a parameter vector of this form:

$$\mathbf{ð_{(1)}= (a_{(j)},b_{(j)},c_{(j)})}$$

Thus, it will be the set of all parameter vectors of the J items, that is:

$$▲ = \mathbf{(ð_{(1)},ð_{(2)},ð_{(3)}......ð_{(j)})}$$

The parameters of the items are invariant in a population. It means that, no matter what the average skill of the group, the parameters of the estimated items will be the same, that is, they are independent of the ability.

Due to the invariance of the parameters of the items, and their independence with IRT, it can be possible to:

(a) Compare a single group with a single test;

(b) Compare a single group, divided into two subgroups, making two completely different (no common item);

(c) To compare a single group, divided into two subgroups, making two tests, partially different;

(d) Compare two groups with a single test;

(e) Compare two groups making two tests, partially different;

(f) monitor the progress of an examinee or group of examined over time;

(g) Evaluate correctly the items. For example, the difficulty parameter of an item estimated by IRT will always be the same regardless of the ability of the group of examinees. Unlike CTT, where the difficulty of an item depends on the skill of the group of examined. If a group of examiners, with high ability, respond to an item, the estimated difficulty may be low while that if the same item is answered by another group, with skill lower, the estimated difficulty may be greater.

(h) Estimate the ability regardless of the parameters of the items. While in CTT the examiner's score depends and varies according to the difficulty of the test (easier or more difficult) in IRT, the ability is always the same regardless of the difficulty of the test.

2.5 Answers classification of examination to an item

The response given by an examiner to an item can be classified as:

a) Dichotomous: In this case, the answers are classified only in certain or wrong;

b) Polynomial: In addition to the right or wrong classifications, probabilities for other response categories;

c) Continuous: Values are used within a range of numbers to sort the response. Used in open questions (BRAGION, 2010).

2.6 IRT Models

There are several IRT models, and those models depend on three factors according to De Andrade et al. (2000):

i- of the nature of the item: dichotomic or non-dichotomous (polynomial or continuous);

ii- the number of populations (or groups) involved: only one or more of one;

iii- the number of skills being measured: only one or more of one.

In this work, only one-dimensional, dichotomous items, for a single population will be considered. One hardly has only one skill being measured in one item, so it is often admitted that a test is a one-dimensional if there is a dominant ability affecting the respondent's response. That
Consideration should be given to the population since a test can be one-dimensional for one population and multidimensional for another.

Another important concept is that of one or more populations (or groups) involved, this concept must take into account if the characteristics of the population are different with respect to the ability to be estimated. For example, if the skill considered is creativity, students in grades 5 and 6 may be taken as a single population. If the researched skill is grammar, it can be considered that these same two populations are distinct. In this work, a group will be considered a sample obtained through the simple random sampling process.

2.7 One-dimensional model of dichotomous items for a single population

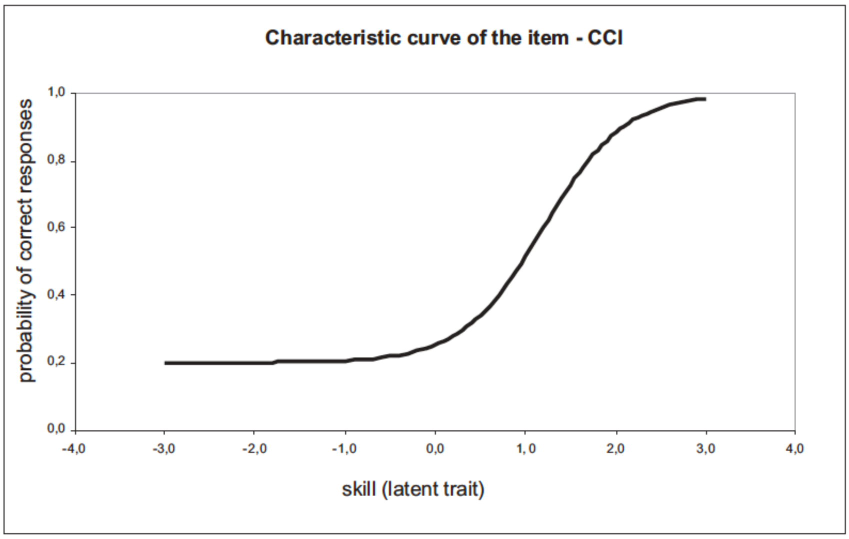
According to De Andrade et al. (2000), the one-dimensional model of dichotomous items for a single population has the following assumptions:

a) Unidimensionality: the assumption that the latent space of the item is one-dimensional, that is, given by the examinee depends on a single competence and the characteristics of the item considered by the model. Generally, the dimensionality of the test is verified through factorial analysis.

b) Local Independence: assumes that for a given ability, the responses to the different items of a test are independent. This means that the response of a certain item, in a test, can not influence the response of other items.

Figure 1 describes the Item Characteristic Curve that indicates the likelihood that people will have to face an item. This probability depends on the level of the person in the measured variable.

Figure 1: Example of item characteristic curve



Source: Pasquali et al. (2000)

Figure 1 shows that as the individual's ability increases, in like manner increases the probability of hitting the item.

2.7.1 Standard two parameter warhead model

Being $\mathbf{Y_{ij}}$ the random variable associated with the hit of individual i on item j with $i = 1, \ldots, n$ (examined), $j = 1, \ldots , J$ (items).

In dichotomous items, the respondent's answer is classified as right or wrong, the value 1 will be associated with the correct answers and the value 0, with wrong answers. The respondent's response to item j is conditionally related to his competence and the parameters of the item, that is, as the competence increases, we expect to have an increase in the probability of that individual to hit the item, in this way:

a) $\mathbf{Y_{ij}} = 1$ For correct answers.
b) $\mathbf{Y_{ij}} = 0$ For wrong answers.

Thus, the variable $\mathbf{Y_{ij}} \approx$*Bernoulli(*π_{ij}), being your probability of distribution given by:

$$\mathbf{f(Y_{ij}= y_{ij} \mid \varnothing = \varnothing_i ;\eth_j)} = \boldsymbol{\pi ij^{yij}(1 - \pi ij)^{1-yij}}$$

i=1,..., n (examined) , j=1, … , J (items)

Being:

a) $\boldsymbol{\pi ij}$ The probability of an individual hit correctly the item j.
b) $\mathbf{\varnothing}$ **Represents** of the random variable ability of individual *i* with skill correctly respond to item j.
c) $\varnothing_{\mathbf{i}}$ Is the individual's ability.
d) $\mathbf{b_j}$ The difficulty parameter of item j.
e) $\mathbf{\eth_j}$ Parameter vector of item j.

If $\mathbf{Y_{ij}} = 1$ we have that:

$$\mathbf{f(Y_{ij}= 1 \mid \varnothing} = \varnothing_i \mathbf{;\eth_j)} = \boldsymbol{\pi ij^{1}(1 - \pi ij)^{1-1} = \pi ij}$$

If $\mathbf{Y_{ij}} = 0$ we have that:

$$\mathbf{f(Y_{ij}= 0 \mid \varnothing} = \varnothing_i \mathbf{;\eth_j)} = \boldsymbol{\pi ij^{0}(1 - \pi ij)^{1-\,0} = 1 - \pi ij}$$

By definition:

$$\boldsymbol{\pi ij}\mathbf{= f(Y_{ij}= 1 \mid \varnothing} = \varnothing_i \mathbf{;\eth_j)} = \mathrm{P}\mathbf{(Y_{ij}= 1 \mid \varnothing} = \varnothing_i \mathbf{;\eth_j)}$$

According to Lord (1968), was the first to develop a model using the normal warhead for more than two items. This model is given by:

$$\boldsymbol{\pi ij}\mathbf{= f(Y_{ij}= 1 \mid \varnothing} = \varnothing_i \mathbf{;\eth_j)} = \mathrm{P}\mathbf{(Y_{ij}= 1 \mid \varnothing} = \varnothing_i \mathbf{;\eth_j)} = \int_{-\infty}^{\boldsymbol{aj(\varnothing i - bj)}} \boldsymbol{\frac{1}{\sqrt{2\pi}} e^{\,.\,t^2/2}\, dt}$$

2.7.2 Normal warhead and logistic models

Birnbaum (1968) used a logistic model to replace the model with a normal warhead. Such substitution was made because of the simplicity of the logistic model by having an explicit integral. The following are the functions of the logistic model and the normal accumulated.

$$\varnothing(\boldsymbol{x}) = \int_{-\infty}^{\infty} \frac{\mathbf{1}}{\sqrt{\mathbf{2}\pi}} \boldsymbol{e}^{\boldsymbol{t}^{\mathbf{2}}/\mathbf{2}} \text{ dt } \textbf{Model with a normal warhead}$$

$$\phi(\boldsymbol{x}) = \boldsymbol{e}^{\boldsymbol{x}} \ / \ \mathbf{1} + \boldsymbol{e}^{\boldsymbol{x}} = (\boldsymbol{e}^{\boldsymbol{x}}/\boldsymbol{e}^{\boldsymbol{x}}) \ / \ \frac{\mathbf{1} + \boldsymbol{e}^{\boldsymbol{x}}}{\boldsymbol{e}^{\boldsymbol{x}}} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-\boldsymbol{x}}} \ \textbf{Normal logistic model}$$

Birnbaum (1968) states that the logistic model can be used to replace function $\phi(\boldsymbol{x})$because it almost coincides with the normal warhead, being defined as:

$$|\ \varnothing(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x})\ [(\mathbf{1},\mathbf{702})\boldsymbol{x}]| \ < \mathbf{0},\mathbf{01};\ -\infty \ < \ \boldsymbol{x} \ < \ +\infty$$

In this case, 1.702 will be represented by D and called the scale, that is, the value that best approximates the graph of the distribution function accumulated logistics of the normal warhead.

In his work, Birnbaum (1968) mentions Halley's work (1952), where there is the demonstration that, when the scale parameter is 1,702, we have the best approximation of the logistic distribution function with relation to the normal warhead.

2.7.3 Logistic model of 1, 2 and 3 parameters

Item 2.4 previously defined that IRT also estimates the parameters of items that are characteristic of the item, such parameters are:

1) b(j) is the difficulty parameter of item j.
2) a(j) is the discrimination parameter of item j.
3) c(j) is the parameter of the random hit probability of item j.

Associated to each item j, we will have a parameter vector of this form:

ð(1)= (a(j),b(j),c(j))

According to De Andrade et al. (2000), the principal models that using these parameters are described as follows (see subsections).

2.7.3.1 Rasch model (one parameter)

This model analyzes that the probability of hitting an item depends only on the level of difficulty of said item and the level of the subject in the measured variable (level of ability). The mathematical expression is:

$$\mathbf{P(Y_{ij}= 1 \mid Ø = Ø_i ;ð_j) =} \frac{\mathbf{1}}{\mathbf{1+}e^{-D(Øi\ -ðj)}}$$

i=1,..., n (examined) , j=1, … , J (items) , $ð_j = (1,b_j,0)$, | **($-\infty<Ø_i<+\infty$)** and **($-\infty<b_j<+\infty$)**

Being:

f) $\mathbf{Y_{ij}}$ random variable associated with the correctness or error of the individual i to item j, can assume the values 0 or 1.

g) $\mathbf{P(Y_{ij}= 1 \mid Ø = Ø_i ;ð_j)}$ is the probability of individual i with skill correctly respond to item j.

h) $\mathbf{Ø_i}$, the parameter of the individual's ability i

i) $\mathbf{b_j}$, the difficulty parameter of item j.

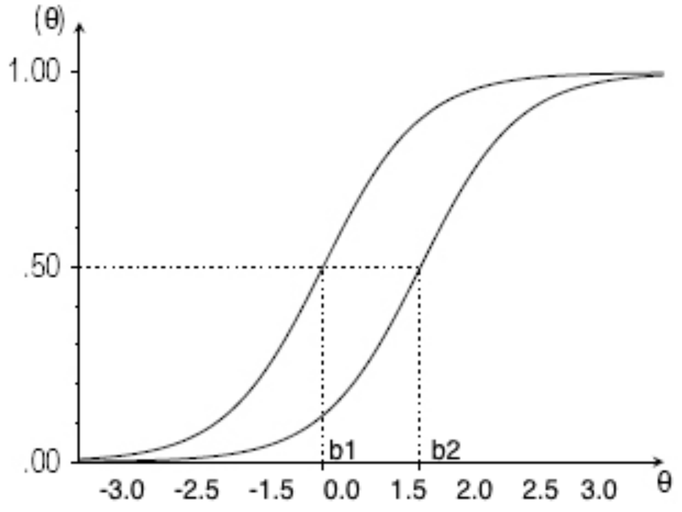
j) **D** 1,702 scale parameter

k) $\mathbf{ð_j}$ parameter vector of item j.

In this model, the only parameter that varies is parameter b, namely the “difficulty parameter”. Such a parameter must be on the same scale of the skill of the student and can assume any value **($-\infty<b<+\infty$).**

In Figure 2, two Item Characteristics Curves (ICC) are presented, the first value (Ø on the left) that corresponds to P(Ø) = 0.5 is about -0.75. Therefore, the difficulty of the first item is b1 = -0.75. The second item, i.e., the (Ø) value that corresponds to P(Ø) = 0.5 is about 1. Therefore the difficulty of the second item is b2 = 1.

Figure 2: Parameters of discrimination (a) and difficulty (b)



Source: Pasquali et al. (2000)

This figure shows that the probability of hitting the item is systematically less in item 2 then item 1. Thus, item 2 is more difficult than item 1 and its difficult indexes are probing that (b2 > b1).

2.7.3.2 Logistic model of two parameters

The mathematical expression for this model is given by:

$$\mathbf{P(Y_{ij}= 1 \mid \text{Ø} = \text{Ø}_i\ ;\text{ð}_j) = \frac{1}{1+e^{D(-a(\text{Ø}i\ -\text{ð}j))}}}$$

i=1 n (examined) , j=1,....J(items) , $\text{ð}_j = (1,b_j,0)$, $\text{ð}_j = (a_j,b_j,0)$,

$(-\infty<\text{Ø}_i<+\infty)$, $(-\infty<b_j<+\infty)$ and $(a>0)$

Being:

a) $\mathbf{Y_{ij}}$ random variable associated with the correctness or error of the individual i to item j, can assume the values 0 or 1.
b) $\mathbf{P(Y_{ij}= 1 \mid \text{Ø} = \text{Ø}_i\ ;\text{ð}_j)}$ is the probability of individual i with skill correctly respond to item j.
c) $\mathbf{\text{Ø}_i}$, the parameter of the individual's ability i.
d) $\mathbf{\text{ð}_j}$ parameter vector of item j.
e) $\mathbf{b_j}$ the difficulty parameter of item j.
f) $\mathbf{a_j}$ discrimination parameter of item j.
g) **D** 1,702 scale parameter.

It is observed that this model differs from the Rasch model, because it is considered the parameter of discrimination of the item, between more the index of discrimination grows, the greater the inclination of its characteristic curve. This behavior can be identified in Figure 2 that shows the ICC of two items of equal difficulty (b1=b2=0.75). The main difference between them is that item 2 in **(Ø)**=0.75 has an inclination greater than item 1, since a2=2.4 and a1=0.4. As the slope is so high, individuals with Ø > 0.75 have almost all a high probability of settling item 2 and individuals with Ø < 0.75 have almost all a near-zero probability of hitting the item. Therefore, item 2 discriminates between those who haveØ > 0.75 and those who have Ø < 0.75. For its part, item 1 has very little inclination when Ø = 0.75. Consequently, although most individuals with Ø > 0.75 will hit, many will fail. Also, although most individuals with Ø < 0.75 will fail the item, many will hit, because the probability of hitting is greater than zero. In item 1 the probability grows very slightly as Ø grows so it is not a good discriminator between individuals with Ø > 0.75 and individuals with Ø < 0.75.

2.7.3.3 Logistic Model of three parameters

In this model, besides the parameters described previously, it is also used a parameter to represent the casual hit of the item by the examined of low ability, denoted by Birnbaum (1968), who introduced this parameter to the model, considering the fact that students with low ability, sometimes give correct answers to the items.

The mathematical expression for this model is given by:

$$\mathbf{P(Y_{ij}= 1 \mid \text{Ø} = \text{Ø}_i ;ð_j) = c_i + (1\text{-}c_j) \left(\frac{1}{1+e^{D(-a(\text{Ø}i - ðj))}}\right)}$$

$i=1\ n$ (examined) , $j=1,....J$(items) , $ð_j = (1,b_j,0)$, $ð_j = (a_j,b_j,0)$,

$(-\infty<\text{Ø}_i<+\infty)$ and **$(-\infty<b_j<+\infty)$** , $(a>0)$ and $(0>c_j>1)$

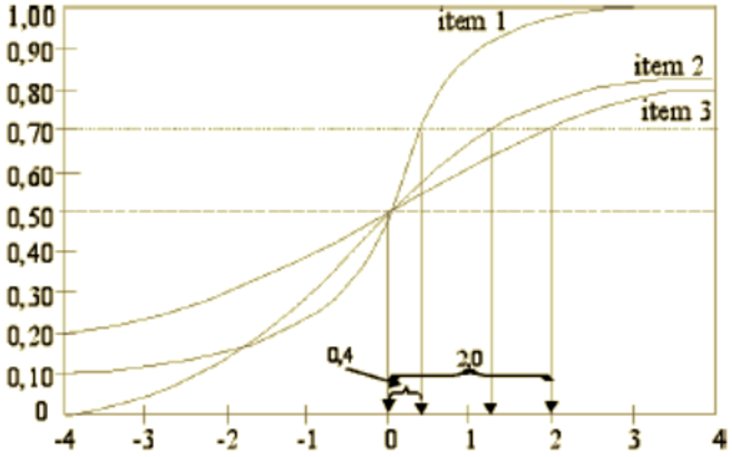
Being:

a) **Y_{ij}** random variable associated with the correctness or error of the individual i to item j, can assume the values 0 or 1.
b) **$P(Y_{ij}= 1 \mid \text{Ø} = \text{Ø}_i ;ð_j)$** is the probability of individual i with skill correctly respond to item j.
c) **Ø_i** , the parameter of the individual's ability i.
d) **$ð_j$** parameter vector of item j.
e) **b_j** , the difficulty parameter of item j.
f) **a_j** discrimination parameter of item j.
g) **c_j** the parameter of the probability of an accidental hit.
h) **D** 1,702 scale parameter.

This model can also be interpreted as: "The probability of the examinee *i* kicking and guessing the probability of the examined *i* did not kick and hit".

In Figure 3, the ICC of several items with different parameter "c" is shown. It has been assumed that the probability of chance matching is not simply the quotient where $1/q$, where q is the number of alternatives of the item, this probability depends on how the item is drawn. Thus, items with the same number of alternatives will have different values of c_j.

Figure 3: Parameter of discrimination of three items.



Source: Pasquali et al. (2000)

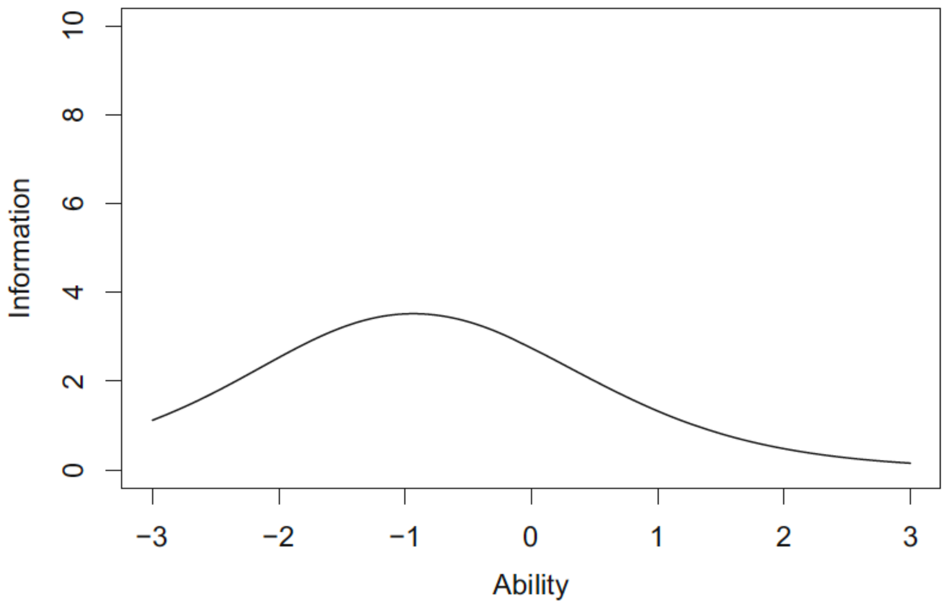
2.8 Information Function

According to Baker et al., (2017, the term information and his statistical meaning were defined as the reciprocal of the variance with which a parameter could be estimated. […] Statistically, the magnitude of precision with which a parameter is estimated is inversely related to the size of the variability of the estimates around the value of the parameter. The variance of the estimator is denoted by $ð^2$. The amount of information, denoted by *I*, then is given by the formula:

$$I = \frac{1}{ð^2} \qquad \text{(Eq. A)}$$

IRT estimates the value of the ability parameter for an examinee. From Eq. (A), the amount of information at a given ability level is the reciprocal of this variance. If the amount of information is large, it means that an examinee whose true ability is at that level can be estimated with precision; that is, all the estimates will be reasonably close to the true value. If the amount of information is small, it means that the ability cannot be estimated with precision and the estimates will be widely scattered about the true ability.

Figure 4: An information Function Calculated for all levels



Source: (BAKER et al., 2017)

An information function level on the ability scale from negative infinity to positive infinity. Because ability is a continuous variable, the information will also be a continuous variable. If the amount of information is plotted against ability, the result is a graph of the information function such as that shown in Figure 4.

Thus, the information function tells us how well each ability level is being estimated. The information function does not depend upon the distribution of examinees over the ability scale. In this regard, it is like the item characteristic curve and the test characteristic curve.

In a general-purpose test, the ideal information function would be a horizontal line at some large value of *I* and all ability levels would be estimated with the same precision. The typical information function looks somewhat like that shown in Fig. 4 and different ability levels are estimated with differing degrees of precision. (Baker et al., 2017)

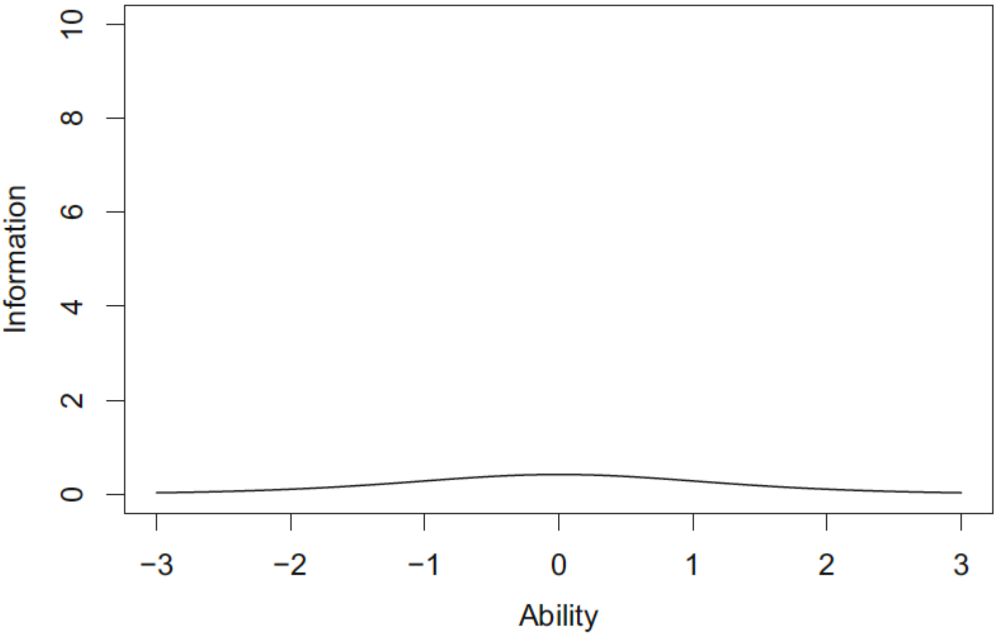
2.8.1 Item Information Function

According to Baker et al. (2017), IRT is an itemized theory because each item of the test measures the underlying latent trait. If analyzed one single item the amount of information.

In this sense, the information function has great importance in the use of the tests since it allows us to choose the one that contributes more information in the range of θ that we are interested in measuring. It is also very useful in building the test. From a bank of calibrated items (that is, from which we have estimated its parameters) we can select those that allow an Information function to fit certain objectives.

Specifically, the Information Function of an item is denoted by I_j(Ø) where j indexes the item. Figure 5 describes the amount of information against ability.

Figure 5: An item information Function



Source: (Baker et al., 2017)

According to Baker et al. (2017), “an item measures ability with the greatest precision at the ability level corresponding to the item’s difficulty parameter. The amount of item information decreases as the ability level departs from the item difficulty and approaches zero at the extremes of the ability scale”. All previous concepts can be referred to the test information function because are also applicable to each of the items separately. In fact, the test information function is no more than the sum of the ICC of each of the items that compose it.

2.9 Parameter estimation

Once an IRT model is selected, it is necessary to apply the test to a large sample, to estimate the parameters of each item and the ability of each subject, based on the obtained response matrix. The estimation of the parameters is the step that allows us to arrive from the known responses of the people to the items, the unknown values of the parameters of the items and the trait levels.

In order to obtain the estimates, the maximum-likelihood method is applied (DE ANDRADE, 2000). In this method, the parameters are estimated in two steps:

1. The parameters of the items are estimated assuming the abilities, in this step the abilities initially supposed, are the notes of the standardized examinations.
2. Skills are estimated by assuming the items, the values assumed for the parameters of the items are those found in step 1.

This two-stage process is repeated until the convergence of skills and parameters of the items. The general logic of estimation is to find the values of the parameters that make the response matrix obtained more likely. If we flip a coin 10 times and get seven faces, the maximum-likelihood estimator of parameter "p" (coin face probability) is 7/10 = 0.7, as shown in traditional Statistics books - see Amon (1984).

The result "seven faces in ten launches" is little compatible with the face probability being 0.1, or 0.2, ... In fact, the probability of obtaining seven faces and three crosses is practically zero if p = 0.1 or if p = 0.2. This probability becomes 0.117 if p = 0.5, and reaches the maximum value (0.267) when p = 0.7. The maximum-likelihood estimator provides the value of "p" under which the event we have found is most likely to occur.

In IRT, the estimation procedure follows a similar logic. We obtain the estimates of the parameters and the levels of θ with which the data matrix found have the maximum compatibility.

In general, a person will respond to a number of items greater than two and will produce a particular sequence of ones and zeros. The probability of obtaining such a sequence of hits and errors can be written as:

$$\mathbf{QP = L\ \Pi\ R\ R\text{-}1}$$

Where R is the result in each item (1, success, 0, failure), P is the probability of success in each item, and Q is the probability of error in each item (Q = 1-P).

Parameter θ is estimated by the maximum-likelihood method, and it will be the value of θ for which the previous expression reaches its maximum value. When we try to estimate the trait level in a real situation, we do not make a search restricted to a few values, we need to find the value of θ that maximizes L among the possible values.

In the case of IRT, there are no formulas that allow estimates to be obtained directly. In the example of the coins, it is known that the maximum-likelihood estimator of the population proportion is the sample proportion. In the IRT, in the absence of such formulas, the estimates are obtained by numerical methods, using computer programs. In the most general case, a function L is established that depends on the parameters of the items and the levels of traits. Computer programs contain algorithms that find the set of estimates for which the function L reaches the maximum value. Parameters of items and trait levels of people will be the values given by the computer program for a particular response matrix.

In Classical Test Theory, once items are applied to a set of people, the score of each person in the test can be obtained by combining the scores in the test items. In IRT, once the items have been applied, the response matrix containing the successes and failures of each person in each test item is generated. Next, a computer program has to be applied which will give us the trait levels and parameters of the items. As we have seen, because these are estimates by the method of maximum-likelihood, the values given by the program are the ones that make the original data matrix more plausible, they are the most compatible with the original data matrix.

2.10 Applications

IRT has enabled the development of computerized adaptive tests (CATs), according to Renom (1993). Such tests differ substantially from tests to use. A CAT consists of a well-calibrated item bank and a computer program in charge of deciding which bank item to present to the person, to present it, to analyze the response issued by the person, to choose a new bank item, and so on. A CAT differs greatly from a pencil and paper test. A first difference is that it is managed by a computer and a second is that each person is evaluated with different items.

However, the fundamental thing about CATs is that items are chosen with the criterion of estimating the level of skill of the person with the highest precision and the lowest number of items.

In brief, a CAT proceeds as follows:

A. Presentation of the first item.
B. Estimation of the level of trait of the person.
C. Search for the most informative bank item for the level of θ estimated in the previous step.
D. Application of the chosen item.
E. Estimation of the level of trait corresponding to the sequence of responses given to the items presented.
F. Again step "c", and so on until a typical estimation error less than a preset stop has been achieved or a predetermined number of items has been administered.

The main achievement of CATs, according to Ponsoda et al. (1994), is that with very few items (twenty, more or less) we can achieve comparable or better measurement accuracies than those obtained in much longer non-adaptive tests. This is because CATs are only administered authentically informative items to determine the level of trait of the person and avoid too easy or difficult items, which barely report on the level of trait.

The relevance given by teachers to the tests used for the evaluation is still very high; a factor that can affect the quality of the evaluation is evident in the tendency they have to use the tests that they prepared much more than any other type of test (DARLING; HAMMOND, 2000). That leads us to think about the additional knowledge that teachers must have to create a test that manages to adequately measure one or several levels of knowledge acquired by the student.

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST)[2] have Standards that evaluate areas such as cognitive complexity, content quality, meaningfulness, language appropriateness, transfer and generalizability, fairness, and reliability.

The aforementioned aspects identify the degree of complexity level for the qualification of learning in the student with additional multidisciplinary aspects at the level of knowledge that the teacher should have in order to create adequate tests that measure the learning on the subject being taught and the meaning such theories as IRT can help in a more adequate evaluation.

In the next chapter, it will be presented a general overview with works that use IRT as an alternative to assess learning in online learning environments and that were taken into account as a conceptual basis for the formation of the methodology proposed in this study.

[2] Is This a Trick Question? A short guide to writing effective questions. Available at http://ksde.org/Portals/0 /CSAS/CSAS%20Home/CTE%20Home/Instructor_Resources/TrickQuestion.pdf

3. RELATED WORK

As already stated, Item Response Theory (IRT) is a framework for modeling student responses on a set of assessments (tests). It is used to describe the relationship between the proficiency of a student and the likelihood of correctly answering a test item. There are two assumptions IRT models make about the tests to be adapted. First, namely, the unidimensional criteria where each question of the test measures one, and only one, skill; second, the local independence criteria, where each question of the test is independent of the other, i.e., any answer provided in another question should not influence the answer of a question.

In the literature, we can find studies combining item analysis methodologies with IRT. One of such works is the one of Santos et al. (2005), which presents a computational tool for the elaboration of adaptive evaluation using as a conceptual base IRT and Computerized Adaptive Tests. In that work, a methodology is proposed for the calculation of the level of difficulty and the ability of the student, using means and medians for the scores obtained from the answers to the questions, the expected values are calculated, and experiments are elaborated with two evaluation models. For the realization of the experiments, all students start the test with a question of the same level of difficulty, which for the case is intermediary. Two evaluation models are presented. The first, namely evaluation model I, alternates the questions of agreements with the answers obtained. If the student is correct, the next question will be of a more challenging level; otherwise, a question will be asked with a level below the level of the question that is being asked at the moment. The evaluation model II starts with a question of the same level of difficulty then the next question is calculated with the proposed methodology by the authors, and it self-adjusts gradually according to the student's answers.

The authors assume that the more students answer a question, the easier it is and the level of difficulty should decrease. In the opposite case, the question is considered the more difficult level and its level should increase. These results are the basis for creating the profile of a group of students or a particular student. According to the analysis of the results, evaluation model II is reached more quickly to the calculation of the student's ability but in cases in which the students are classified as Advanced or Basic. Additionally, the first three questions were not used in the analyzes because they state that the students are in the process of adaptation and learning of the evaluation method.

The results of the investigation show that model II is more favorable when compared with the model I. Authors also postulate that it would be interesting to reformulate the experiments using a random level question at the beginning of the tests. It would ratify results or show if the polarization of the difficulty levels of the basic or advanced students the result of is always choosing a question of intermediate level. Additionally, an additional factor that can influence the level of difficulty of the tests and the number of questions used can be added.

According to Cook (2014), an adequate number of questions in a questionnaire is ten questions, this affirmation was the result of experiments with students who answered questionnaires of random sizes varying among 1, 5, 10 or 15 questions. The analysis showed the test had a higher number of questions presented a lower performance in the student than the students who answered a smaller number of questions and the best average result was for those who answered questionnaires of 10 questions. It is worth noting that the scale used is five units but what can be evidenced is that if we have a test with several questions, we can have better performance of students if we reduce less than 35% of the number of questions in the questionnaire.

Tian et al. (2017) present a study of the reading development level of Chinese students. In this case, IRT parameters are applied to find a relationship between their values and a method to modify item's options that do not have a reasonable behavior to the data. However, the study focuses on the modification of the items that could generate an addressing in the learning methodology and lose the test's ability to identify interest points for the teacher, who could make essential findings of the students' answers.

A multidimensional IRT temporal model named T-BMIRT is proposed by Huang-Hu (2017), and it is compared with traditional IRT in online learning studies. The study raises the importance that students, during different moments of time, may have different levels of knowledge. Because of that, the parameters must be estimated for each of those moments as part of the Temporal IRT, describing the student learning trajectories in an online education system.

The results obtained are better than the ones obtained by traditional IRT, and temporal IRT performs better when the dataset contains learning videos interactions. However, the set of learning objects cannot be limited to just one type in online environments, and it would be necessary to test the approach with other types of objects to find similarities in the results. Additionally, the proposed model does not include information from the students during their interaction with videos in the dataset used for the estimation of the parameters. Authors propose to perform a more extensive analysis of the data using different moments in which the students recorded their answers, using the analysis criteria of the extended IRT model.

The application of IRT in online environments still needs to be studied and disseminated in the research environment, as indicated by Jatobá et al. (2017), and this field generates a large study space to use techniques that implement IRT. In this case, results showed that the use of online environments based on CAT and IRT is still quite limited, which motivates us to go deeper into the subject and try to contribute to the research process since the methodology proposed generates input for the creation of a CAT with VLE's question banks.

The authors and their aforementioned works leave open the need to know if it is possible to identify which questions are more important within an evaluation process and how we could link the evaluated contents with the questions of the questionnaires. The proposed methodology aims to initially address this classification process so that in a next stage the resulting information can be used as input into the creation of more accurate adaptive tests.

4. METHODOLOGY

The methodology presented in this section has the goal of providing a helpful tool for teachers. This tool can pick and rank the most important questions in a questionnaire. Such selection is made by using IRT algorithms and a set of strategies, defined in this work, to select and rank the questions. By providing the teacher with this set of the most important questions, it is expected to help in the improvement of their questionnaires, enhancing the students' performance in the tests. Since question selection is made by following IRT parameters and not only by text analysis along with answers statistics, this method can guarantee the selection of questions that really cover essential parts of the knowledge, and also can contribute for decreasing the errors.

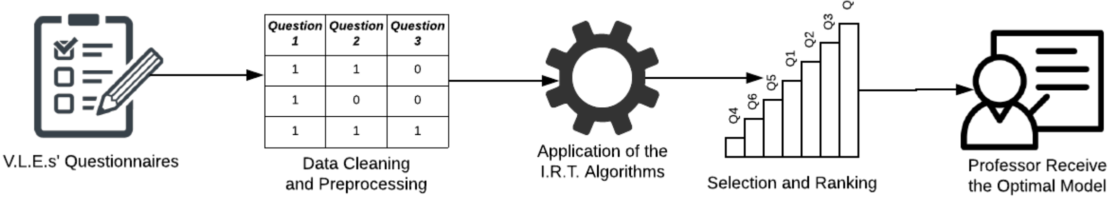
Due to the characteristics and procedures that were used for the analysis of the data, it is aligned with the conditions proposed by the experimental methodology, which according to Fonseca et al. (2002), It enables an approach and an understanding of reality to investigate as an unfinished permanent process and is the result of a thorough examination carried out with the objective of solving a problem, using scientific procedures, investigating a trained person or group addressing an aspect of reality in the sense of experimentally testing hypotheses.

In the same way Gil et al. (2007), experimental research consists in determining an object of study, selecting the variables that would be able to influence it, defining the forms of control and observation of the effects that the variable reduces on the object.

So after presenting the problem that wants to be addressed and the hypotheses formulated to solve it, we described the methodology for ranking questions used in virtual learning environments.

In Figure 6, it is provided an overview of the proposed methodology for questions analysis, selection, and ranking. Each one of the phases is better explained in the next subsections.

Figure 6: Overview of the proposed process



Question 1	Question 2	Question 3
1	1	0
1	0	0
1	1	1

V.L.E.s' Questionnaires
Data Cleaning and Preprocessing
Application of the I.R.T. Algorithms
Q4
Q6
Q5
Q1
Q2
Q3
Q7
Selection and Ranking
Professor Receive the Optimal Model

Source: the author.

4.1. Data Cleaning and Preprocessing

The first step of the methodology is data acquisition and preparation. Even though in this methodology it is assumed the dataset will come from the questionnaires of existing VLE such as Moodle, it is necessary to analyze the answers provided by the students as well as the questions' content before the IRT algorithms could be applied.

The first constraint for the collected data is that it should be aggregated into questionnaires. To such aggregation, it is necessary to guarantee that the contents and the order of questions do not change when applied to different sets of students (of different classes, for instance). Each questionnaire now can be organized into a matrix where each line represents a student, and each column represents a question. Matrix cells are filled with the students' grades to each question. If the constraint is guaranteed, the data can be analyzed and cleaned (BONG NA et al., 1999).

Sometimes VLEs (KOSKELA et al., 2005) allow a student to answer the same questionnaire more than once. In such situations, only the first attempt is considered valid because once the student is familiar with the questions, such previous knowledge can insert a bias over the question's importance analysis. The attempts not finalized neither submitted for evaluation are also discarded.

Once it is guaranteed that questionnaires aggregate the data, and it is cleaned, it should be binarized (CORBETT; ANDERSON, 1994). Such a step is essential because the analysis is made relying on the information of the student success or failure while answering each question. So, it is assumed if a student has reached a score of, at least, 75% of the total score[3], her grade will be replaced by one (1), representing success in answering such question. Otherwise, it will be replaced by zero (0). Such binarized matrices will be the input for the IRT algorithms, which will perform an analysis of importance for each question based on the success or failure each student has in answering the questions.

4.2 Application of Item Response Theory

The questionnaires are then submitted for analysis by using MIRT, an R library that analyzes dichotomous data (RIZOPOULOS, 2006) and computes the IRT parameters, including difficulty, discrimination, guessing, and the maximum amount of information. The process involves two phases:

a. Calculation of the Parameter Logistic Models (Rasch, 2PL, 3PL): *Using the binary* matrix of each questionnaire as input, three parameter logistic models are calculated.

[3] In this case are used this percentage (75%) because are the university utilize this measure like evaluating criteria.

b. Creation of optimal model: Once the best model is known, it is executed a statistical analysis is made over the results. Such analysis output is the optimal model, i.e., a subset of the initial questionnaire formed by the questions that maximize the coverage of the knowledge topics presented in the questionnaire.

4.3 Selection and Ranking

The methodology final step refers to the selection of the final subset of more important questions. Since IRT returns an optimal model for each questionnaire, and each questionnaire could be applied to different sets (scholar classes) of students, the output of IRT needs to be processed before handle the final set of more important questions to the professor.

According to De Andrade et al. (2000), an item presents a greater amount of information when it has a high discriminative index and a low success rate. In the analysis of the Item Characteristic and Item Information Function curves applied, it is possible to construct a classification in relation to the amount of information and discrimination. highlighting the following elements:

1. Good information, good discrimination and a reasonable chance of success.
2. Lots of information, high discrimination and low probability of chance.
3. Low information and low discrimination.
4. Out of trial standards by IRT.

Under these conditions, items that meet conditions 1 and 2 were considered. For the other items (3 and 4), there is a need to reevaluate their elaborations, with a view to correcting problems such as cohesion, clarity of skill required, correction of alternatives or even layout and layout of the item in the test. Ultimately the item is discarded as it does not meet the required criteria. Finally, the ranking was created giving a priority to the questions that had a higher difficulty level.

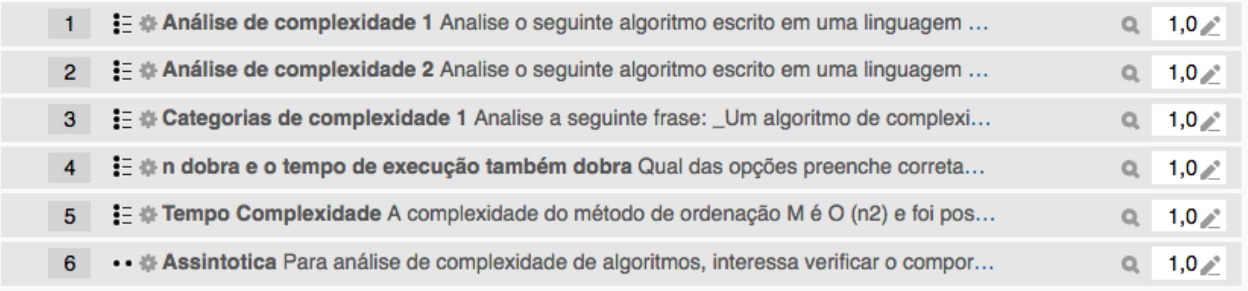
5. EXPERIMENTS

To validate the proposed methodology, two experiments were performed; The first one considers a course on “Data Classification and Searching”, offered at the Institute of Informatics at UFRGS, considering the periods 2016-1, 2016-2, and 2017-2. From this course were obtained the answers of 2 questionnaires: Algorithms Complexity and Hashing (showed in Figure 7 and Figure 8), which were applied during the course in a different order. The second experiment was performed for the course of Electrical Engineering, also at UFRGS, during the periods 2016-2, 2017-1, 2017-2, and 2018-1, the answers of 12 questionnaires were obtained.

5.1 Experiment 1: “Data Classification and Searching” course

This experiment considered a dataset collected from a course named “Data Classification and Searching” at UFRGS. Such course applied many questionnaires using Moodle, but two of them were used in this experiment; they were related to (a) “Hashing” and (b) “Algorithms Complexity” subjects. The questionnaires were answered by students over the periods of 2016-1 with 31 Students, 2016-2 with 44 students, and 2017-2 with 25 students. The “Hashing” questionnaire is composed of 11 questions, while the “Algorithms Complexity” is composed of 6 questions (Figure 7).

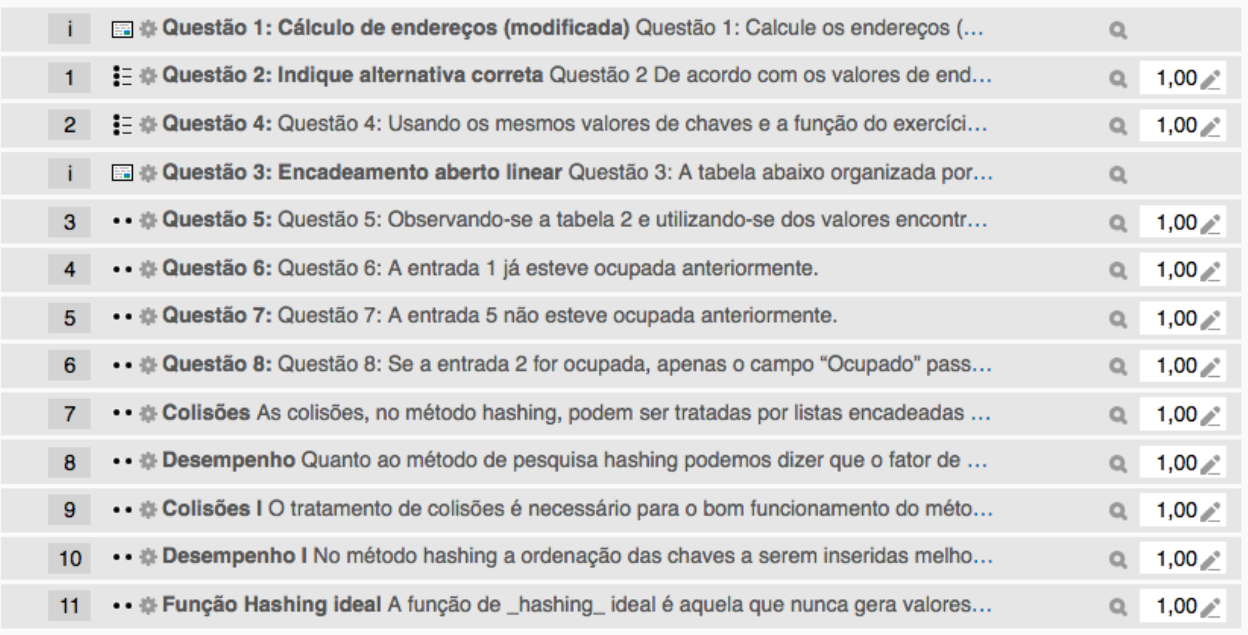
Figure 7: Questions of the Algorithms Complexity Questionnaire



1	**Análise de complexidade 1** Analise o seguinte algoritmo escrito em uma linguagem …	1,0
2	**Análise de complexidade 2** Analise o seguinte algoritmo escrito em uma linguagem …	1,0
3	**Categorias de complexidade 1** Analise a seguinte frase: _Um algoritmo de complexi…	1,0
4	**n dobra e o tempo de execução também dobra** Qual das opções preenche correta…	1,0
5	**Tempo Complexidade** A complexidade do método de ordenação M é O (n2) e foi pos…	1,0
6	**Assintotica** Para análise de complexidade de algoritmos, interessa verificar o compor…	1,0

Source: the author.

Figure 8: Questions of the Hashing Questionnaire



i	**Questão 1: Cálculo de endereços (modificada)** Questão 1: Calcule os endereços (…	
1	**Questão 2: Indique alternativa correta** Questão 2 De acordo com os valores de end…	1,00
2	**Questão 4:** Questão 4: Usando os mesmos valores de chaves e a função do exercíci…	1,00
i	**Questão 3: Encadeamento aberto linear** Questão 3: A tabela abaixo organizada por…	
3	**Questão 5:** Questão 5: Observando-se a tabela 2 e utilizando-se dos valores encontr…	1,00
4	**Questão 6:** Questão 6: A entrada 1 já esteve ocupada anteriormente.	1,00
5	**Questão 7:** Questão 7: A entrada 5 não esteve ocupada anteriormente.	1,00
6	**Questão 8:** Questão 8: Se a entrada 2 for ocupada, apenas o campo "Ocupado" pass…	1,00
7	**Colisões** As colisões, no método hashing, podem ser tratadas por listas encadeadas …	1,00
8	**Desempenho** Quanto ao método de pesquisa hashing podemos dizer que o fator de …	1,00
9	**Colisões I** O tratamento de colisões é necessário para o bom funcionamento do méto…	1,00
10	**Desempenho I** No método hashing a ordenação das chaves a serem inseridas melho…	1,00
11	**Função Hashing ideal** A função de _hashing_ ideal é aquela que nunca gera valores…	1,00

Source: the author.

Each questionnaire was extracted from Moodle in a CSV file. The "Algorithms Complexity", for instance, presented six questions in a score scale varying from 0.0 to 1.7. While the "Hashing" questionnaire presented 11 questions varying from 0.0 to 0.91. In Table 1, it is shown a sample of the data collected from Moodle during the 2017-2 period, and it was composed of the following items:

1. **ID**: Auto Numerical value assigned for the student.
2. **Status**: Information value for questions completed answered.
3. **Date**: Date value when the student responds to the questionnaire.
4. **Used Time**: Time elapsed during the questionnaire.
5. **Grade**: Total score for the questionnaire.
6. **Q1 … Q6**: *score* individual for each question (*score* = Number of questions/10).

Table 1. A sample of "Algorithms Complexity" CSV file for the 2017-2 period

ID	Status	Date	Used Time	Grade	Q1	Q2	Q3	Q4	Q5	Q6
1	Finished	28-3	9m 15s	10.0	1.7	1.7	1.7	1.7	1.7	1.7
3	Finished	4-4	56s	10.0	1.7	1.7	1.7	1.7	1.7	1.7
5	Finished	9-3	57s	10.0	1.7	1.7	1.7	1.7	1.7	1.7
8	Finished	12-4	4m 47s	8.3	1.7	1.7	1.7	1.7	1.7	0
12	Finished	6-4	1m 17s	10.0	1.7	1.7	1.7	1.7	1.7	1.7
14	Finished	11-4	1m 16s	8.3	1.7	1.7	1.7	1.7	1.7	0

Source: the author.

It can be seen that no information concerning the students involved in the process was used (i.e., they cannot be identified). Finally, each period has generated two CSV files, one for each questionnaire, totaling six files were used with the methodology proposed.

5.1.1 Methodology Application

In this section, we explain the steps applied, i.e., data cleaning and preprocessing, application of IRT, and the selection and ranking of questions.

5.1.1.2 Methodology (Step 1): Data Cleaning and Preprocessing

The questionnaires were cleaned to eliminate attempts that were not finalized and multiple attempts. The performed changes are:

Questionnaire 1: Hashing

(a) From 46 records originally presented in the 2016-1 period, 15 records were excluded.
(b) From 40 records originally presented in 2016-2 period, 14 records were excluded.

Questionnaire 2: Algorithms Complexity There were no changes.

During cleaning, only 72% of the 111 records collected from Moodle were kept, which represents 81 attempts. Once cleaned, it is necessary to binarize CSV files, considering the following criteria: if the student has reached at least 70% of the grade of a question, the grade is then replaced by 1, otherwise by 0.

5.1.1.3 Methodology (Step 2): Application of Item Response Theory

From the graphic analysis of the calculated logistic models described in Figure 9 and Figure 10, were identified the questions with the highest index of discrimination in questionnaires Hashing and Complexity respectively; According to Pasquali et al. (2003) “ [...] if an item presents a perfect discrimination, then the angle of incidence of the curve would be 90 degrees, that is, a perpendicular. In this case, the item is able to discriminate infinitesimally minimal differences in the levels of theta [...]” in addition was used the value of the discrimination coefficient for each logistic model (2PL - 3PL) and its angle of incidence, were made the selections of questions in each period analyzed. Then, were searched matches between the most repeated questions, with the aforementioned characteristics. a criterion of descending importance is established from higher to lower among the indices of discrimination.

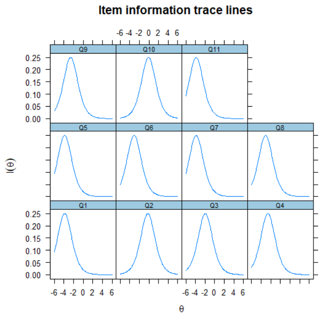
The graphics of the rasch model were not used because the discrimination index is constant by definition (a = 1).

Table 2 presents the coefficients between periods 2016-1,2016-2 and 2017-2, the highlighted values represent the highest value for each logistic model. The most difficult questions were: Q3 with a difficulty level (b) = -1.944, the question Q2 (b) = -0.49 and the question Q2(b)=-1.91 for Rasch model; the questions Q9 (b) = 10.041, Q9(b) = 1.433 and the question Q6 = -1.008 for the two parameter model (2PL); Finally the questions Q10 (b)= 327, Q9 (b)= 322 and Q7=0.19 for the three parameter (3PL) model; similarly the Questions more discriminative were Q5 with a parameter (a) = 33.224, Q6 (a) = 39.36, Q3,Q5 (a) = 58.262 for the two parameter model (2PL); the questions Q4 (a) = 57.209, Q10 (a) = 896, Q2 (a) = 754 for the three parameter (3PL) model; the period with more variability in the success of hit was 2016-2 with values in the “c” parameter, only the question Q6 was excluded.

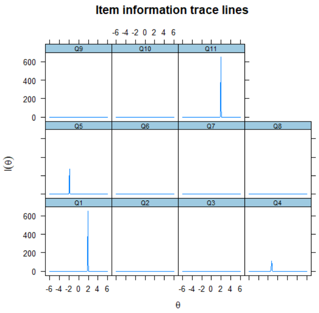
Questionnaire 1: Hashing

Figure 9: IRT Analysis for Hashing Questionnaire

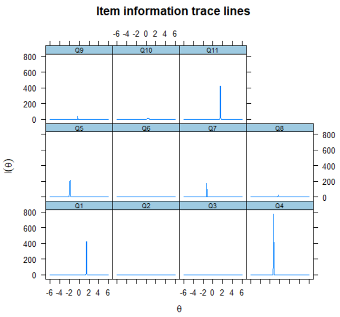
Period:2016-1



Rasch Model

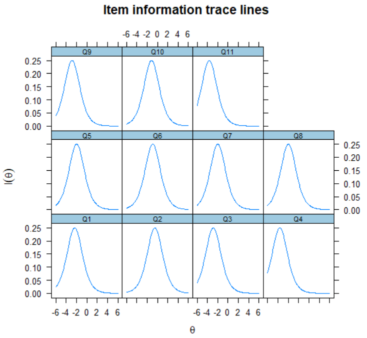


Logistic Model of two parameters

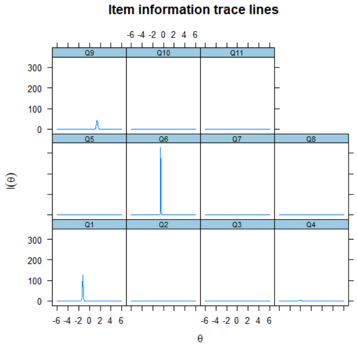


Logistic Model of three parameters

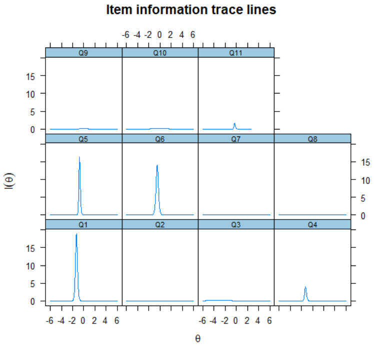
Period:2016-2



Rasch Model

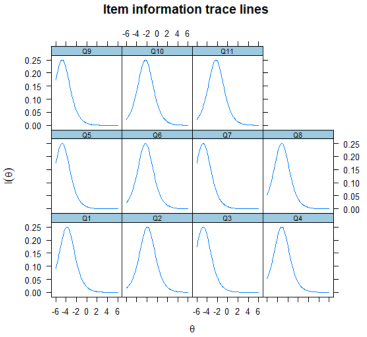


Logistic Model of two parameters

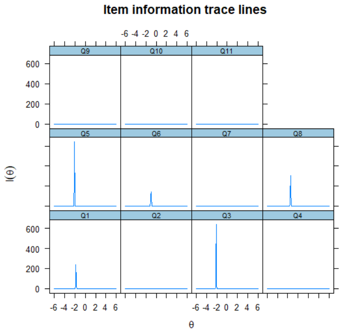


Logistic Model of three parameters

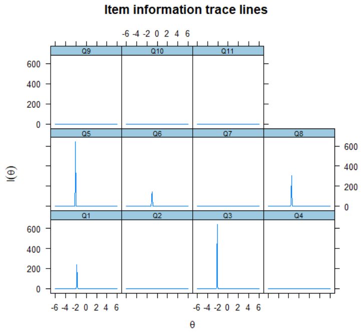
Period:2017-2



Rasch Model



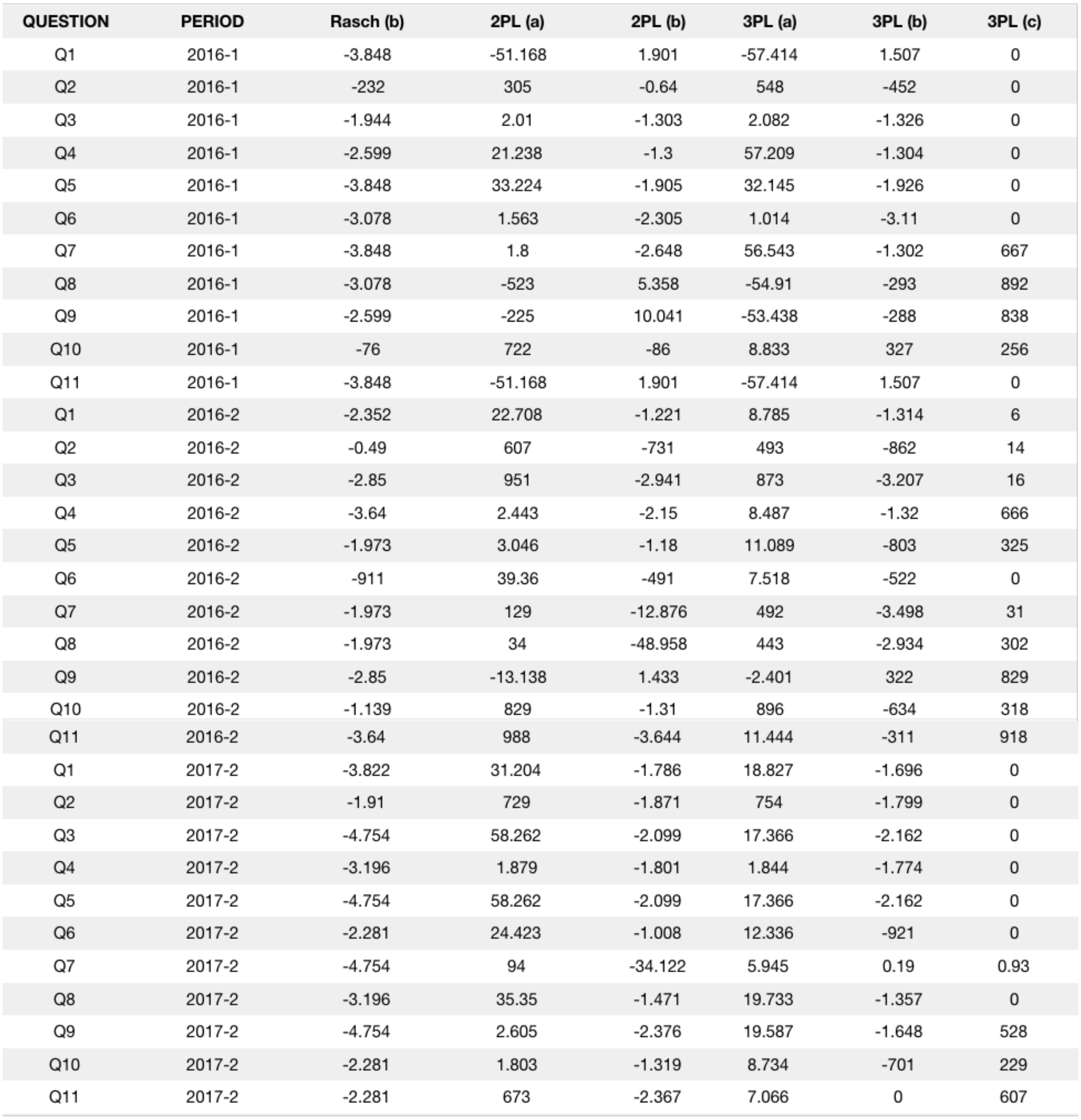
Logistic Model of two parameters



Logistic Model of three parameters

Source: the author.

Table 2: Coefficients of Logistic Models - Hashing - Questionnaire



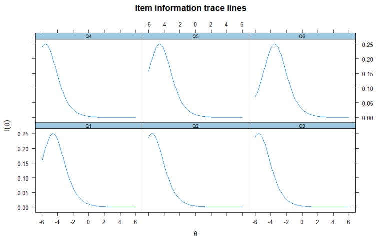
QUESTION	PERIOD	Rasch (b)	2PL (a)	2PL (b)	3PL (a)	3PL (b)	3PL (c)
Q1	2016-1	-3.848	-51.168	1.901	-57.414	1.507	0
Q2	2016-1	-232	305	-0.64	548	-452	0
Q3	2016-1	-1.944	2.01	-1.303	2.082	-1.326	0
Q4	2016-1	-2.599	21.238	-1.3	57.209	-1.304	0
Q5	2016-1	-3.848	33.224	-1.905	32.145	-1.926	0
Q6	2016-1	-3.078	1.563	-2.305	1.014	-3.11	0
Q7	2016-1	-3.848	1.8	-2.648	56.543	-1.302	667
Q8	2016-1	-3.078	-523	5.358	-54.91	-293	892
Q9	2016-1	-2.599	-225	10.041	-53.438	-288	838
Q10	2016-1	-76	722	-86	8.833	327	256
Q11	2016-1	-3.848	-51.168	1.901	-57.414	1.507	0
Q1	2016-2	-2.352	22.708	-1.221	8.785	-1.314	6
Q2	2016-2	-0.49	607	-731	493	-862	14
Q3	2016-2	-2.85	951	-2.941	873	-3.207	16
Q4	2016-2	-3.64	2.443	-2.15	8.487	-1.32	666
Q5	2016-2	-1.973	3.046	-1.18	11.089	-803	325
Q6	2016-2	-911	39.36	-491	7.518	-522	0
Q7	2016-2	-1.973	129	-12.876	492	-3.498	31
Q8	2016-2	-1.973	34	-48.958	443	-2.934	302
Q9	2016-2	-2.85	-13.138	1.433	-2.401	322	829
Q10	2016-2	-1.139	829	-1.31	896	-634	318
Q11	2016-2	-3.64	988	-3.644	11.444	-311	918
Q1	2017-2	-3.822	31.204	-1.786	18.827	-1.696	0
Q2	2017-2	-1.91	729	-1.871	754	-1.799	0
Q3	2017-2	-4.754	58.262	-2.099	17.366	-2.162	0
Q4	2017-2	-3.196	1.879	-1.801	1.844	-1.774	0
Q5	2017-2	-4.754	58.262	-2.099	17.366	-2.162	0
Q6	2017-2	-2.281	24.423	-1.008	12.336	-921	0
Q7	2017-2	-4.754	94	-34.122	5.945	0.19	0.93
Q8	2017-2	-3.196	35.35	-1.471	19.733	-1.357	0
Q9	2017-2	-4.754	2.605	-2.376	19.587	-1.648	528
Q10	2017-2	-2.281	1.803	-1.319	8.734	-701	229
Q11	2017-2	-2.281	673	-2.367	7.066	0	607

Source: the author.

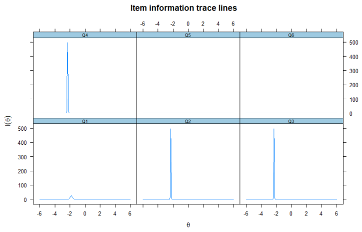
Questionnaire 2: Algorithms Complexity

Figure 10: IRT Analysis for Algorithms Complexity Questionnaire

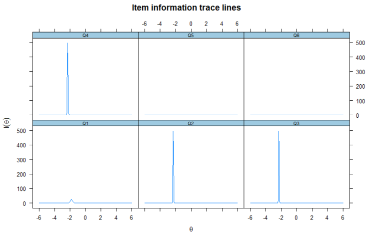
Period:2016-1



Rasch Model

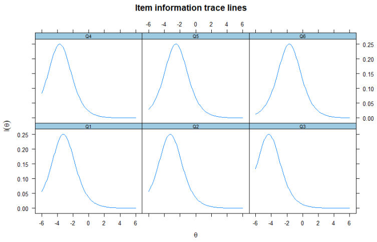


Logistic Model of two parameters

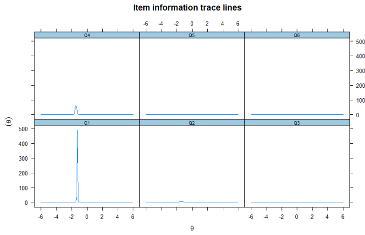


Logistic Model of three parameters

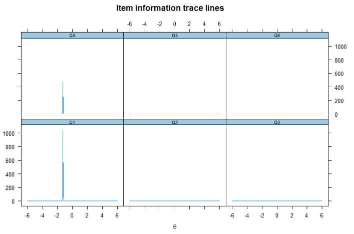
Period:2016-2



Rasch Model

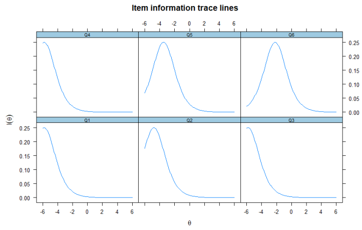


Logistic Model of two parameters

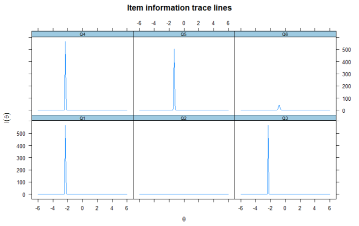


Logistic Model of three parameters

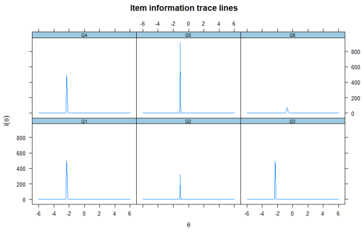
Period:2017-2



Rasch Model



Logistic Model of two parameters

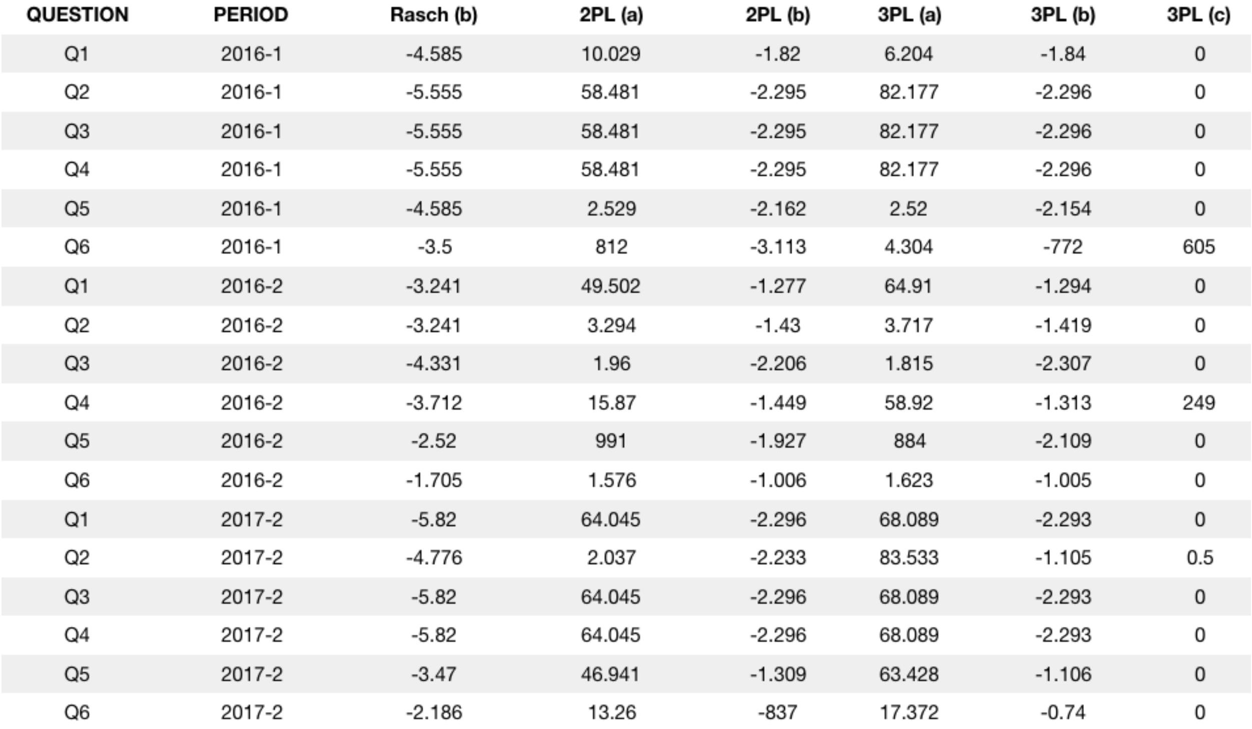


Logistic Model of three parameters

Source: the author.

Table 3 presents the coefficients between periods 2016-1,2016-2 and 2017-2, the highlighted values represent the highest value for each logistic model. The most difficult questions was: Q6 with a difficulty level (b) = -3.5, Q6 (b) = -1.705 and Q6 (b) = -2.186 for Rasch model; the questions Q1 (b) = 1.82, Q6 = -1006 and Q5 = -1.309 for the two parameter model (2PL); Finally the questions Q1 (b) = -1.84, Q6 = -1.005 and Q6 = -0.74 for the three parameter model; similarly the questions more discriminative were Q2,Q3 and Q4 with a parameter (a) = 58.481, Q1 (a) = 49.502 and Q1,Q2 and Q4 (a) = 64.045 for the three parameter model (3PL); The random success of hit was founded in the questions Q6 with a parameter (c) = 685, Q4 (c) = 249 and Q2 (c) = 0.5.

Table 3: Coefficients of Logistic Models - Algorithm Complexity - Questionnaire



QUESTION	PERIOD	Rasch (b)	2PL (a)	2PL (b)	3PL (a)	3PL (b)	3PL (c)
Q1	2016-1	-4.585	10.029	-1.82	6.204	-1.84	0
Q2	2016-1	-5.555	58.481	-2.295	82.177	-2.296	0
Q3	2016-1	-5.555	58.481	-2.295	82.177	-2.296	0
Q4	2016-1	-5.555	58.481	-2.295	82.177	-2.296	0
Q5	2016-1	-4.585	2.529	-2.162	2.52	-2.154	0
Q6	2016-1	-3.5	812	-3.113	4.304	-772	605
Q1	2016-2	-3.241	49.502	-1.277	64.91	-1.294	0
Q2	2016-2	-3.241	3.294	-1.43	3.717	-1.419	0
Q3	2016-2	-4.331	1.96	-2.206	1.815	-2.307	0
Q4	2016-2	-3.712	15.87	-1.449	58.92	-1.313	249
Q5	2016-2	-2.52	991	-1.927	884	-2.109	0
Q6	2016-2	-1.705	1.576	-1.006	1.623	-1.005	0
Q1	2017-2	-5.82	64.045	-2.296	68.089	-2.293	0
Q2	2017-2	-4.776	2.037	-2.233	83.533	-1.105	0.5
Q3	2017-2	-5.82	64.045	-2.296	68.089	-2.293	0
Q4	2017-2	-5.82	64.045	-2.296	68.089	-2.293	0
Q5	2017-2	-3.47	46.941	-1.309	63.428	-1.106	0
Q6	2017-2	-2.186	13.26	-837	17.372	-0.74	0

Source: the author.

5.1.1.4 Methodology (Step 3): Selection and Ranking

Finally, criteria 1 and 2 previously described in section 4.3 were applied to Tables 4 and 5, and the discrimination index for each question are analyzed, as described in the following paragraphs.

Hashing Questionnaire:

By criterium 1: Good information, good discrimination and a reasonable chance of success
Questions selected: Q3, Q4

By criterium 2: Lots of information, high discrimination and low probability of success
Questions selected: Q3, Q5, Q6

Final selection (criteria 1 and 2): **Q3, Q4, Q5, Q6**

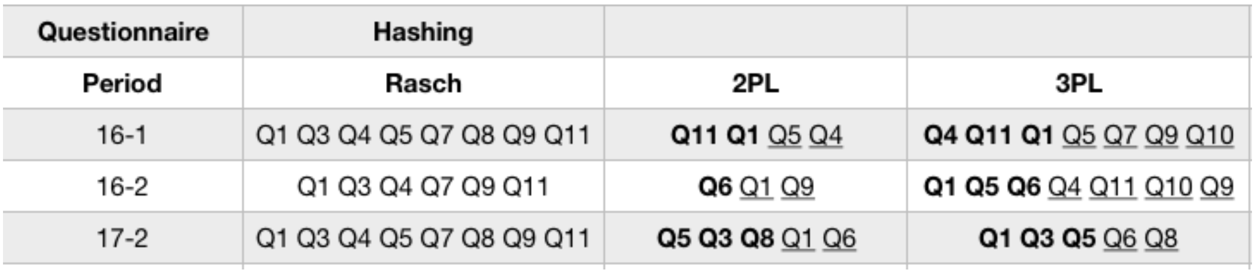
Algorithm Complexity Questionnaire:

By criterium 1: Good information, good discrimination and a reasonable chance of success
Questions selected: Q1, Q3, Q4

By criterium 2: Lots of information, high discrimination and low probability of success
Questions selected: Q1, Q2, Q3, Q4

Final selection (criteria 1 and 2): **Q1, Q3**
In this case, questions Q2 and Q4 were discarded by a high probability for random success.

Table 4: Question ranked by the amount of information - Hashing Questionnaire

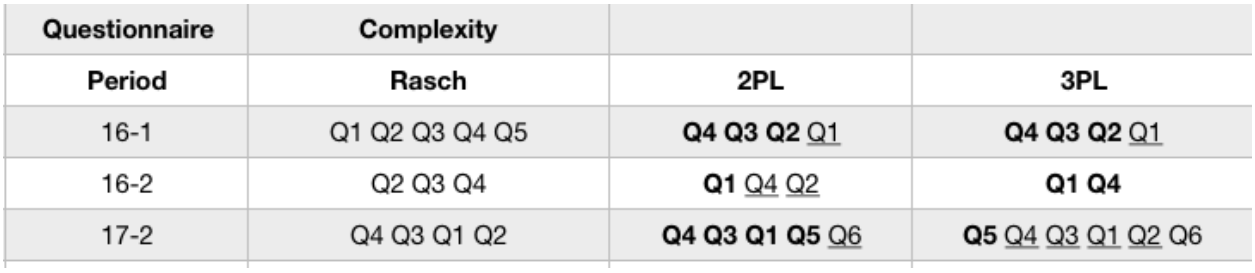


Questionnaire	**Hashing**		
Period	**Rasch**	**2PL**	**3PL**
16-1	Q1 Q3 Q4 Q5 Q7 Q8 Q9 Q11	**Q11 Q1** Q5 Q4	**Q4 Q11 Q1** Q5 Q7 Q9 Q10
16-2	Q1 Q3 Q4 Q7 Q9 Q11	**Q6** Q1 Q9	**Q1 Q5 Q6** Q4 Q11 Q10 Q9
17-2	Q1 Q3 Q4 Q5 Q7 Q8 Q9 Q11	**Q5 Q3 Q8** Q1 Q6	**Q1 Q3 Q5** Q6 Q8

Q: Low amount of info | **Q**: big amount of info **|** Q: Good amount of information

Source: the author.

Table 5: Question ranked by the amount of information - Algorithm Complexity Questionnaire



Questionnaire	Complexity		
Period	Rasch	2PL	3PL
16-1	Q1 Q2 Q3 Q4 Q5	**Q4 Q3 Q2** Q1	**Q4 Q3 Q2** Q1
16-2	Q2 Q3 Q4	**Q1** Q4 Q2	**Q1 Q4**
17-2	Q4 Q3 Q1 Q2	**Q4 Q3 Q1 Q5** Q6	**Q5** Q4 Q3 Q1 Q2 Q6

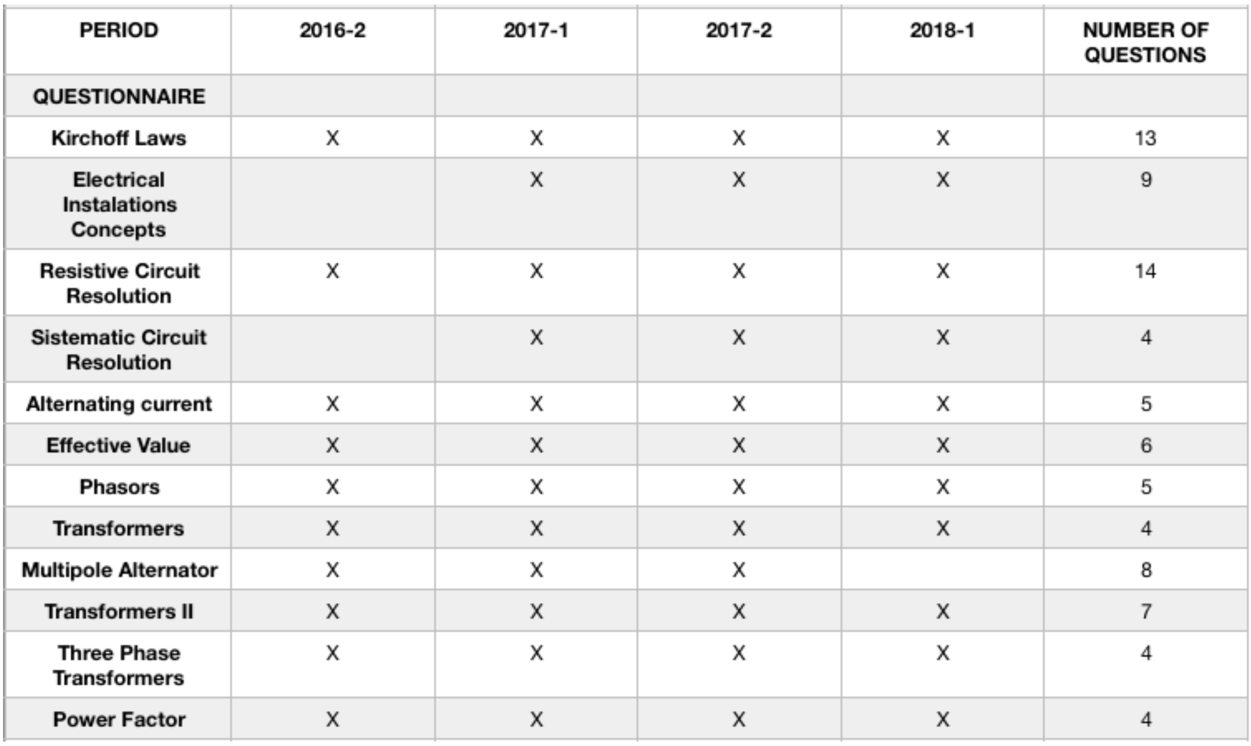
Q: Low amount of info | **Q**: big amount of info **|** Q: Good amount of information

Source: the author.

5.2 Experiment 2: “Electrical Engineering” course

The dataset collected from the “Electrical Engineering” course included twelve questionnaires extracted from Moodle. The questionnaires were answered over the periods of 2016-2 with 32 students, 2017-1 with 56 students, 2017-2 with 79 students and 2018-1 with 40 students. Table 6 describes the questionnaires, the periods and the questions used for the analysis.

Table 6: Questionnaires of “Electrical Engineering” course

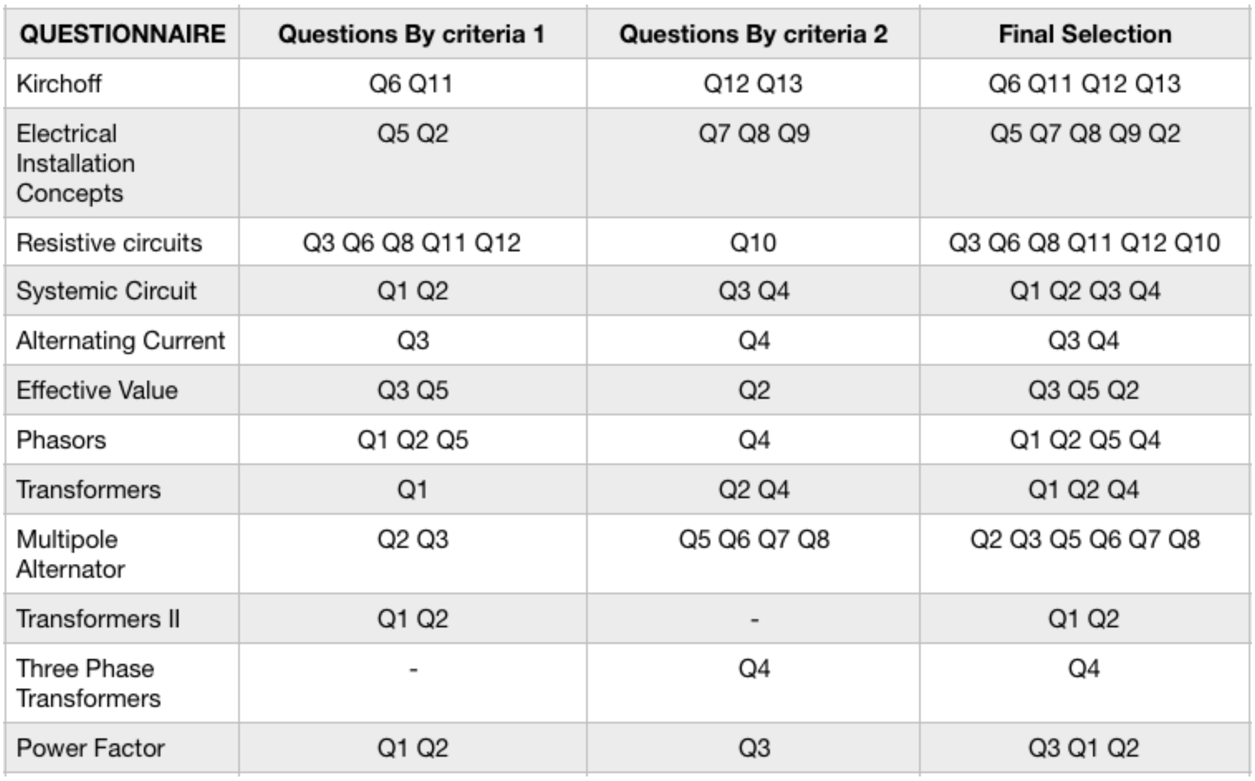


PERIOD	2016-2	2017-1	2017-2	2018-1	NUMBER OF QUESTIONS
QUESTIONNAIRE					
Kirchoff Laws	X	X	X	X	13
Electrical Instalations Concepts		X	X	X	9
Resistive Circuit Resolution	X	X	X	X	14
Sistematic Circuit Resolution		X	X	X	4
Alternating current	X	X	X	X	5
Effective Value	X	X	X	X	6
Phasors	X	X	X	X	5
Transformers	X	X	X	X	4
Multipole Alternator	X	X	X		8
Transformers II	X	X	X	X	7
Three Phase Transformers	X	X	X	X	4
Power Factor	X	X	X	X	4

Source: the author.

After the graphic analysis and creation of the tables according to criteria 1 (food information, good discrimination and a reasonable chance of success), and 2 (lots of information, high discrimination and low probability of success) (**see Appendix**), in table 7 are described the questions selected in each questionnaire.

Table 7: Question selected By criteria 1 and 2 for the **“Electrical Engineering”** course



QUESTIONNAIRE	Questions By criteria 1	Questions By criteria 2	Final Selection
Kirchoff	Q6 Q11	Q12 Q13	Q6 Q11 Q12 Q13
Electrical Installation Concepts	Q5 Q2	Q7 Q8 Q9	Q5 Q7 Q8 Q9 Q2
Resistive circuits	Q3 Q6 Q8 Q11 Q12	Q10	Q3 Q6 Q8 Q11 Q12 Q10
Systemic Circuit	Q1 Q2	Q3 Q4	Q1 Q2 Q3 Q4
Alternating Current	Q3	Q4	Q3 Q4
Effective Value	Q3 Q5	Q2	Q3 Q5 Q2
Phasors	Q1 Q2 Q5	Q4	Q1 Q2 Q5 Q4
Transformers	Q1	Q2 Q4	Q1 Q2 Q4
Multipole Alternator	Q2 Q3	Q5 Q6 Q7 Q8	Q2 Q3 Q5 Q6 Q7 Q8
Transformers II	Q1 Q2	-	Q1 Q2
Three Phase Transformers	-	Q4	Q4
Power Factor	Q1 Q2	Q3	Q3 Q1 Q2

Source: the author.

More Details of the experiment are described in the appendix section**.**

6. ANALYSIS AND DISCUSSION OF THE RESULTS

The software SPSS for windows (version 21) was used. The results of each questionnaire question were analyzed by means of one-way ANOVA followed by Tukey post hoc test, in order to determine differences between evaluating periods. In addition, differences among evaluating periods and questions were established for the Item Response Theory models. Differences between IRT models considering the parameter of difficulty level were assessed by one-way ANOVA, while differences using discriminatory index were verified by student's t-test. P-values <0.05 were considered significant and data are expressed as mean ± standard error (S.E).

6.1 Experiment 1: Analysis of "Data Classification and Searching" questionnaires

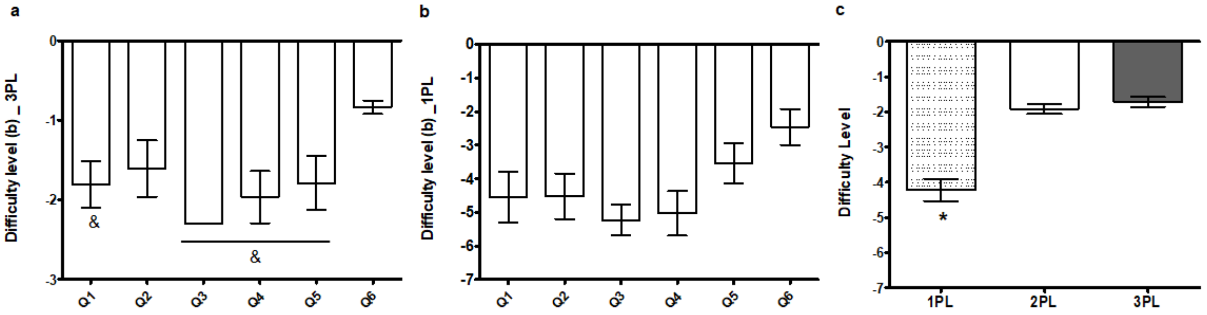
In this section, the analysis of the first experiment concerning the course on "Data Classification and Searching" is described.

6.1.1 Statistical analysis - questionnaire: Algorithms' complexity

The Analysis of the difficulty level of the "Algorithms' Complexity" questionnaire considering the logistic model of three parameters (3PL) of IRT showed that question number 6 had a higher difficulty level compared to questions 1, 3, 4 and 5 ($F_{(5,17} = 3.256$, $p< 0.05$ – Figure 11a). A similar pattern was seen in the logistic model of one parameter (1PL), evidencing once again that the question number 6 was more difficult than the other ones ($F_{(5,17)} = 2.833$, $p= 0.65$, n.s - Figure 11b). In order to analyze the general difficulty levels of the questionnaire, the three IRT logistic models were compared. There was a significant difference between model 1PL as compared to the other models ($F_{(5,17)}= 3.256$, $p< 0.05$ – Figure 11c), suggesting that this model could be more sensible to determine the difficulty level of the questionnaire, which in this case was classified as easy.

Taking each question of this questionnaire into consideration, one way-ANOVA did not show significant differences between them. Moreover, no significant differences ($P > 0,05$) were found between the evaluating periods when the following variables were evaluated: total qualification and time spent in the questionnaire.

Figure 11. Algorithms' Complexity questionnaire



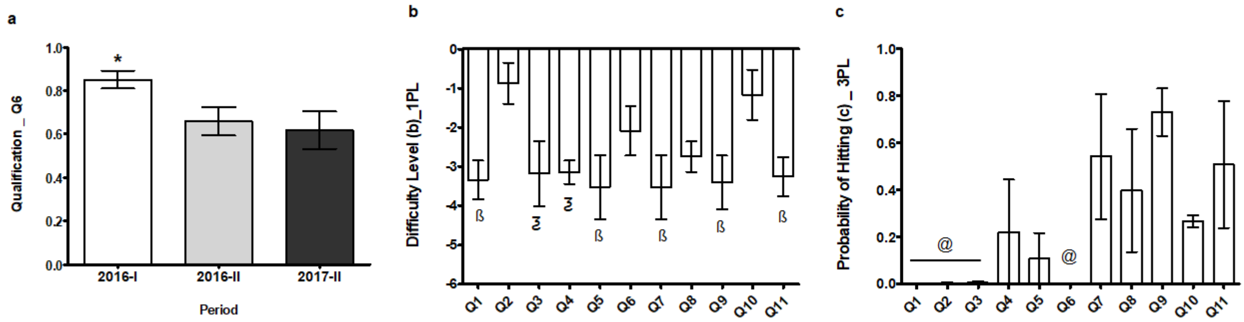
Difficulty levels by each question using the logistic model of one parameter (3PL) (a), and one parameter - 1PL (b). Evaluation of difficulty level by means of IRT models (c). $^{\&}$ Significant differences from question 6; *Significant differences from the other evaluating periods. Data analyzed using one-way ANOVA. Significance accepted p<0.05. Source: the author.

6.1.2 Statistical analysis - questionnaire: Hashing

As depicted in Figure 12a, there was a significant difference on question 6 of the questionnaire, indicating that it was perceived as more difficult in the period 2016-1 when compared to the other periods ($F_{(2,93)}$= 3.408, p< 0.05). No additional differences were found, considering the participant' results along the evaluating periods.

One-way ANOVA revealed a significant difference between the questions when the difficulty level of the questionnaire was established by means of the 1PL model ($F_{(10,32)}$= 2.311, p< 0.05); pairwise comparison showed that questions 2 and 10 were more difficult than questions 1, 5, 7, 9 and 11. In addition, questions 3 and 4 were easier when compared to question number 2, as shown in Figure 12b. No significant differences were observed between question number 6 and questions 2 and 10. The 3PL model was carried out in order to identify the probability of hitting. There was a significant difference between question 9 when compared to questions number 1, 2, 3, 5 and 6, as shown in Figure 12c.

Figure 12. Hashing Questionnaire



Result of a specific question by each period (a). Difficulty levels by each question using the logistic model of one parameter - 1PL (b). Probability of hitting by each question using the logistic model of three parameter - 3PL (c). *Significant differences from the other evaluating periods. ^{ß}Significant differences from questions 2 and 10. ^{ʒ}Significant differences from question

2. $^{@}$Significant differences from question 9. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

6.2 Experiment 2: Analysis of "Electrical Engineering" questionnaires

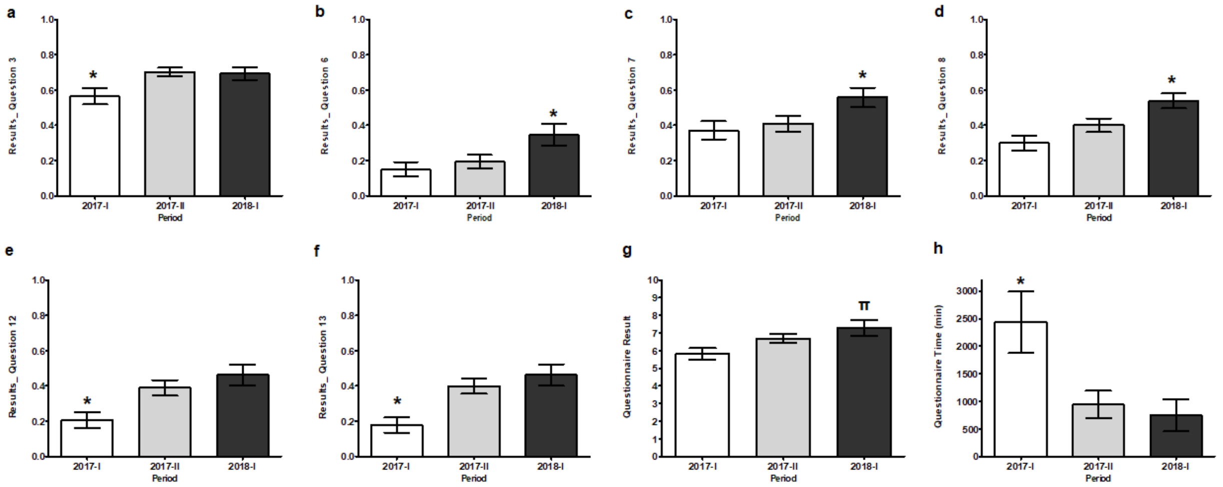
In this section, the analysis of the first experiment concerning the course on "Electrical Engineering" is described.

6.2.1 Statistical analysis - questionnaire 1: Kirchoff Laws

There was a significant effect of the evaluating period when each question was considered, indicating that during the period 2017-I the students had fewer hits for questions 3, 12, and 13 ($F_{(2,174)}$= 4.822; $F_{(2,174)}$= 6.415 and $F_{(2,174)}$= 8.601; p< 0.05, respectively - Figure 13a,e-f). Nevertheless, there were more hits for questions 6 to 8 ($F_{(2,174)}$= 4.096; $F_{(2,174)}$= 3.082 and $F_{(2,174)}$= 5.597; p< 0.05, respectively) when compared to the other periods (Figure 13b-d).

Participant's performance was better during the period 2018-I when compared to 2017-I, without significant difference from 2017-II ($F_{(2,174)}$= 4.642, p<0.05 - Figure 13g). In addition, the students spent more time completing the Kirchoff Laws questionnaire on the period 2017-I ($F_{(2,174)}$= 5,455, p<0.05) as shown in Figure 13h). These data could suggest modifications in teaching strategies of the questionnaire issues throughout the different academic periods.

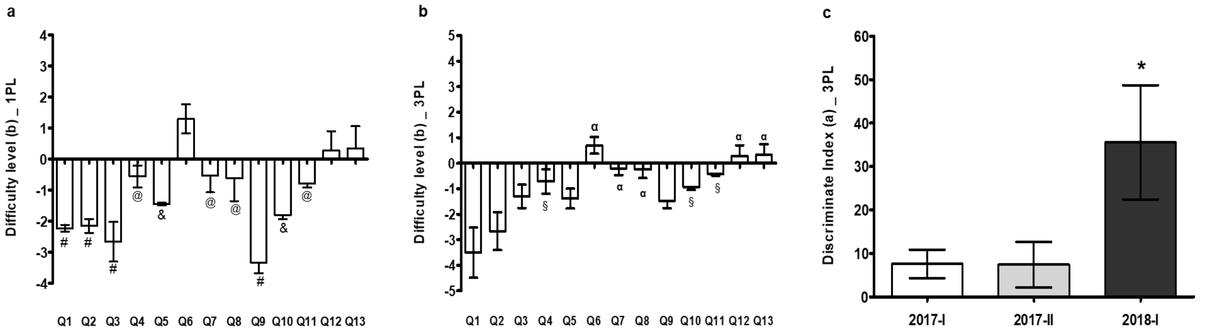
Figure 13. Kirchoff Laws Questionnaire.



Result of a specific question by each period (a-f); general result of the questionnaire (g); Time spent to complete the questionnaire (h). *Significant differences from the other evaluating periods. $^{\pi}$Significant difference from the period 2017-I. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

Using the IRT model in Questionnaire 1, statistical analysis showed significant differences for the logistic models 1PL and 3PL, as observed in figures 14a and 14b ($F_{(12,38)}$=8.492 and $F_{(12,38)}$=6.616; p< 0.05, respectively). Both models (using the difficulty level parameter - b) showed that questions 6, 12 and 13 were most difficult and questions 1 and 2 the easiest when compared to the other questions. Interestingly, the logistic model 1PL also indicated significant differences on question 9 as compared to questions 4, 7, 8 and 11. Moreover, when the logistic model 3PL was considered, the Kirchoff Laws questionnaire was more discriminative during the period 2018-I as compared to the other two periods ($F_{(2,38)}$=3.728, p< 0.05- Figure 14c).

Figure 14. Kirchoff Laws Questionnaire - IRT model

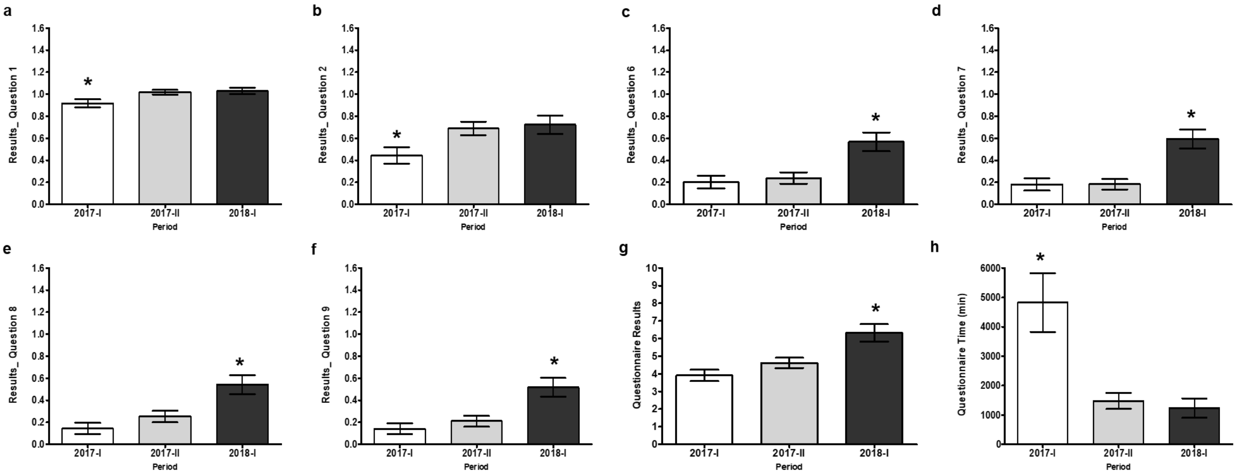


Difficulty levels by each question using the logistic model of one parameter - 1PL (a) and the one of three parameter - 3PL (b). Discrimination index by each period, using 3PL model. $^{\#}$Significant differences from questions 6, 12 and 13; $^{\&}$ Significant differences from question 6; $^{@}$Significant differences from question 9; §significant differences from question 1; $^{\alpha}$Significant differences from questions 1 and 2. *Significant differences from the other evaluating periods. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

6.2.2 Statistical analysis - questionnaire 2: Electrical Installations Concepts

One-way ANOVA revealed significant differences in most of the questions of the questionnaire concerning electrical installations concepts with the exception of questions 3, 4 and 5 (P> 0.05). As depicted in Figure 15 a-b, questions 1 and 2 had the lowest number of hits during the period 2017-I ($F_{(2,175)}$=4.290 and $F_{(2,175)}$=3.970; p< 0.05, respectively). In addition, significant differences were observed for questions 6, 7, 8 and 9, which had the highest number of hits by students on the 2018-I period (($F_{(2,175)}$=8.383; $F_{(2,175)}$=13.205; $F_{(2,175)}$=9.123 and $F_{(2,175)}$=9,071; p< 0.01 - Figure 15 c-f). The best qualification of the questionnaire and the shorter time required to completed it was observed in the period of 2018-I ($F_{(2,175)}$= 9.656 and $F2_{(2,175)}$= 9.718; p< 0.01 - Figure 15g-h).

Figure 15. Electrical Installations Concepts Questionnaire

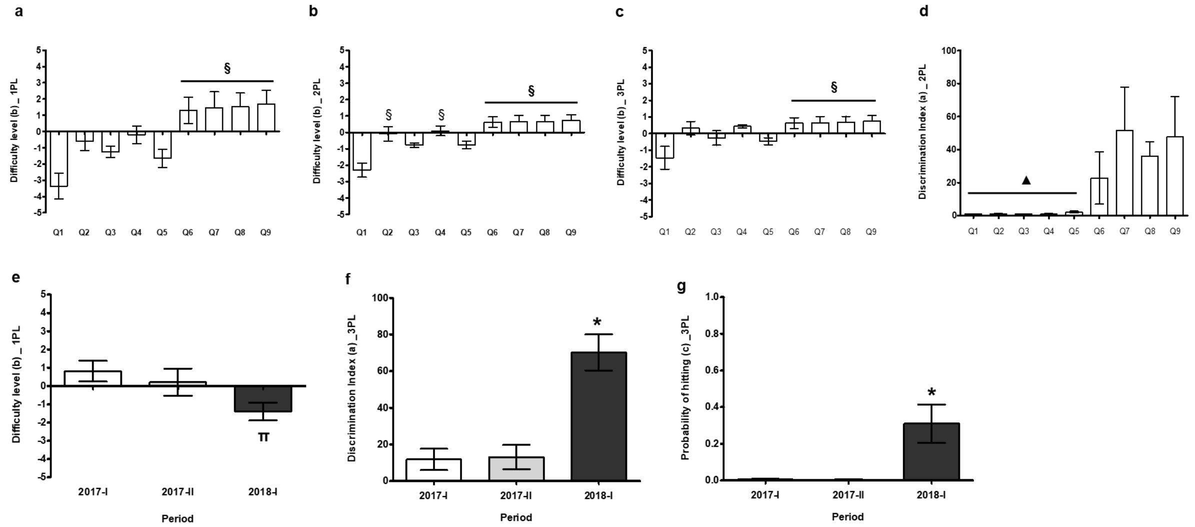


Result of a specific question by each period (a-f); general result of the questionnaire (g); Time spent to complete the questionnaire (h). *Significant differences from the other evaluating periods. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

Using IRT analysis, questions 6 to 9 were more difficult than the other ones and this behavior was observed in the three logistic models (1PL, 2PL, and 3PL), confirming the result ($F_{(8,26)}$=5.901; $F_{(8,26)}$=9.068 and $F_{(8,26)}$=3.648; p< 0.05), as shown in the Figures 16a-c. The discrimination index of the 2PL model ($F_{(8,26)}$= 2.619, p< 0.05) is represented in Figure 16d, indicating that questions 7 and 9 were more discriminative than questions 1 to 5. No significant differences were observed in question 6 and 8.

Analyzing the difficulty level by the 1PL model, evidenced that during the period 2018-I the questionnaire was easier in comparison to the period 2017-1, without significant differences from 2017-II ($F_{(2,26)}$=3.559, p< 0.05 - Figure 16e). The discrimination index in the same period was higher ($F_{(2,26)}$=19.238, p< 0.01 - Figure 16f) as well as the probability of hitting when compared to the other periods, as shown in Figure 16g ($F_{(2,26)}$= 8.559, p< 0.05).

Figure 16. Electrical Installations Concepts Questionnaire - IRT model



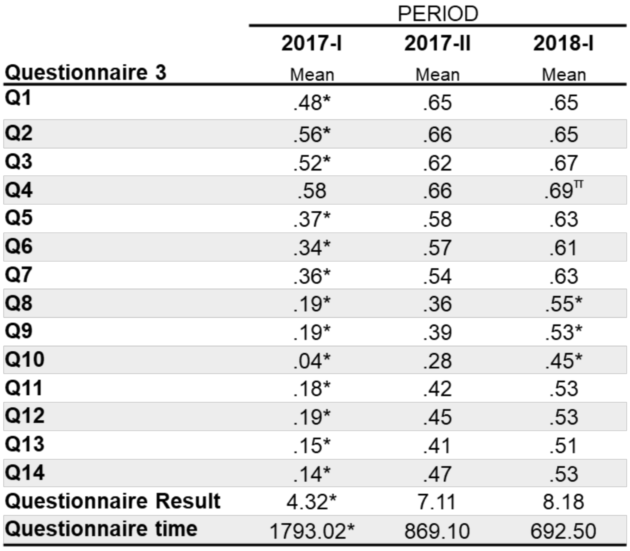
Difficulty level by each question using IRT model (a-c). Discrimination index using the logistic model of two parameter – 2PL (d). Evaluation of difficulty level (e), discrimination index (f) and the probability of hitting (g) by each evaluating period. §Significant differences from question 1; ▲Significant differences from questions 7 and 9; πSignificant difference from the period 2017-I; *Significant differences from the other evaluating periods. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

6.2.3 Statistical analysis - questionnaire 3: Resistive Circuit Resolution

The results of the resistive circuit resolution questionnaire indicated that during the 2017-I period, the students had the worse performance in all questions ($F_{(2,164)}=24.320$, $p< 0.01$) and required also more time to complete it ($F_{(2,164)}=4.233$, $p< 0.05$). In addition, on the 2018-I period, the students showed an increase in the number of hits for questions 8, 9 and 10 when compared to the other periods (Table 8).

Using the three models of IRT (1PL, 2PL and 3PL - $F_{(13,41)}=15.639$; $F_{(13,41)}=6.467$ and $F_{(13,41)}=7.361$; $p< 0.05$, respectively), it was observed that questions 1 to 4 were the easiest, following by questions 5 to 7 (Figure 17a-c). When models 2PL and 3PL were applied ($F_{(2,41)}=51.420$; $F_{(2,,41)}=47.303$; $p< 0.001$) the questionnaire was more discriminative on the 2017-I period in comparison to 2017-II and 2018-I (Figure 17d-e). Finally, in Figure 17f, the 3PL model revealed that the 2017-II period had significant differences from the other ones, being considered with a higher difficulty level ($F_{(2,41)}=3.407$; $p< 0.05$).

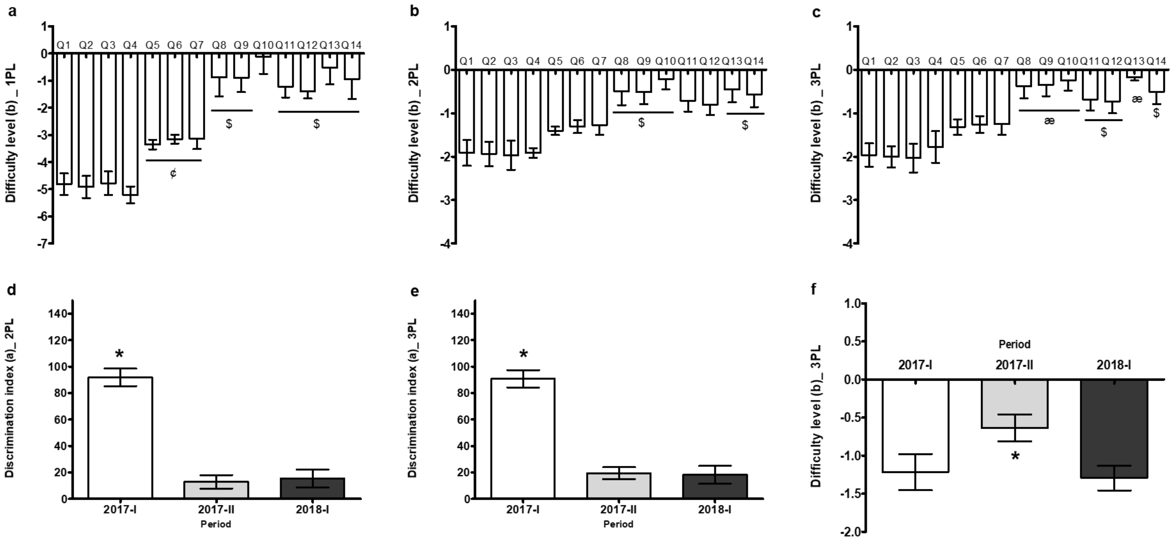
Table 8: Description of questionnaire result by each evaluating period



	PERIOD		
	2017-I	2017-II	2018-I
Questionnaire 3	Mean	Mean	Mean
Q1	.48*	.65	.65
Q2	.56*	.66	.65
Q3	.52*	.62	.67
Q4	.58	.66	.69π
Q5	.37*	.58	.63
Q6	.34*	.57	.61
Q7	.36*	.54	.63
Q8	.19*	.36	.55*
Q9	.19*	.39	.53*
Q10	.04*	.28	.45*
Q11	.18*	.42	.53
Q12	.19*	.45	.53
Q13	.15*	.41	.51
Q14	.14*	.47	.53
Questionnaire Result	4.32*	7.11	8.18
Questionnaire time	1793.02*	869.10	692.50

πSignificant difference from the period 2017-I. *Significant differences from the other evaluating periods. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

Figure 17. Resistive Circuit Resolution Questionnaire – IRT model

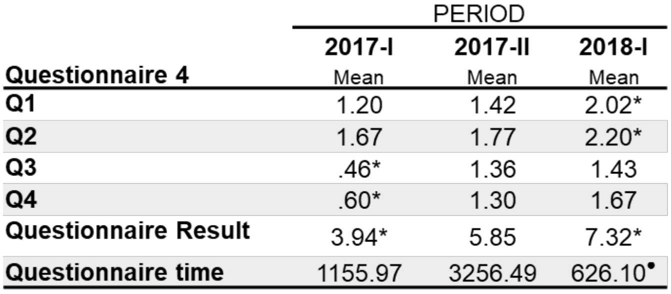


Difficulty level by each question using IRT model (a-c). Evaluation of discrimination index using the logistical model of two parameter – 2PL (d) and three parameter -3PL (e) by each evaluating period. Difficulty level by each period according to the logistical model of three parameter-3PL (f). $Significant differences from the questions 1 to 4. ¢ Significant difference from the questions 10 and 13. æ Significant differences from question 1 to 7 *Significant differences from the other evaluating periods. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

6.2.4 Statistical analysis - questionnaire 4: Systemic Circuit Resolution

The outcomes on 2017-I showed that questions 3-4 had lower hits by the students ($F_{(2,174)}=11.515$; $F_{(2,174)}=10.278$; $p< 0.05$, respectively), while on period 2018-I an increase on the number of correct answers for the questions 1 and 2 was observed ($F_{(2,174)}=5.851$; $F_{(2,1741)}=3.155$; $p< 0.05$). In addition, at 2018-I the time used to complete the questionnaire was lower in comparison to the other evaluating periods. Surprisingly, students spent more time resolving the questionnaire on the 2017-II period ($F_{(2,41)}=3.354$; $p< 0.05$ - Table 9).

Table 9. Description of questionnaire result by each evaluating period.

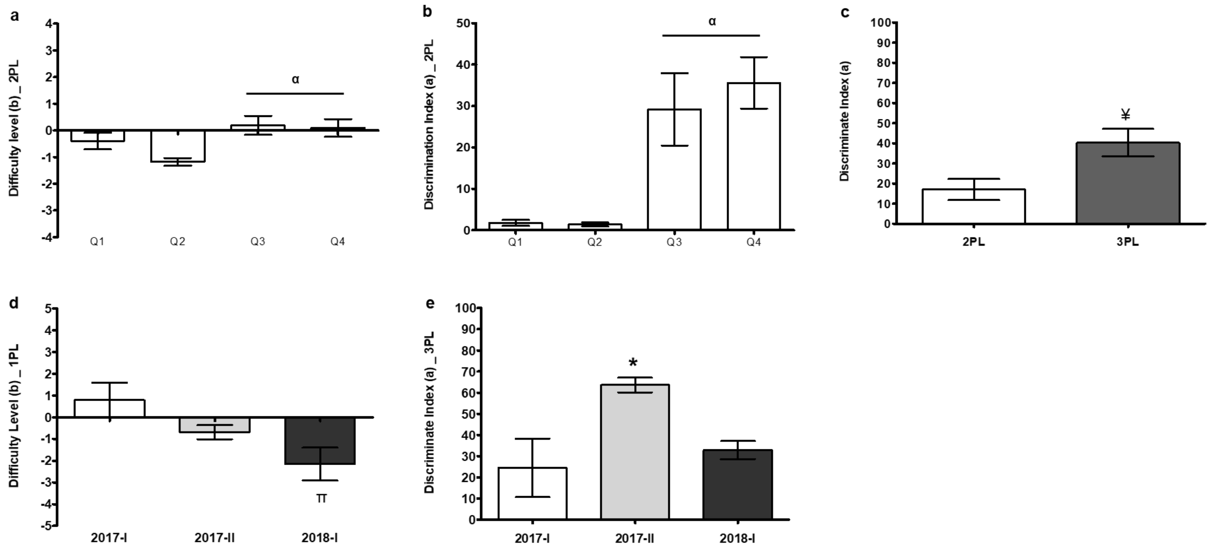


	PERIOD		
	2017-I	**2017-II**	**2018-I**
Questionnaire 4	Mean	Mean	Mean
Q1	1.20	1.42	2.02*
Q2	1.67	1.77	2.20*
Q3	.46*	1.36	1.43
Q4	.60*	1.30	1.67
Questionnaire Result	3.94*	5.85	7.32*
Questionnaire time	1155.97	3256.49	626.10•

*Significant differences from the other evaluating periods. •Significant difference from the period 2017-II. Data analyzed by one-way ANOVA. Significance accepted $p<0.05$. Source: the author.

Model 2PL evidenced that questions 3 and 4 had higher difficulty level ($F_{(3,11)}=4.452$, $P<0.05$ - Figure 18a) as well as higher discrimination levels ($F_{(3,11)}=11.139$, $P<0.01$-Figure 18b) when compared to the other two questions. However, when considering the discrimination index using 2PL and 3PL models, the last one was shown to be more discriminating than 2PL ($t_{(22)}=2.729$, $P<0.01$), which could be related to intrinsic characteristics of the 3PL model, due to other variables included (Figure 18c). As observed in previous questionnaires, on the 2018-I period, the systemic circuit resolution questionnaire had also lower difficulty level when compared to 2017-I ($F_{(2,11)}=4.919$, $P<0.05$ - Figure 18d). Finally, 3PL for 2017-II was more discriminative than the other periods analyzed ($F_{(2,11)}=5.785$, $P<0.05$ - Figure 18e).

Figure 18. Systemic Circuit Resolution Questionnaire - IRT model

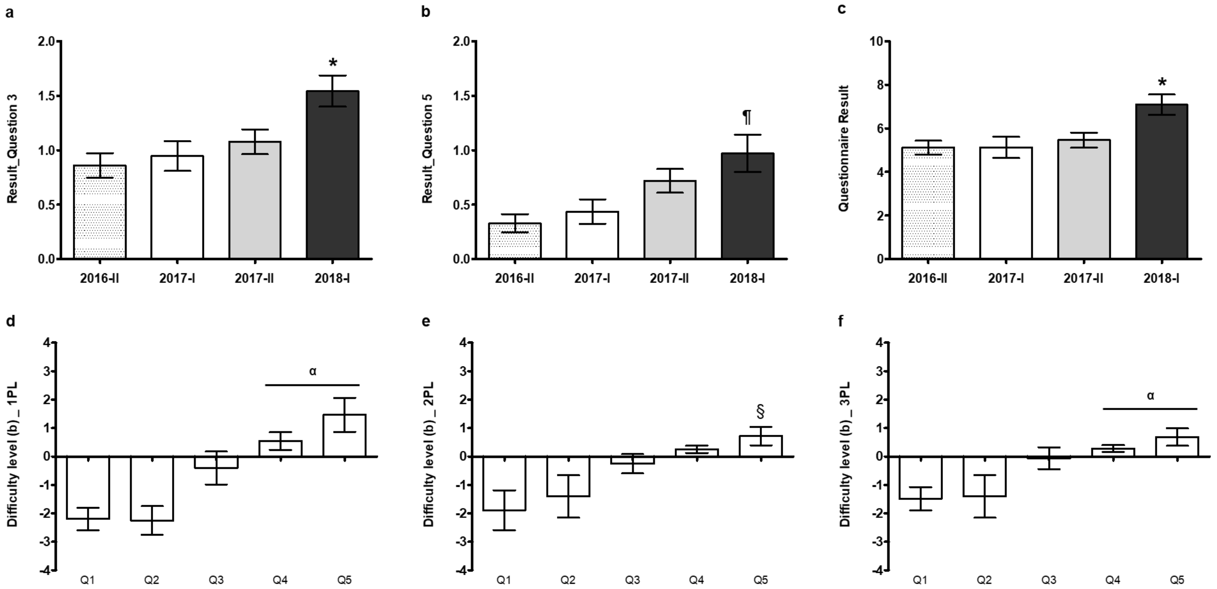


Difficulty level according to the logistic model of one parameter – 2PL (a); Discrimination index using the logistic model of two parameter – 2PL (b) Evaluation of discrimination index by means of two models (c). Results considering the evaluating period by difficulty level (d) and discrimination index (e). $^{\alpha}$Significant differences from question 1 and 2. $^{\pi}$Significant differences from the period 2017-I. *Significant differences from the other evaluating periods. $^{\yen}$Significant differences from the other models. Data analyzed by one-way ANOVA and *t*-test by independent samples. Significance accepted p<0.05. Source: the author.

6.2.5 Statistical analysis - questionnaire 5: Alternating Current

One-way Anova for the alternating current questionnaire showed significant differences for questions 3, 5 and the general qualification ($F_{(3,245)}=4.015$; $F_{(3,245)}=5.459$ and $F_{(3,245)}=3.662$; P<0.05), without significant differences by questions 1, 2 and the time spent during the evaluation (Figure 19a-c). There was observed a consistent result regarding the difficulty level of the alternating current questionnaire when IRT was used. The three models (1PL: $F_{(4,19)}=11.320$; 2PL: $F_{(4,19)}=4.708$ and 3PL: $F_{(4,19)}=4.977$; p< 0.01) revealed that questions 4 and 5 were the most difficult while questions 1 and 2 the easiest (Figure 19d-f). No significant differences were identified in other parameters of the IRT model as well as in the evaluating period (p>0.05).

Figure 19. Alternating Current Questionnaire



Result of specific questions by each period (a-c). Difficulty level using IRT model (d-f). *Significant differences from the other evaluating periods. ¶Significant differences from the period's 2016-2 and 2017-I. §Significant differences from question 1. $^{\alpha}$Significant differences from questions 1 and 2. Data analyzed by one-way ANOVA. Significance accepted $p<0.05$. Source: the author.

6.2.6 Statistical analysis - questionnaire 6: Effective Value

During the periods of 2016-2 and 2017-I, there was a significant reduction in the number of hits by all the questions, except for question 6 ($F_{(3,262)}$= 7.662 -22.097; $p<0.001$). Although in the period 2018-I the performance per question was better, the time used by the students was longer in comparison with the other periods ($F_{(3,262)}$= 3,263; $p<0.05$ - Table 10). This could suggest that time and final result are independent variables and could not be strictly related to student performance.

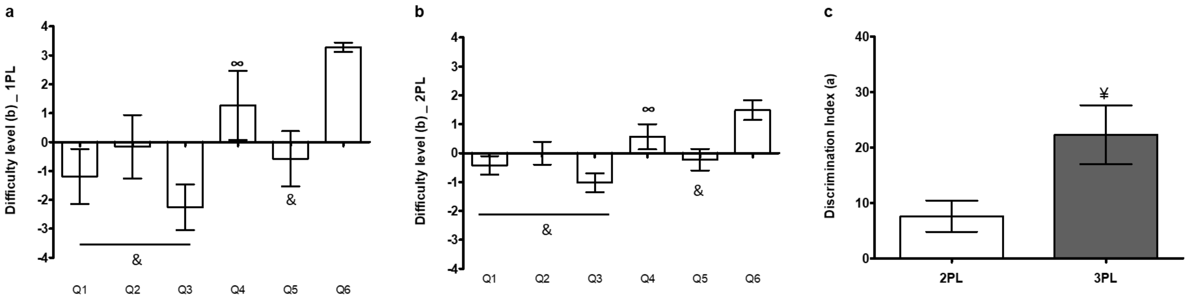
Table 10: Description of questionnaire result by each evaluating period.

	PERIOD			
	2016-II	**2017-I**	**2017-II**	**2018-I**
Questionnaire 6	Mean	Mean	Mean	Mean
Q1	.86	.67	1.28¶	1.43¶
Q2	.47	.49	1.13¶	1.34¶
Q3	1.12	.97	1.42¶	1.51¶
Q4	.37	.21	.52	1.22*
Q5	.65	.55	1.28¶	1.30¶
Q6	.22	.21	.17	.29
Questionnaire Result	3.68	3.09	5.84¶	7.07*
Questionnaire time	943.94	1132.88	680.12	1960.74*

¶Significant differences from the period's 2016-2 and 2017-I. *Significant differences from the other evaluating periods. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

There was a similar result regarding the degree of difficulty when the 1PL and 2PL models were used ($F_{(5,23)}$=4.511; $F_{(5,23)}$=5.634 P<0.05, respectively). Both models indicating that question 1 to 3 were easy when compared to question 6 (Figure 20a-b). The analysis of the discrimination index using 2PL and 3PL models showed a significant difference between them, indicating that 3PL had an upper index ($t_{(46)}$=2.447, P<0.05 - Figure 20c). No significant differences were found when the evaluating period was considered or when the other parameters of the IRT model were included.

Figure 20. Effective Value Questionnaire - IRT Model

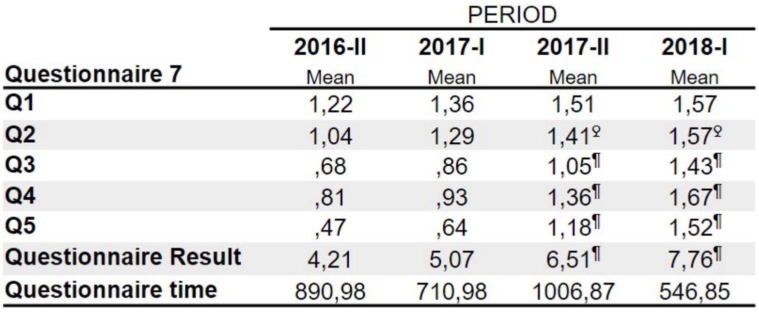


Difficulty level by each question using the logistic model of one parameter – 1PL (a); two parameter – 2PL (b). Evaluation of the discrimination index by means of two models (c). $^{\&}$ Significant differences from question 6. $^{\infty}$Significant differences from question 3. $^{\text{¥}}$Significant differences from the other model. Data analyzed by one-way ANOVA and *t*-test by independent samples. Significance accepted p<0.05. Source: the author.

6.2.7 Statistical analysis - questionnaire 7: Phasors

The analysis of each question of the Phasors questionnaire showed that, during the periods of 2017-II and 2018-I almost all the questions were resolved by students ($F_{(3,252)}$=3.484-16.024, p< 0.05), being the questionnaire considered easy (Table 11). As observed in Figures 21c-e, this information was similar when the difficulty level was determined by means of the three IRT models (1P: $F_{(3,19)}$=6.700, p< 0.01; 2PL:$F_{(3,19)}$=9.909, p< 0.001 and 3PL: $F_{(3,19)}$= 13.681, p< 0.0001), indicating that IRT is able to identify significant differences as observed through other methods. Interestingly, only one question (number 4) had a high discriminatory index for the Pharsors questionnaire when analyzed through 2PL and 3PL ($F_{(4,19)}$=5.452 and $F_{(4,19)}$=7.324, p<0,01, respectively - Figure 21a-b).

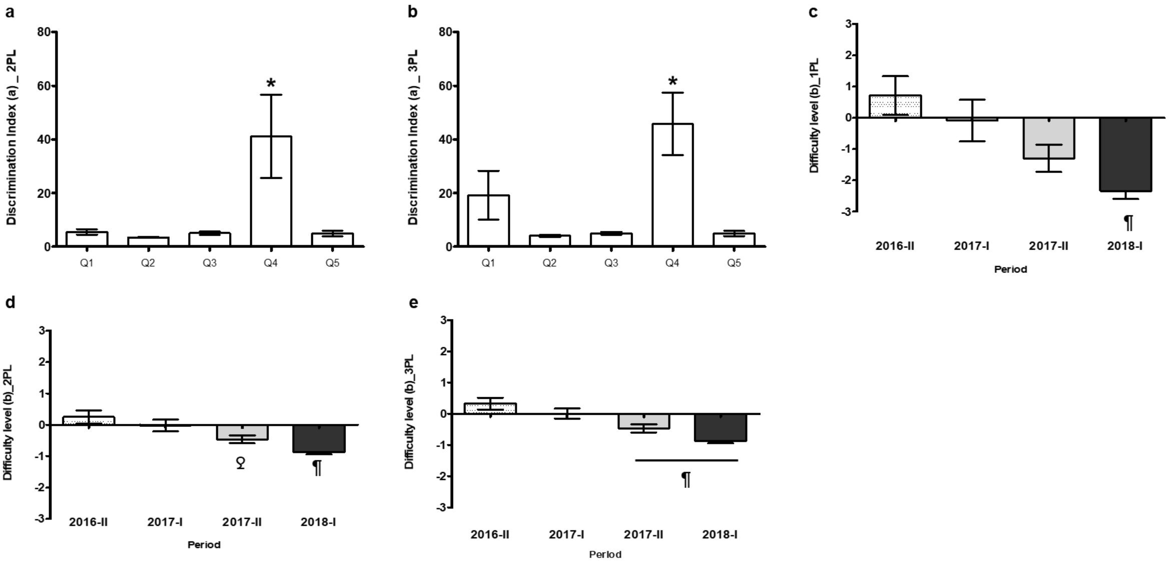
Table 11. Description of questionnaire result by each evaluating period.



	PERIOD			
	2016-II	**2017-I**	**2017-II**	**2018-I**
Questionnaire 7	Mean	Mean	Mean	Mean
Q1	1,22	1,36	1,51	1,57
Q2	1,04	1,29	1,41º	1,57º
Q3	,68	,86	1,05¶	1,43¶
Q4	,81	,93	1,36¶	1,67¶
Q5	,47	,64	1,18¶	1,52¶
Questionnaire Result	4,21	5,07	6,51¶	7,76¶
Questionnaire time	890,98	710,98	1006,87	546,85

¶Significant differences from the period's 2016-2 and 2017-I. ºSignificant differences from the period 2016-2. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

Figure 21. Phasors Questionnaire - IRT Model

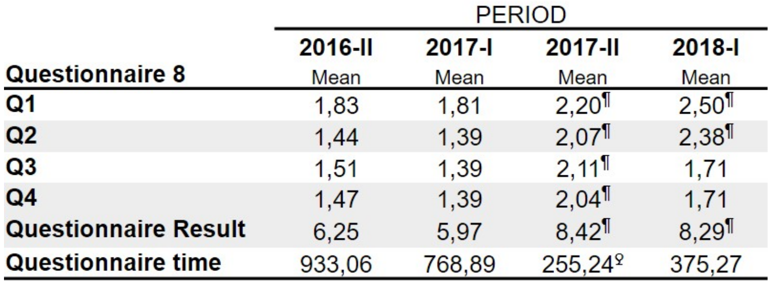


Discrimination index by each question using the logistic model of one parameter – 1PL (a) and the two-parameter model – 2PL (b). Difficulty level using IRT model by each evaluating period (c-e). *Significant differences from the other questions. ¶Significant differences from the period's 2016-2 and 2017-I. ºSignificant differences from the period 2016-2. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

6.2.8 Statistical analysis - questionnaire 8: Transformers

The analysis of each question using as a fixed factor the evaluating period revealed that during the periods 2017-II and 2018-I the students' performance was better when compared with previous periods ($F_{(3,248)}$=5.286 -11.113, p<0.001 - Table 12). Moreover, during the 2017-II period, students required very few minutes to complete the questionnaire ($F_{(3,248)}$=2.729, p<0.05).

Table 12. Description of questionnaire result by each evaluating period.

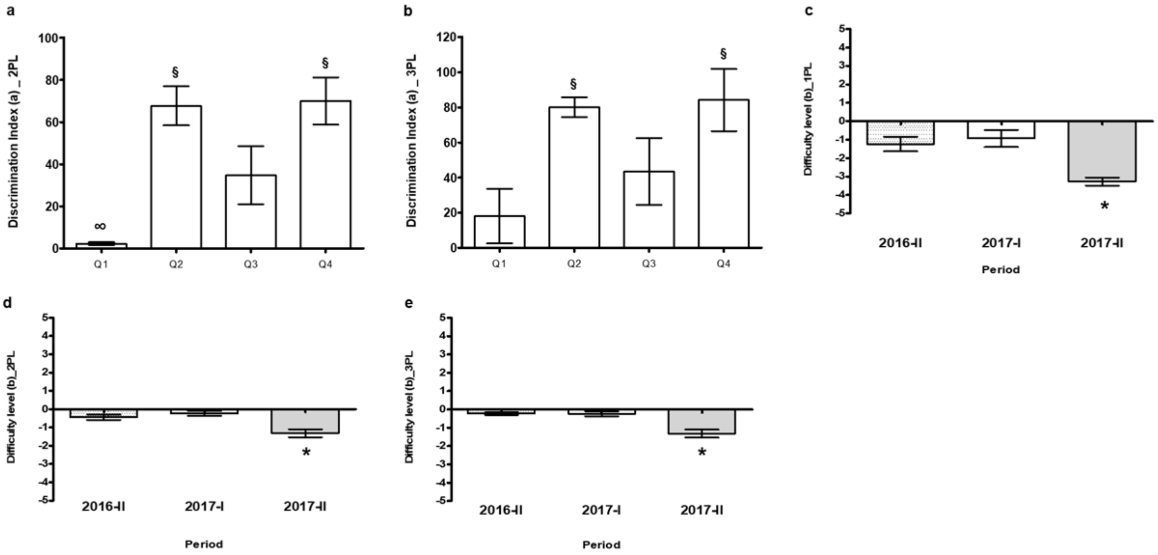


	PERIOD			
	2016-II	**2017-I**	**2017-II**	**2018-I**
Questionnaire 8	Mean	Mean	Mean	Mean
Q1	1,83	1,81	2,20¶	2,50¶
Q2	1,44	1,39	2,07¶	2,38¶
Q3	1,51	1,39	2,11¶	1,71
Q4	1,47	1,39	2,04¶	1,71
Questionnaire Result	6,25	5,97	8,42¶	8,29¶
Questionnaire time	933,06	768,89	255,24♀	375,27

¶Significant differences from the period's 2016-2 and 2017-I. ♀Significant differences from the period 2016-2. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

As depicted in Figures 22a-b, the discrimination index using 2PL and 3PL models showed that question 1 was the least discriminating when compared to the other points ($F_{(3,11)}$=10.134 and $F_{(3,11)}$=4.161, p<0.05 respectively). In the period 2018-I, it was not possible to use IRT due to the lower variability in the results. However, previous periods (2016-2 to 2017-II) were considered. Thereby, it was observed that the difficulty level of the questionnaire was low in 2017-II, as mentioned above (1PL: $F_{(211)}$=11.571, 2PL: $F_{(211)}$=11.176 and 3PL: $F_{(211)}$=16.500, p<0.01 - Figures 22c-e). No additional differences were found considering IRT parameters.

Figure 22. Transformers Questionnaire - IRT Model



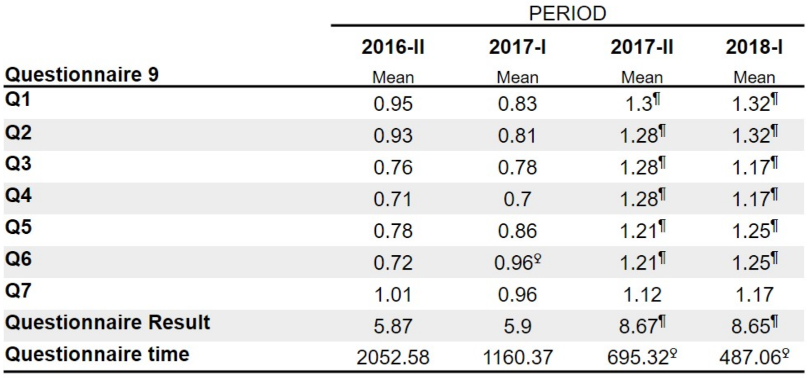
Discrimination index by each question using the logistic model of one parameter – 1PL (a) and the two-parameter mode – 2PL (b). Difficulty level using IRT model by each evaluating period (c-e). §Significant differences from question 1. ∞Significant differences from question 3. *Significant differences from the other evaluating periods. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

6.2.9 Statistical analysis - questionnaire 9: Transformers II

Table 13 describes the performance of the students over four different periods. Thus, it was observed that all the questions were resolved in the periods 2017-II and 2018-I ($F_{(3,245)}$=8.824 -15.088, p< 0.05). This result was in agreement with IRT analysis, using 1PL ($F_{(3,27)}$=25,196, p< 0.0001 - Figure 23b), 2PL ($F_{(3,27)}$=47.009, p< 0.0001 - Figure 23c) and 3PL models ($F_{(3,27)}$=46.263, p< 0.0001 - Figure 23d), indicating lower difficulty level in these two periods.

It is interesting to note that there were few questions with a high discrimination index, being the questions 3 and 4 the most significant in comparison with questions 1, 2 and 7 ($F_{(6,27)}$=2.929, p< 0.05). No significant differences were observed in questions 5 and 6, as shown in Figure 23a.

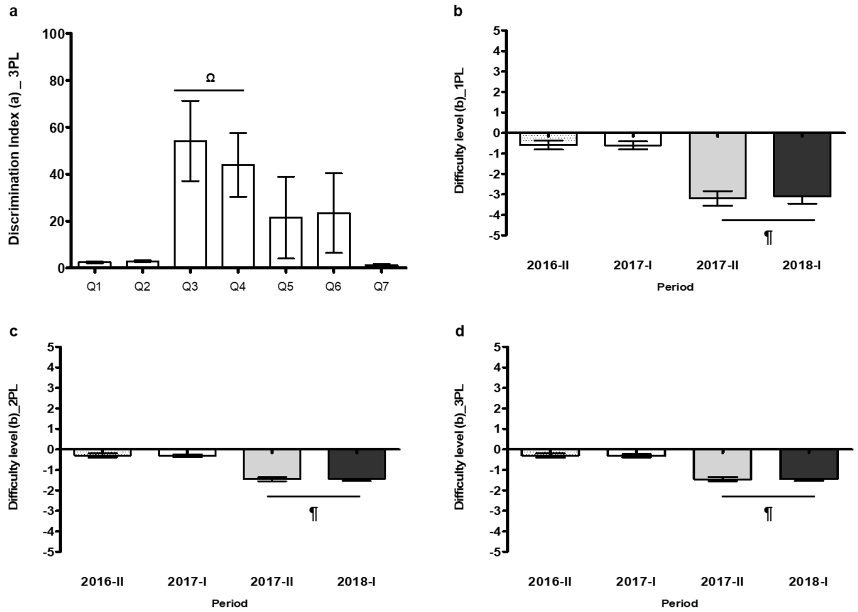
Table 13. Description of questionnaire result by each evaluating period.



	PERIOD			
	2016-II	2017-I	2017-II	2018-I
Questionnaire 9	Mean	Mean	Mean	Mean
Q1	0.95	0.83	1.3¶	1.32¶
Q2	0.93	0.81	1.28¶	1.32¶
Q3	0.76	0.78	1.28¶	1.17¶
Q4	0.71	0.7	1.28¶	1.17¶
Q5	0.78	0.86	1.21¶	1.25¶
Q6	0.72	0.96º	1.21¶	1.25¶
Q7	1.01	0.96	1.12	1.17
Questionnaire Result	5.87	5.9	8.67¶	8.65¶
Questionnaire time	2052.58	1160.37	695.32º	487.06º

¶Significant differences from period's 2016-2 and 2017-I. ºSignificant differences from period 2016-2. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

Figure 23. Transformers II Questionnaire - IRT Model



Discrimination index by each question using the logistic model of three parameter – 3PL (a). Difficulty level using IRT models by each evaluating period (b-de). ΩSignificant differences from questions 1, 2 and 7. ¶Significant differences from periods 2016-2 and 2017-I. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

6.2.10 Statistical analysis - questionnaire 10: Three-phase Transformers

One-way Anova evidenced significant differences by each question when the evaluation period was considered ($F_{(3,245)}$=3.703-26.624, p< 0.01). Pairwise comparison indicated that during the 2016-2 period students had the worst performance for the Three-Phase Transformer questionnaire. Peculiarly, during periods 2017-I, 2017-II and 2018-II the number of hits by each question as well as the general questionnaire results were relatively similar between them, as shown in Table 14. Nevertheless, on 2017-II, the time used to complete the questionnaire was the lowest in comparison with the other periods.

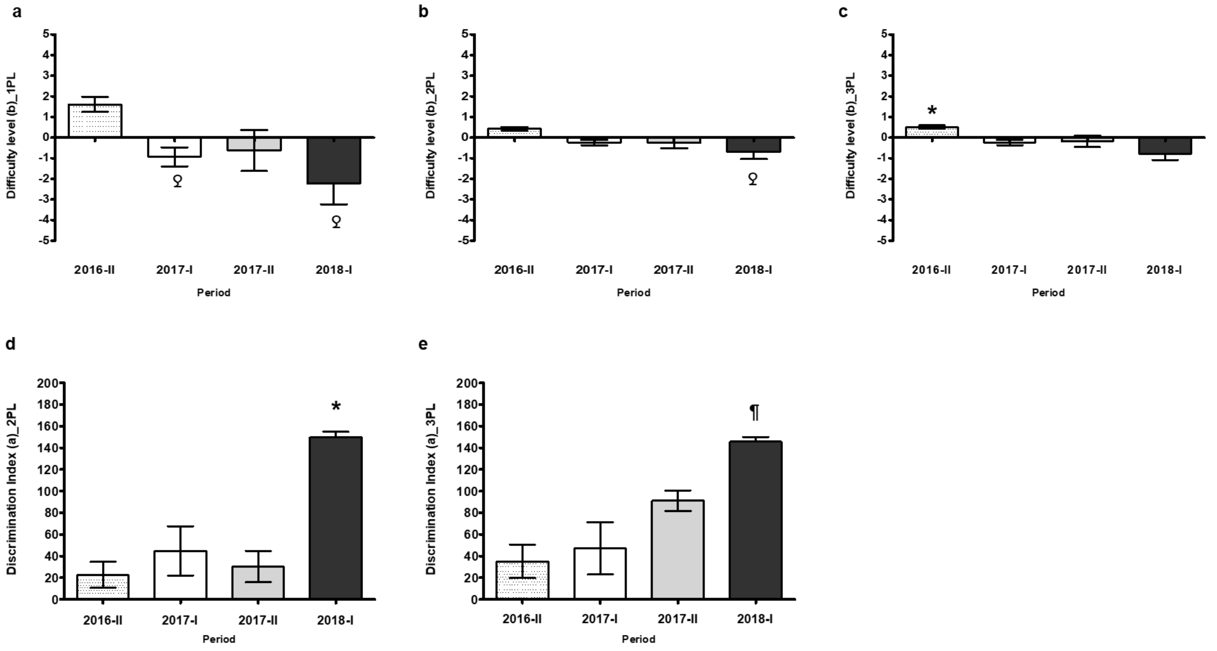
Table 14. Description of questionnaire result by each evaluating period.

	PERIOD			
	2016-II	2017-I	2017-II	2018-I
Questionnaire 10	Mean	Mean	Mean	Mean
Q1	0.9	1.49Φ	0.97	1.45Φ
Q2	0.9	1.54♀	1.83♀	2.17¶
Q3	0.83	1.59Φ	0.97	1.45Φ
Q4	0.59	1.59♀	1.87♀	2.17¶
Questionnaire result	3.21*	6.01	5.77	7.24
Questionnaire time	1368.29	758.5	42.78♀	562.81

ΦSignificant differences from periods 2016-2 and 2017-II. ¶Significant differences from the periods 2016-2 and 2017-I. ♀Significant differences from the period 2016-2. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

Using the parameter difficulty level of the three IRT models, it was observed that the Three-Phase Transformers questionnaire was perceived as more difficult on 2016-2, especially when it was compared with the periods of 2017-I and 2018-II (1PL: $F_{(3,15)}=4.348$, 2PL: $F_{(3,15)}=3.962$ and 3PL: $F_{(3,15)}=6.155$, $p< 0.05$ - Figures 24a-c). No significant differences were found when questions were used as a fixed variable after IRT analysis. There was a significant effect of the period in which the questionnaire was applied by the discriminatory index. The 2PL and 3PL models ($F_{(3,15)}=15.673$ and $F_{(3,15)}=10.812$, $p< 0.001$) showed a similar result, indicating a high discrimination on the 2018-I period, as indicated in Figure 24d-e.

Figure 24. Three Phase Transformers Questionnaire - IRT Model



Difficulty level using IRT model by each evaluating period (a-c). Discrimination index using the logistic model of two parameters – 2PL (d) and 3PL (e). ♀Significant differences from 2016-2. ¶Significant differences from 2016-2 and 2017-I. *Significant differences from the other evaluating periods. Data analyzed by one-way ANOVA. Significance accepted $p<0.05$. Source: the author.

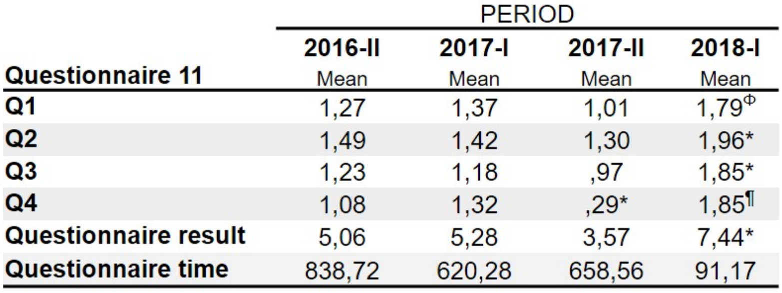
6.2.11 Statistical analysis - questionnaire 11: Power Factor

The analysis of the Power factor questionnaire showed a similar pattern as observed in previous questionnaires (Table 15). In the period of 2018-I, there was a better student' performance for all the questions in comparison with 2016-2 ($F_{(3,250)}=2.836\text{-}8.048$, $p< 0.05$). Moreover, some differences were observed between 2018-I and the periods of 2017-I and 2017-II, especially for questions 1 ($F_{(3,250)}=3.706$, $p< 0.05$) and 4 ($F_{(3,250)}=20.156$, $p< 0.001$).

One-way ANOVA evidenced a significant difference when the evaluation period was used as a fixed variable. The logistic model of one parameter (1PL) showed that the difficulty level was low on 2018-I and high in 2017-II in comparison to the other periods ($F_{(3,15)}$=10.580 p< 0.001-Figure 25a). The logistic model of three parameters (3PL) confirmed the significant differences between these two time points ($F_{(3,15)}$=5.064, p< 0.05- Figure 25b).

The discrimination index of the Power Factor questionnaire showed significant differences when the 2PL model was used, indicating a high index in the 2018-I period when compared to the other periods ($F_{(3,15)}$=4.174, p< 0.05-Figure 25c**)**. Model 3PL identified significant differences between 2018-I and 2017-I ($F_{(3,15)}$=5.051, p< 0.05- Figure 25d). Thereby, these subtle differences among them could be related to the variables included by each model during IRT analysis. No significant differences were found when questions were used as a fixed variable.

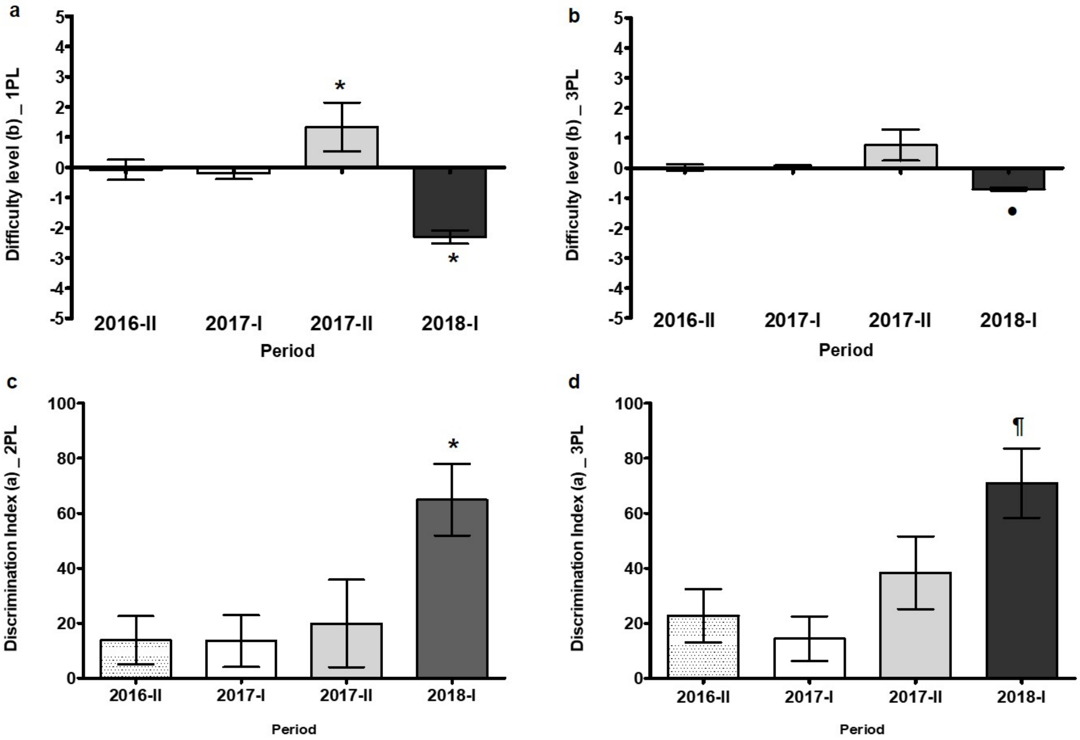
Table 15. Description of questionnaire 11 results by each evaluating period.



	PERIOD			
	2016-II	**2017-I**	**2017-II**	**2018-I**
Questionnaire 11	Mean	Mean	Mean	Mean
Q1	1,27	1,37	1,01	1,79$^{\Phi}$
Q2	1,49	1,42	1,30	1,96*
Q3	1,23	1,18	,97	1,85*
Q4	1,08	1,32	,29*	1,85¶
Questionnaire result	5,06	5,28	3,57	7,44*
Questionnaire time	838,72	620,28	658,56	91,17

$^{\Phi}$Signifcant differences from periods 2016-2 and 2017-II. ¶Significant differences from periods 2016-2 and 2017-I. *Significant differences from the other evaluating periods. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

Figure 25. Power Factor Questionnaire - IRT Model



Difficulty level using 1PL (a) and 2PL by each evaluating period (b). Discrimination index using 2PL (c) and 3PL (d). *Significant differences from the other evaluating periods. •Significant difference from period 2017-II. ¶Significant differences from periods 2016-2 and 2017-I. Data analyzed by one-way ANOVA. Significance accepted p<0.05. Source: the author.

6.3 Final Considerations

From the analysis to the questionnaires of a group of students it was possible to apply a methodology for the selection of questions based on IRT. To discuss the results, the following questions were proposed:

RQ1: Are all the questions necessary for a questionnaire? Are some of them more important than others?
RQ2: Can it be established a ranking of importance (or contribution) of the questions of a questionnaire?
RQ3: Can the position of the question in the ranking indicate badly formulated questions?
RQ4: Some questions can be easier or more difficult for one group of students?
RQ5: Can the concentration made by IRT analysis be evaluated for identifying if a question is badly formulated, i.e., if a great population of students failed, or if it is a difficult question, i.e., if some students hit and other students fail.

In the case of RQ1 and following IRT, all questions are necessary if the assumptions of unidimensionality and local independence are fulfilled, although these questions can be classified according to the criteria previously mentioned in 4.3, from the Hashing and Complexity questionnaires were obtained:

Experiment 1:
Complexity (6 Questions): **Q1, Q3**
Hashing (11 Questions): **Q3, Q4, Q5, Q6**

For experiment 2, see table 7.

The questions mentioned above were classified by different criteria such as: the ranking of the amount of information, level of discrimination and random success, these classifications can be questioned in case of RQ2.

The position of the ranking cannot indicate directly questions with errors as asked in RQ3, otherwise the questions that did not comply with the criteria proposed by the methodology cannot be measured with the IRT and there is a need to reevaluate their elaborations, with a view to correcting problems such as cohesion, clarity of skill required, correction of alternatives or even layout and layout of the item in the test as proposed by De Andrade et al. (2000). Questions RQ4 and RQ5 can be addressed with the statistical analysis and the results with significant differences because the factors used were the difficulty level in the three logistic models for all periods.

7. CONCLUSION

The generation of an optimal model, ranked by questions from a course with a lower mean knowledge rate to higher ones, can be a useful tool for teachers. Knowing which questions are contributing more to the learning of their students can help teachers in the task of refining and adapt their questionnaires. The objective of this investigation was to propose a methodology to analyze questionnaires used in online courses, to detect the behaviors of students according to their answers that could serve as a decision tool of the variability of a question and if this affects, or not, the responses of the students.

A series of questions were formulated for responding to this objective and it was possible to conclude that a ranking and a classification of the questions can be established when this methodology was used, but it is important to know that during the creation of the questionnaire the principles of one-dimensionality and local independence postulated by the IRT are totally necessary. The statistical results showed that differences can be found between the questions of each questionnaire based on the level of difficulty between the different periods analyzed. In all the questionnaires it was possible to identify questions and they could be classified according to their discrimination and difficulty. These results can be an indicator of question`s variability and can be used as a decision tool for the modification of the questionnaires used and the order of them.

A limitation of the methodology is the manual interpretation that must be done of the graphic analyzes. Later, a virtual agent in charge of this task could be thought of being trained from the tables of results of the coefficients generated by IRT, automating the process.

As future work, we propose performing online tests, using the output of the methodology at the beginning of a class to have the feedback from the professor about the helpfulness of the methodology on both the questionnaires refinement and in the decrease of the mean error rate as a whole. Also, we believe that an ontology could be created for mapping the knowledge area of each questionnaire helping the professor to identify the areas their students are performing worse. It would also allow making comparisons between the evaluated groups, as suggested by Millán et al. (2013). These efforts would allow a better classification between the evaluated content and the questionnaires which would allow identifying more efficiently the most important questions.

REFERENCES

AMÓN, Jesús. **Estadística para psicólogos: Estadística descriptiva**. Probabilidad; estadística inferencial (534 p.). Pirámide, 1985.

ARAUJO, Eutalia Aparecida Candido de; ANDRADE, Dalton Francisco de; BORTOLOTTI, Silvana Ligia Vincenzi. **Item response theory**. Revista da Escola de Enfermagem da USP, 2009, vol. 43, no SPE, p. 1000-1008.

BAKER, Frank B.; KIM, Seock-Ho. **The basics of item response theory using R**. New York, NY: Springer, 2017.

BONG NA, Woon; MARSHALL, Roger; LANE KELLER, Kevin. **Measuring brand power: validating a model for optimizing brand equity**. Journal of product & brand management, 1999, vol. 8, no 3, p. 170-184.

BRAGION, M.L.L.de. **Análise combinada de exames vestibulares da Universidade Federal de Lavras usando a teoria de resposta ao item**. 2010. 187 p. Tese (Doutorado em Estatística e Experimentação Agropecuária),Universidade Federal de Lavras, Lavras, 2010.

BRUSILOVSKY, Peter. **User modeling and user-adapted interaction**. 2001.

CHALHOUB–DEVILLE, Micheline; DEVILLE, Craig. **Computer adaptive testing in second language contexts**. Annual Review of Applied Linguistics, 1999, vol. 19, p. 273-299.

CORBETT, Albert T.; ANDERSON, John R. **Knowledge tracing: Modeling the acquisition of procedural knowledge**. User modeling and user-adapted interaction, 1994, vol. 4, no 4, p. 253-278.

COOK, David A.; THOMPSON, Warren G.; THOMAS, Kris G. **Test-enhanced web-based learning: optimizing the number of questions (a randomized crossover trial)**. Academic Medicine, 2014, vol. 89, no 1, p. 169-175.

DA SILVA NUNES, Carlos Henrique Sancineto; PRIMI, Ricardo. **Impacto do tamanho da amostra na calibração de itens e estimativa de escores por teoria de resposta ao item**. Avaliaçao Psicologica: Interamerican Journal of Psychological Assessment, 2005, vol. 4, no 2, p. 141-153.

DE ANDRADE, Dalton Francisco; TAVARES, Heliton Ribeiro; DA CUNHA VALLE, Raquel. **Teoria da Resposta ao Item: conceitos e aplicações**. ABE, Sao Paulo, 2000.

DREYFUS, Stuart E.; DREYFUS, Hubert L. **A five-stage model of the mental activities involved in directed skill acquisition**. California Univ Berkeley Operations Research Center, 1980.

EKANADHAM, Chaitanya; KARKLIN, Yan. **T-skirt: Online estimation of student proficiency in an adaptive learning system**. ArXiv preprint arXiv:1702.04282, 2017.

FERNÁNDEZ, José Muñiz; HAMBLETON, Ronald K. **Medio siglo de teoría de respuesta a los ítems**. Anuario de psicología/The UB Journal of psychology, 1992, no 52, p. 41-66.

FONSECA, João José Saraiva. **Metodologia da Pesquisa Científica**. 2002.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. São Paulo, 2002, vol. 5, no 61, p. 16-17.

GUIDE, A. Short. **Is This a Is This a Is This a Is This a Trick Question?**. 2001. Available at: <https://ksde.org/Portals/0/CSAS/CSAS%20Home/CTE%20Home/Instructor_Resources/TrickQuestion.pdf>. Access 01 Dec. 2018.

HAMBLETON, Ronald K.; SWAMINATHAN, Hariharan; ROGERS, H. Jane. **Fundamentals of item response theory**. Sage, 1991.

HUANG, Jiankun; WU, Wenjun. **T-BMIRT: Estimating representations of student knowledge and educational components in online education**. En 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017. p. 1301-1306.

JATOBÁ, Victor, et al. **Testes Adaptativos Computadorizados baseados na Teoria de Resposta ao Item em Sistemas e-learning: Uma revisão sistemática da literatura**. En Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2017. p. 273.

KOSKELA, Marileena, et al. **Suitability of a Virtual Learning Environment for Higher Education. Electronic Journal of e-Learning**. 2005, vol. 3, no 1, p. 23-32.

LORD, Frederic M. **On the Statistical Treatment of Football Numbers**. 1953.

LORD, Frederic M. **Item characteristic curves estimated without knowledge of their mathematical form—a confrontation of Birnbaum's logistic model**. Psychometrika, 1970, vol. 35, no 1, p. 43-50.

MILLÁN, Eva, et al. **Using Bayesian networks to improve knowledge assessment**. Computers & Education, 2013, vol. 60, no 1, p. 436-447.

NOVICK, Melvin R. **The axioms and principal results of classical test theory**. Journal of mathematical psychology, 1966, vol. 3, no 1, p. 1-18.

PASQUALI, L.; PRIMI, R. **Basic theory of Item Response Theory (IRT)**. Avaliação Psicol, 2003, vol. 2, no 2, p. 99-110.

PONSODA, V., et al. **Teoría de la Respuesta al Ítem. Psicometría I**. Facultad de Psicología, UAM. Madrid: España. Ediciones de la Universidad Autónoma de Madrid, 1998, p. 1-23.

PONSODA, Vicente; OLEA, Julio; REVUELTA, Javier. **ADTEST: A computer-adaptive test based on the maximum information principle**. Educational and Psychological Measurement, 1994, vol. 54, no 3, p. 680-686.

RENOM, J. **Test adaptativos computarizados (TAC)**. En Aportaciones recientes a la evaluación psicológica. 1993. p. 71-94.

RIZOPOULOS, Dimitris. **ltm: An R package for latent variable modeling and item response theory analyses**. Journal of statistical software, 2006, vol. 17, no 5, p. 1-25.

SANTOS, Fabrícia Damando; DE REZENDE GUEDES, Leonardo Guerra. **Testes Adaptativos Informatizados baseados em teoria de resposta ao item utilizados em ambientes virtuais de aprendizagem**. RENOTE, 2005, vol. 3, no 2.

SPEARMAN, Charles. **"General Intelligence", objectively determined and measured**. The American Journal of Psychology, 1904, vol. 15, no 2, p. 201-292.

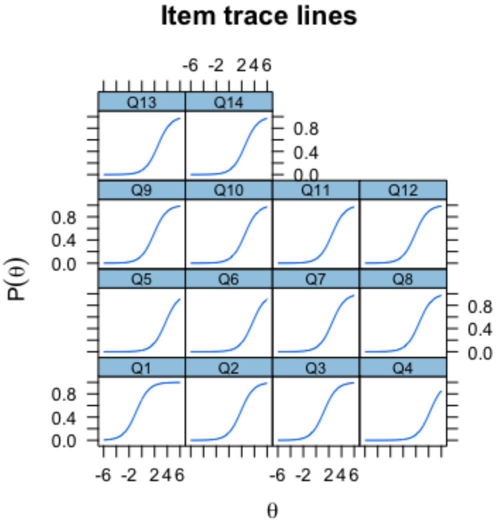
TIAN, Xinyun, et al. **Applying Item Response Theory to Analyzing and Improving the Item Quality of an Online Chinese Reading Assessment**. En 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). IEEE, 2017. p. 754-759.

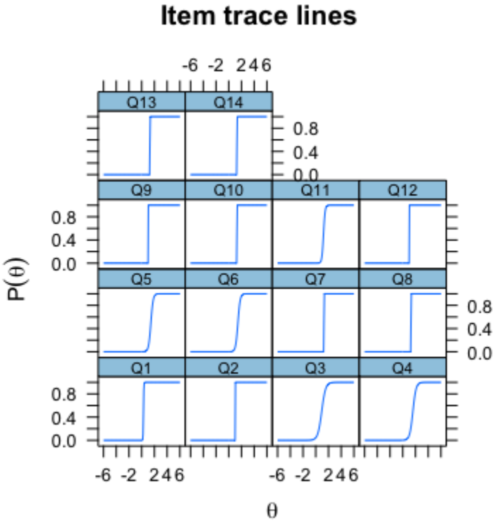
VENDRAMINI, C. M. M. **Aplicação da Teoria de Resposta ao Item na avaliação educacional. Temas em avaliação psicológica**. 2002, vol. 1, p. 116-130.

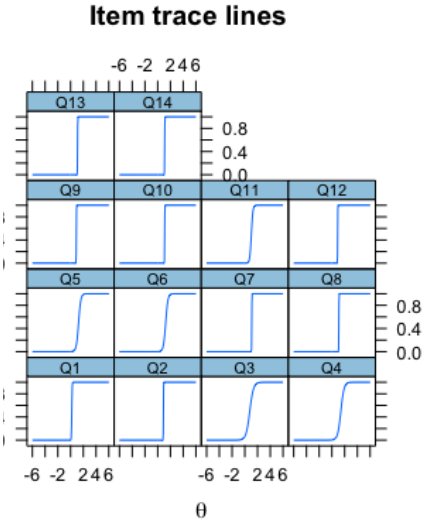
APPENDIX A <Details for Results of Experiment 2: "Electrical Engineering" course>

Experiment 2: "Electrical Engineering" course
Graphical Analysis of IRT for Questionnaires from the "Electrical Engineering" Course
Detail information was described in the following tables (**Tables A to L**).
Graphical IRT Analysis: Questionnaire 1 - Kirchoff

Period: 2016-2

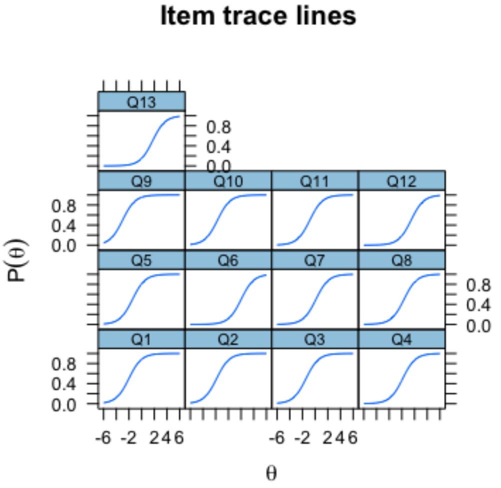


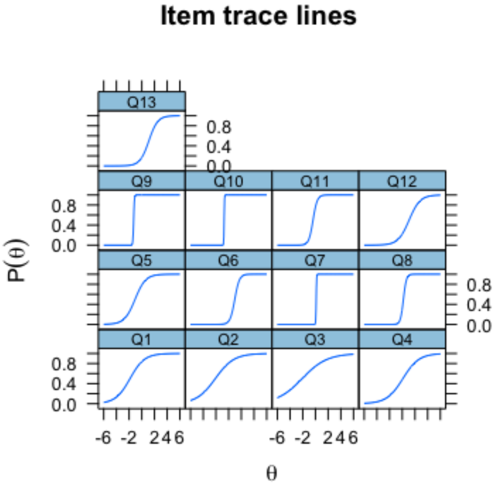


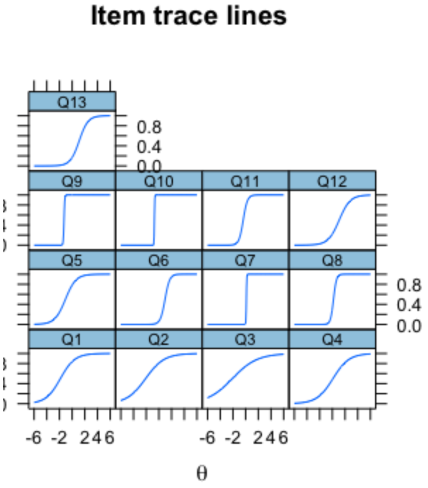


Rasch model 2PL
3PL

Period: 2017-1

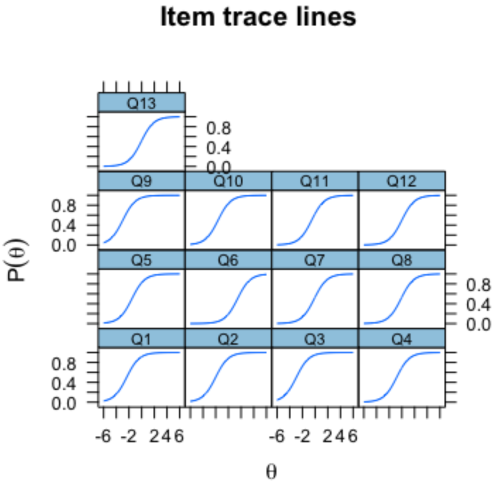


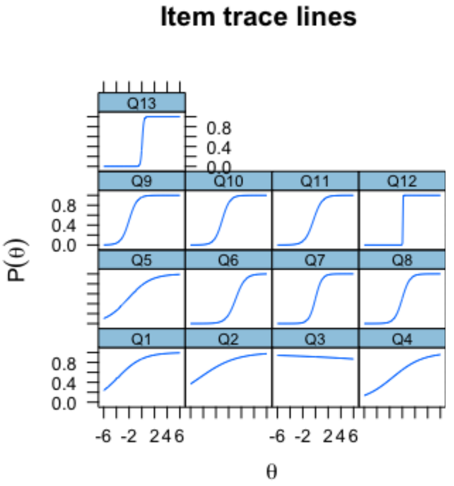


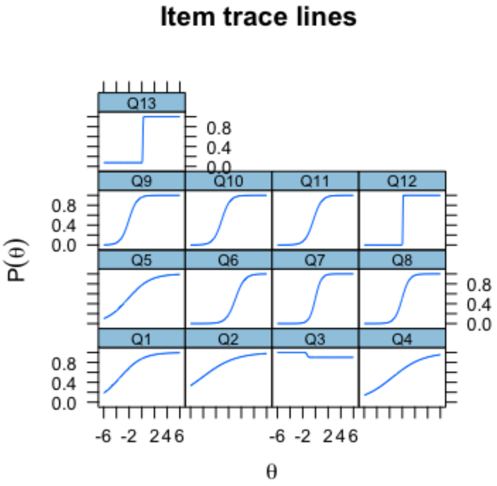


Rasch model 2PL
3PL

Period: 2017-2



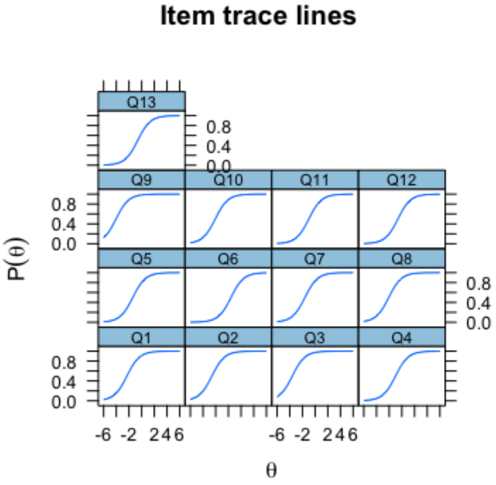


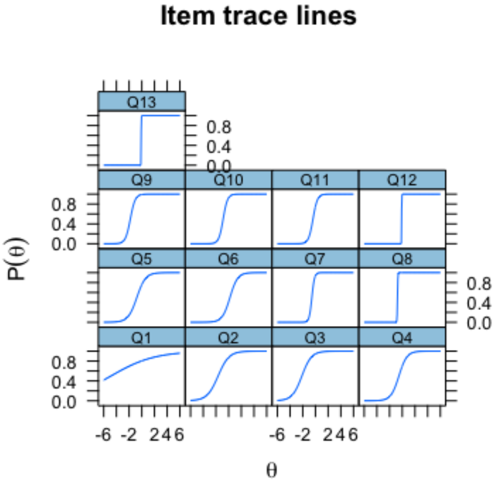


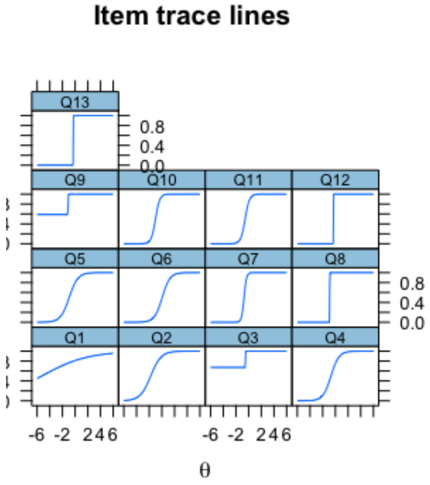
Rasch model
3PL

2PL

Period: 2018-1



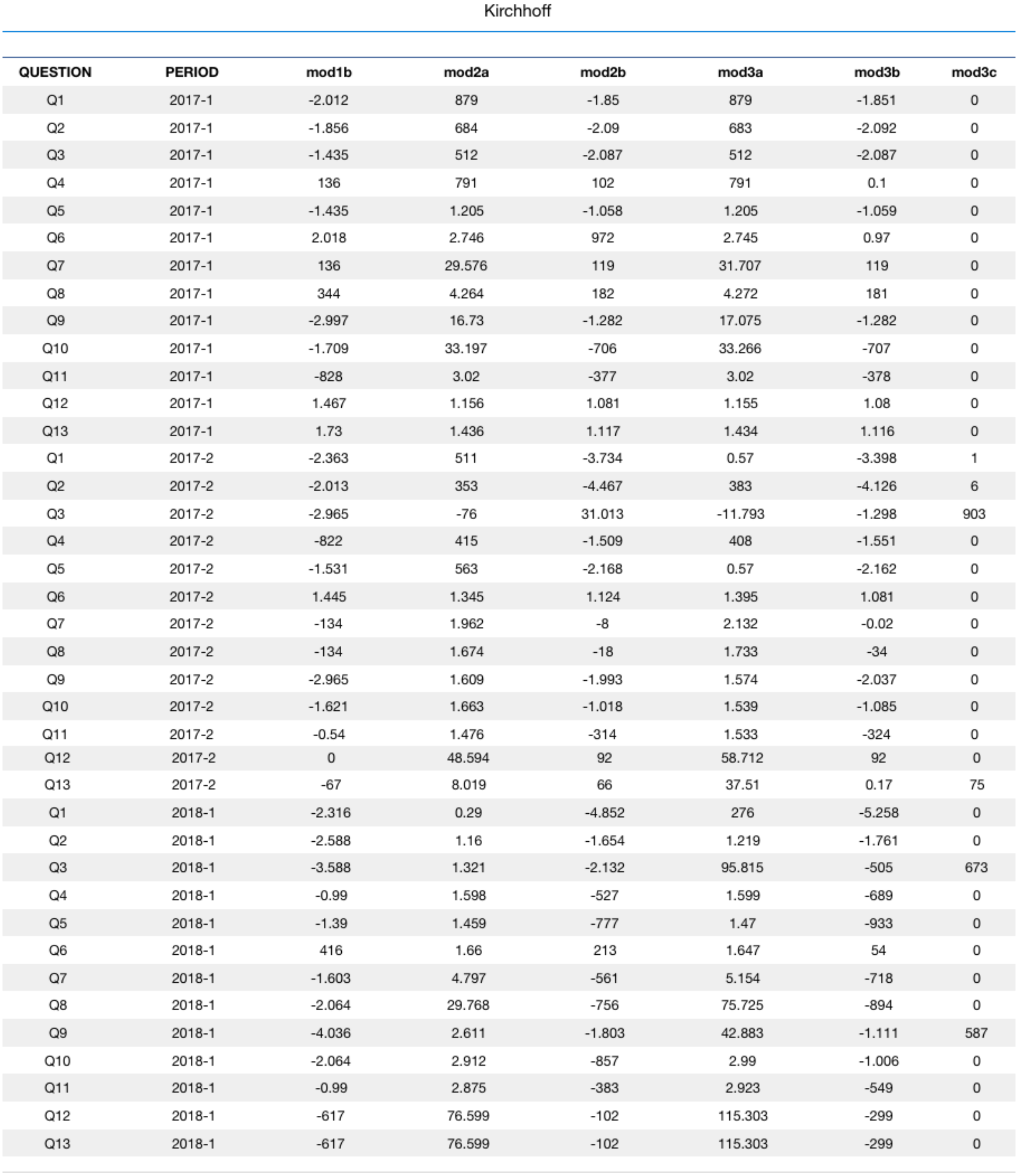




Rasch model
3PL

2PL

Table A: Coefficients of Logistic Models for kirchoff questionnaire

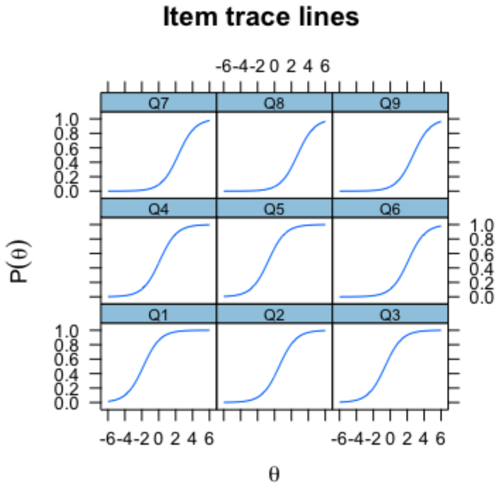


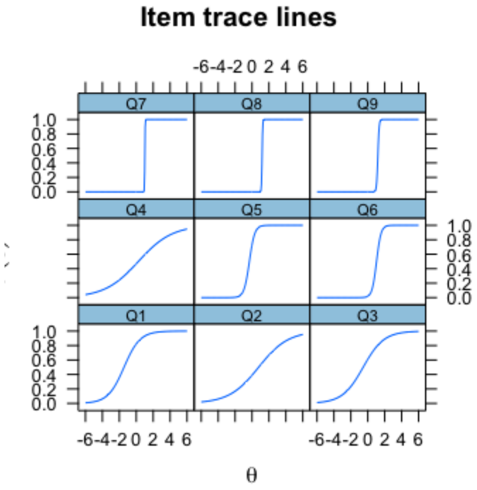
Kirchhoff

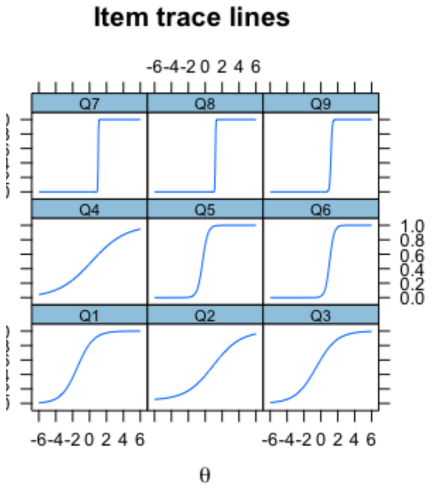
QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2017-1	-2.012	879	-1.85	879	-1.851	0
Q2	2017-1	-1.856	684	-2.09	683	-2.092	0
Q3	2017-1	-1.435	512	-2.087	512	-2.087	0
Q4	2017-1	136	791	102	791	0.1	0
Q5	2017-1	-1.435	1.205	-1.058	1.205	-1.059	0
Q6	2017-1	2.018	2.746	972	2.745	0.97	0
Q7	2017-1	136	29.576	119	31.707	119	0
Q8	2017-1	344	4.264	182	4.272	181	0
Q9	2017-1	-2.997	16.73	-1.282	17.075	-1.282	0
Q10	2017-1	-1.709	33.197	-706	33.266	-707	0
Q11	2017-1	-828	3.02	-377	3.02	-378	0
Q12	2017-1	1.467	1.156	1.081	1.155	1.08	0
Q13	2017-1	1.73	1.436	1.117	1.434	1.116	0
Q1	2017-2	-2.363	511	-3.734	0.57	-3.398	1
Q2	2017-2	-2.013	353	-4.467	383	-4.126	6
Q3	2017-2	-2.965	-76	31.013	-11.793	-1.298	903
Q4	2017-2	-822	415	-1.509	408	-1.551	0
Q5	2017-2	-1.531	563	-2.168	0.57	-2.162	0
Q6	2017-2	1.445	1.345	1.124	1.395	1.081	0
Q7	2017-2	-134	1.962	-8	2.132	-0.02	0
Q8	2017-2	-134	1.674	-18	1.733	-34	0
Q9	2017-2	-2.965	1.609	-1.993	1.574	-2.037	0
Q10	2017-2	-1.621	1.663	-1.018	1.539	-1.085	0
Q11	2017-2	-0.54	1.476	-314	1.533	-324	0
Q12	2017-2	0	48.594	92	58.712	92	0
Q13	2017-2	-67	8.019	66	37.51	0.17	75
Q1	2018-1	-2.316	0.29	-4.852	276	-5.258	0
Q2	2018-1	-2.588	1.16	-1.654	1.219	-1.761	0
Q3	2018-1	-3.588	1.321	-2.132	95.815	-505	673
Q4	2018-1	-0.99	1.598	-527	1.599	-689	0
Q5	2018-1	-1.39	1.459	-777	1.47	-933	0
Q6	2018-1	416	1.66	213	1.647	54	0
Q7	2018-1	-1.603	4.797	-561	5.154	-718	0
Q8	2018-1	-2.064	29.768	-756	75.725	-894	0
Q9	2018-1	-4.036	2.611	-1.803	42.883	-1.111	587
Q10	2018-1	-2.064	2.912	-857	2.99	-1.006	0
Q11	2018-1	-0.99	2.875	-383	2.923	-549	0
Q12	2018-1	-617	76.599	-102	115.303	-299	0
Q13	2018-1	-617	76.599	-102	115.303	-299	0

Graphical IRT Analysis: Questionnaire 2 - Electrical Installations Concepts

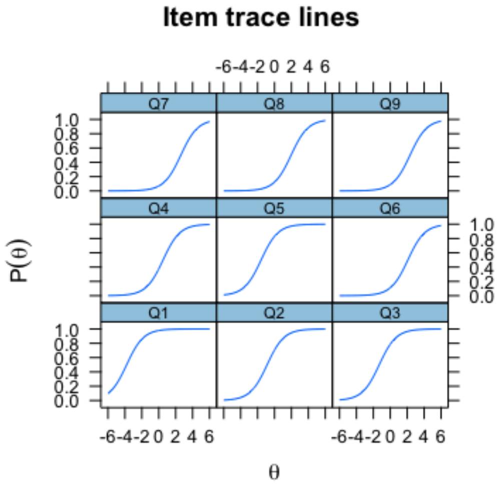
Period: 2017-1

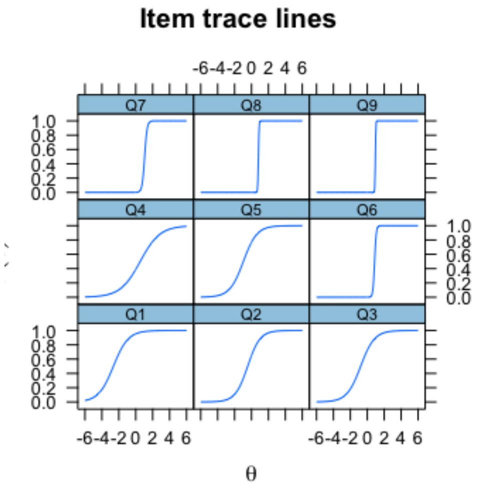


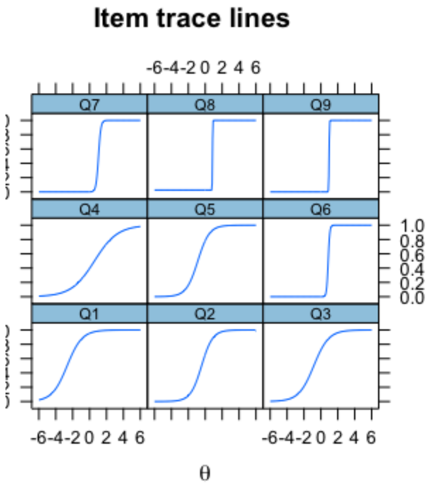




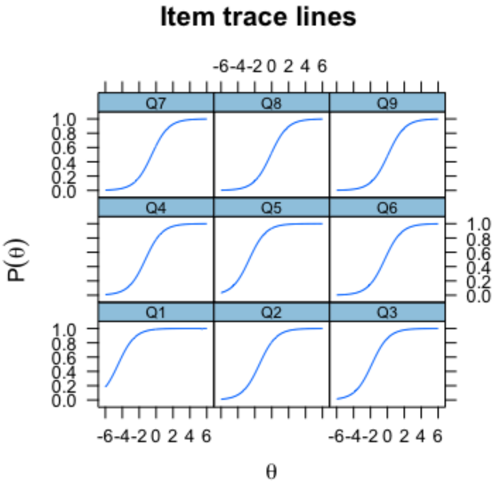
Period: 2017-2

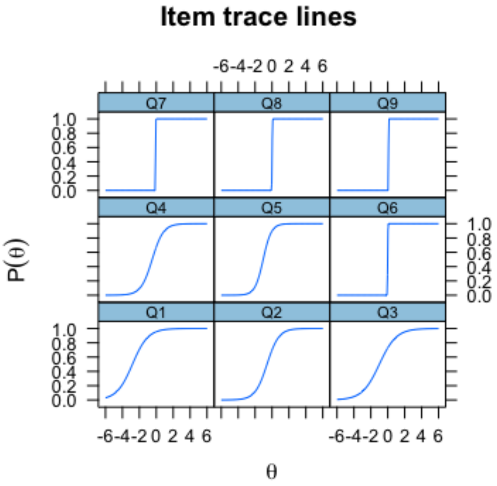






Period: 2018-1





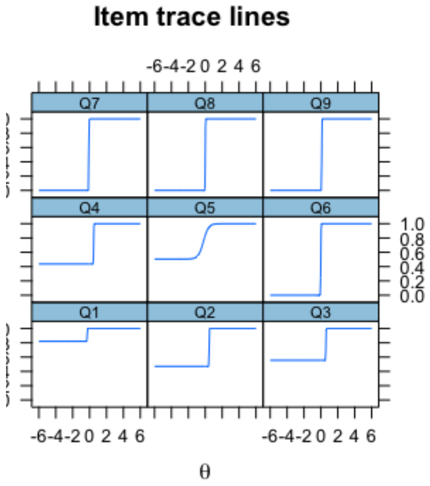
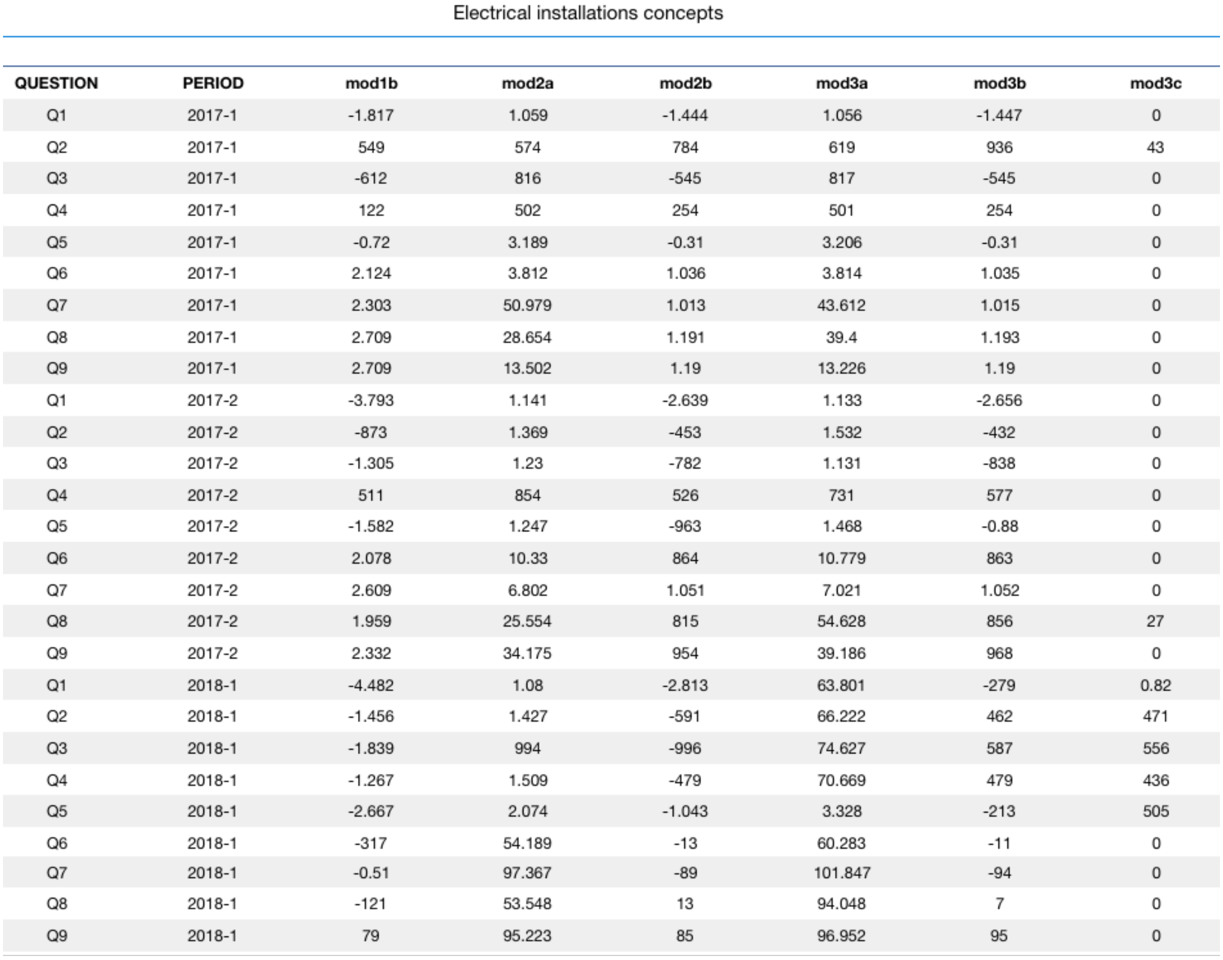


Table B: Coefficients of Logistic Models for Electrical Installations Concepts questionnaire

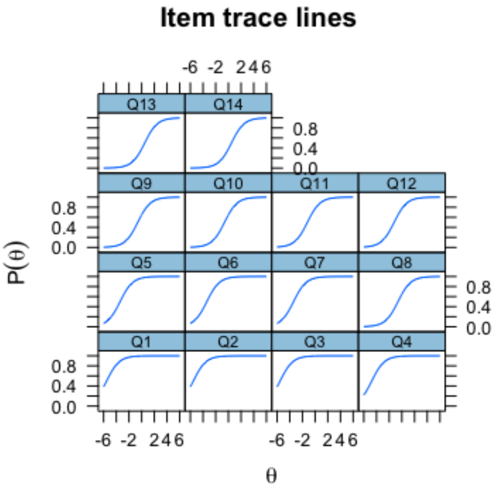


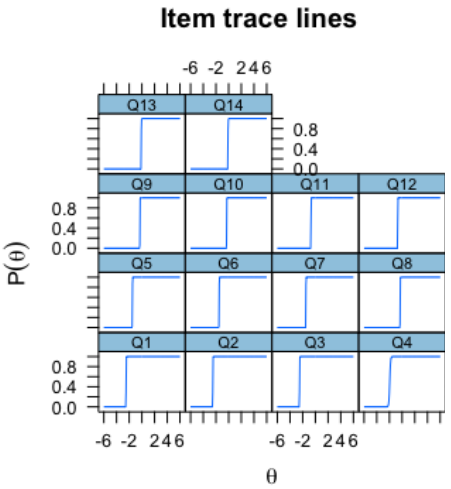
Electrical installations concepts

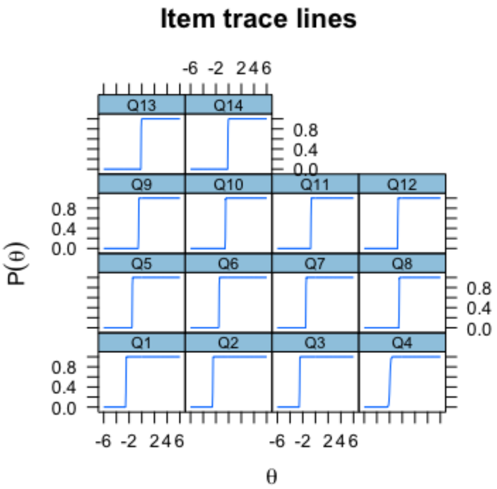
QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2017-1	-1.817	1.059	-1.444	1.056	-1.447	0
Q2	2017-1	549	574	784	619	936	43
Q3	2017-1	-612	816	-545	817	-545	0
Q4	2017-1	122	502	254	501	254	0
Q5	2017-1	-0.72	3.189	-0.31	3.206	-0.31	0
Q6	2017-1	2.124	3.812	1.036	3.814	1.035	0
Q7	2017-1	2.303	50.979	1.013	43.612	1.015	0
Q8	2017-1	2.709	28.654	1.191	39.4	1.193	0
Q9	2017-1	2.709	13.502	1.19	13.226	1.19	0
Q1	2017-2	-3.793	1.141	-2.639	1.133	-2.656	0
Q2	2017-2	-873	1.369	-453	1.532	-432	0
Q3	2017-2	-1.305	1.23	-782	1.131	-838	0
Q4	2017-2	511	854	526	731	577	0
Q5	2017-2	-1.582	1.247	-963	1.468	-0.88	0
Q6	2017-2	2.078	10.33	864	10.779	863	0
Q7	2017-2	2.609	6.802	1.051	7.021	1.052	0
Q8	2017-2	1.959	25.554	815	54.628	856	27
Q9	2017-2	2.332	34.175	954	39.186	968	0
Q1	2018-1	-4.482	1.08	-2.813	63.801	-279	0.82
Q2	2018-1	-1.456	1.427	-591	66.222	462	471
Q3	2018-1	-1.839	994	-996	74.627	587	556
Q4	2018-1	-1.267	1.509	-479	70.669	479	436
Q5	2018-1	-2.667	2.074	-1.043	3.328	-213	505
Q6	2018-1	-317	54.189	-13	60.283	-11	0
Q7	2018-1	-0.51	97.367	-89	101.847	-94	0
Q8	2018-1	-121	53.548	13	94.048	7	0
Q9	2018-1	79	95.223	85	96.952	95	0

Graphical IRT Analysis: Questionnaire 3 - Resistive circuit resolution

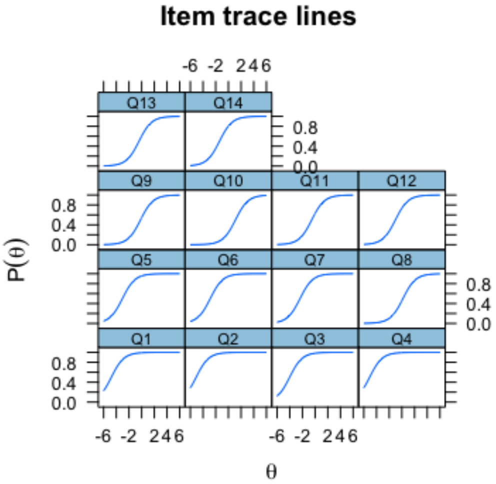
Period: 2017-1

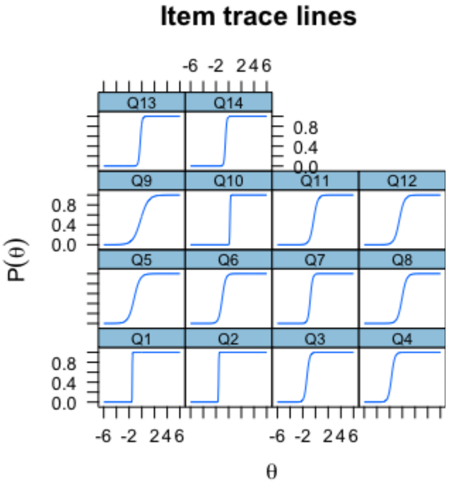


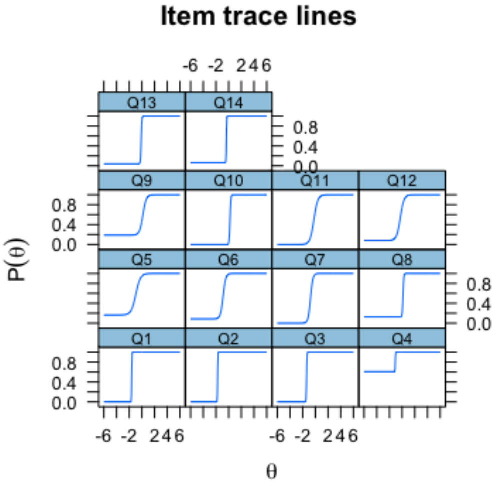




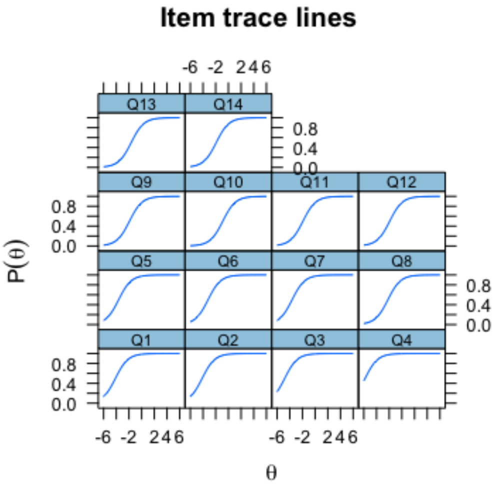
Period: 2017-2

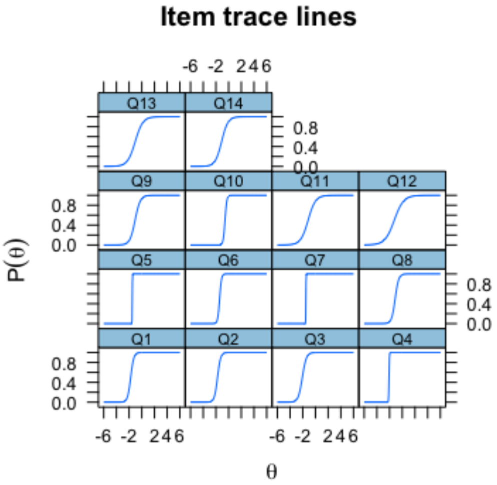






Period: 2018-1





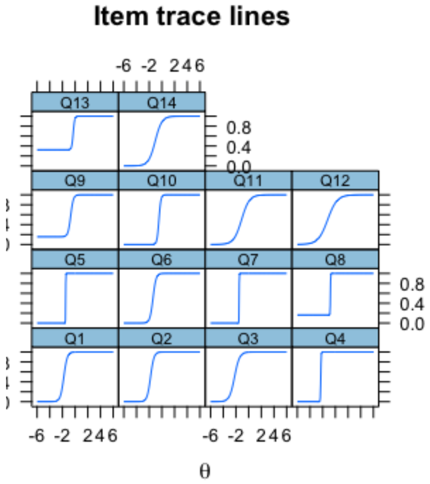
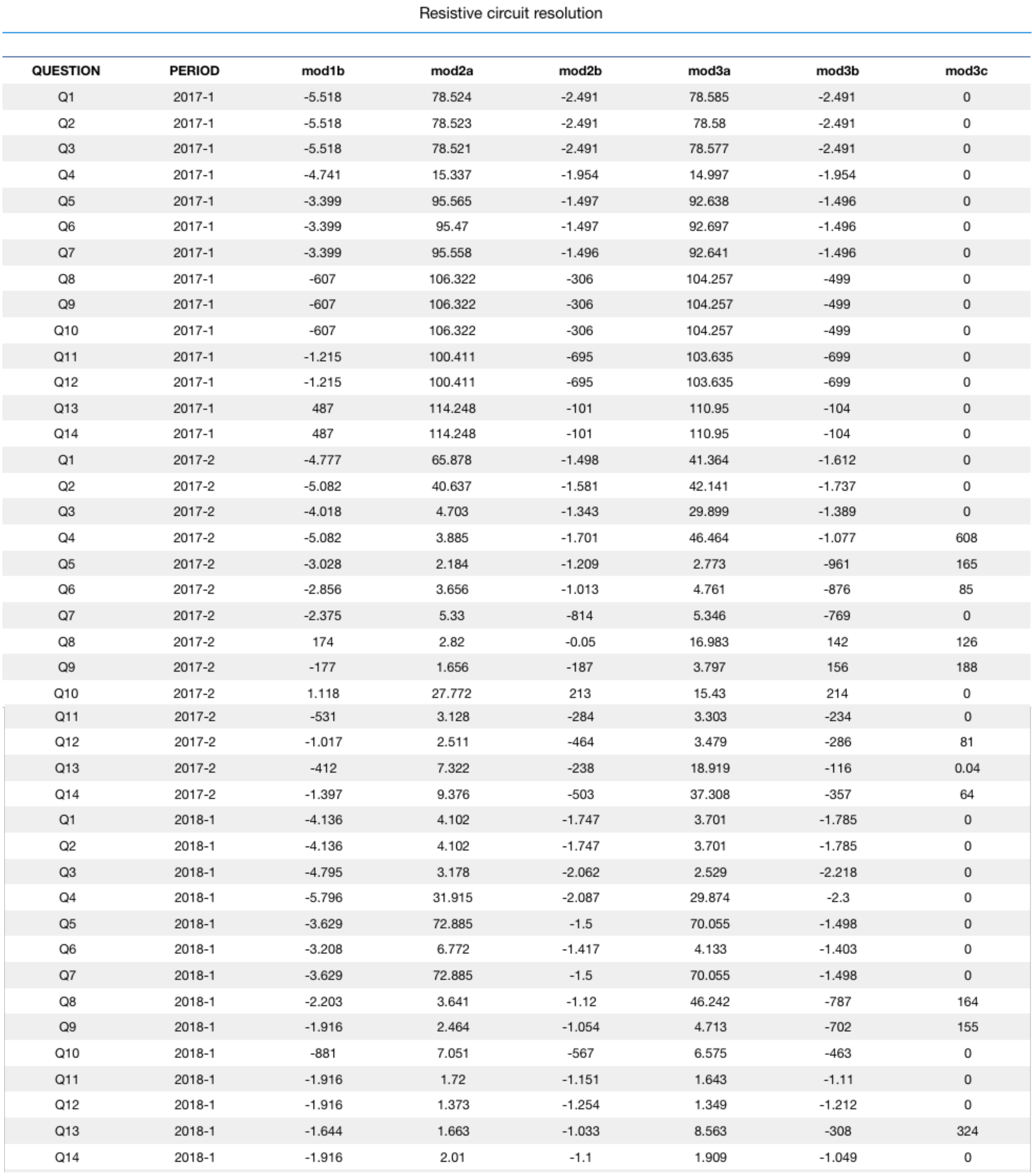


Table C: Coefficients of Logistic Models for Resistive Circuit Resolution questionnaire

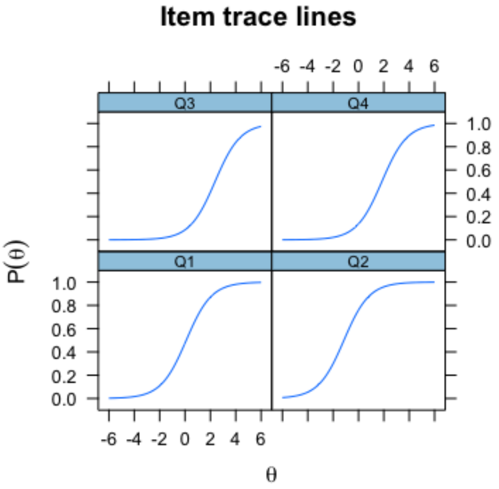


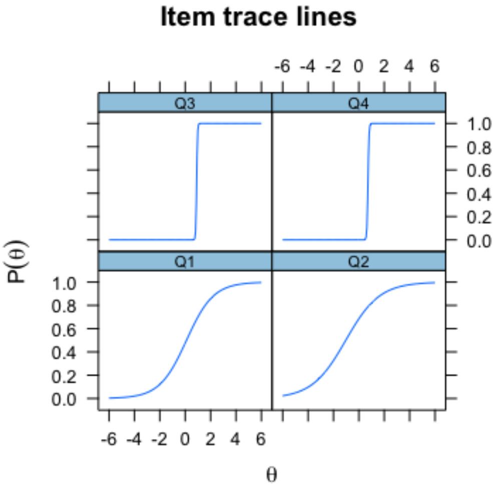
Resistive circuit resolution

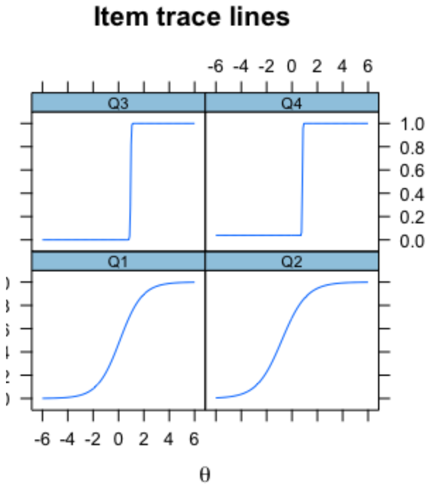
QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2017-1	-5.518	78.524	-2.491	78.585	-2.491	0
Q2	2017-1	-5.518	78.523	-2.491	78.58	-2.491	0
Q3	2017-1	-5.518	78.521	-2.491	78.577	-2.491	0
Q4	2017-1	-4.741	15.337	-1.954	14.997	-1.954	0
Q5	2017-1	-3.399	95.565	-1.497	92.638	-1.496	0
Q6	2017-1	-3.399	95.47	-1.497	92.697	-1.496	0
Q7	2017-1	-3.399	95.558	-1.496	92.641	-1.496	0
Q8	2017-1	-607	106.322	-306	104.257	-499	0
Q9	2017-1	-607	106.322	-306	104.257	-499	0
Q10	2017-1	-607	106.322	-306	104.257	-499	0
Q11	2017-1	-1.215	100.411	-695	103.635	-699	0
Q12	2017-1	-1.215	100.411	-695	103.635	-699	0
Q13	2017-1	487	114.248	-101	110.95	-104	0
Q14	2017-1	487	114.248	-101	110.95	-104	0
Q1	2017-2	-4.777	65.878	-1.498	41.364	-1.612	0
Q2	2017-2	-5.082	40.637	-1.581	42.141	-1.737	0
Q3	2017-2	-4.018	4.703	-1.343	29.899	-1.389	0
Q4	2017-2	-5.082	3.885	-1.701	46.464	-1.077	608
Q5	2017-2	-3.028	2.184	-1.209	2.773	-961	165
Q6	2017-2	-2.856	3.656	-1.013	4.761	-876	85
Q7	2017-2	-2.375	5.33	-814	5.346	-769	0
Q8	2017-2	174	2.82	-0.05	16.983	142	126
Q9	2017-2	-177	1.656	-187	3.797	156	188
Q10	2017-2	1.118	27.772	213	15.43	214	0
Q11	2017-2	-531	3.128	-284	3.303	-234	0
Q12	2017-2	-1.017	2.511	-464	3.479	-286	81
Q13	2017-2	-412	7.322	-238	18.919	-116	0.04
Q14	2017-2	-1.397	9.376	-503	37.308	-357	64
Q1	2018-1	-4.136	4.102	-1.747	3.701	-1.785	0
Q2	2018-1	-4.136	4.102	-1.747	3.701	-1.785	0
Q3	2018-1	-4.795	3.178	-2.062	2.529	-2.218	0
Q4	2018-1	-5.796	31.915	-2.087	29.874	-2.3	0
Q5	2018-1	-3.629	72.885	-1.5	70.055	-1.498	0
Q6	2018-1	-3.208	6.772	-1.417	4.133	-1.403	0
Q7	2018-1	-3.629	72.885	-1.5	70.055	-1.498	0
Q8	2018-1	-2.203	3.641	-1.12	46.242	-787	164
Q9	2018-1	-1.916	2.464	-1.054	4.713	-702	155
Q10	2018-1	-881	7.051	-567	6.575	-463	0
Q11	2018-1	-1.916	1.72	-1.151	1.643	-1.11	0
Q12	2018-1	-1.916	1.373	-1.254	1.349	-1.212	0
Q13	2018-1	-1.644	1.663	-1.033	8.563	-308	324
Q14	2018-1	-1.916	2.01	-1.1	1.909	-1.049	0

Graphical IRT Analysis: Questionnaire 4 - Systemic Circuit Resolution

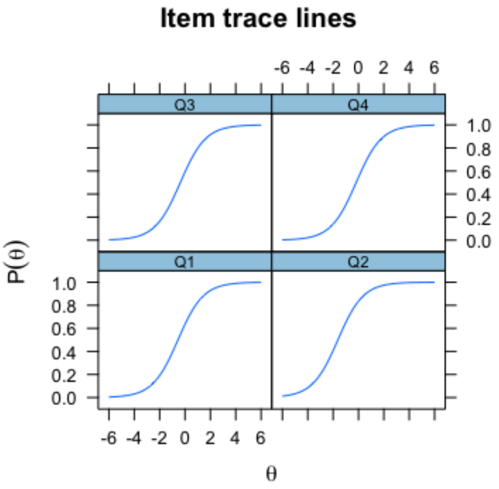
Period: 2017-1

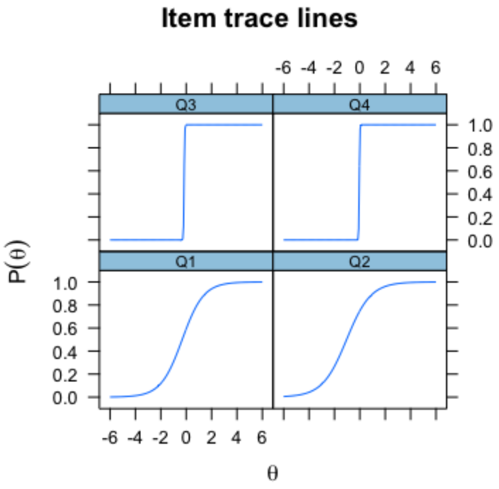


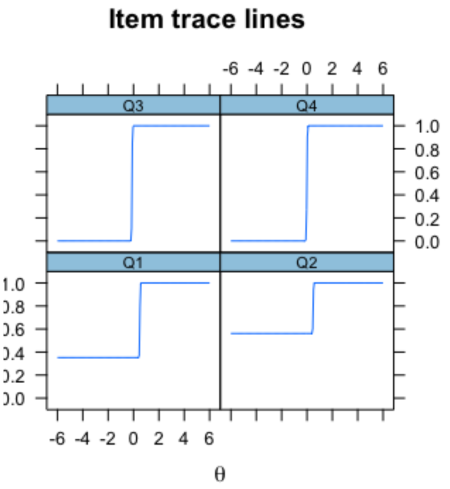




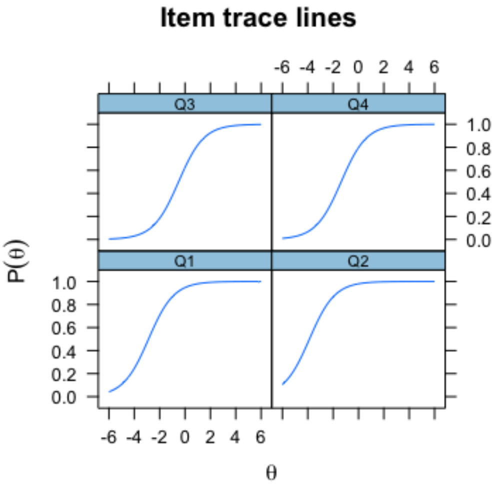
Period: 2017-2

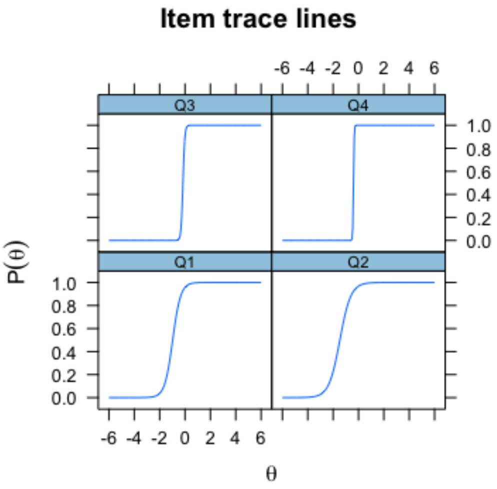






Period: 2018-1





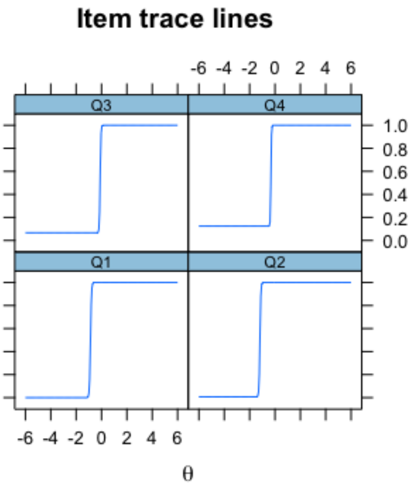


Table D: Coefficients of Logistic Models for Systemic Circuit Resolution questionnaire

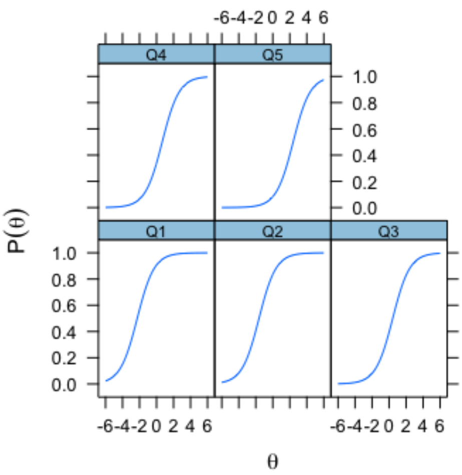
Systematic circuit resolution

QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2017-1	97	942	102	1.133	109	0
Q2	2017-1	-1.149	0.75	-1.031	1.021	-802	0
Q3	2017-1	2.367	29.729	902	40.652	959	0
Q4	2017-1	1.843	24.79	719	54.95	819	39
Q1	2017-2	-555	1.187	-314	68.753	501	352
Q2	2017-2	-1.624	1.052	-1.042	68.317	485	562
Q3	2017-2	-369	44.085	-166	63.51	-118	0
Q4	2017-2	-184	46.321	-43	53.91	-12	0
Q1	2018-1	-2.875	3.175	-973	25.754	-0.87	0
Q2	2018-1	-3.867	2.313	-1.469	28.139	-1.186	7
Q3	2018-1	-0.51	13.75	-163	32.728	-107	67
Q4	2018-1	-1.367	35.632	-394	44.918	-322	124

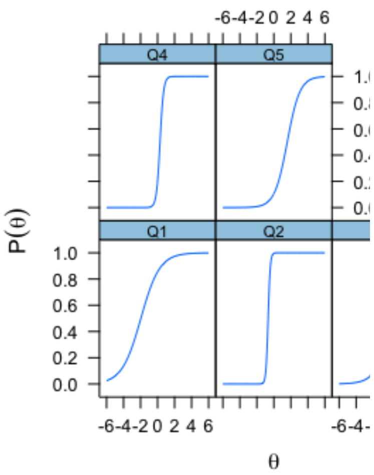
Graphical IRT Analysis: Questionnaire 5 - Alternating Current

Period: 2016-2

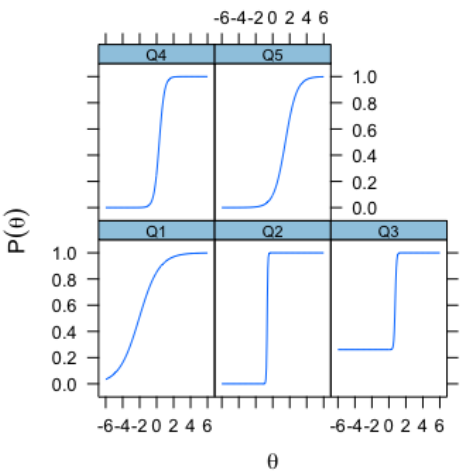
Item trace lines



Item trace lines

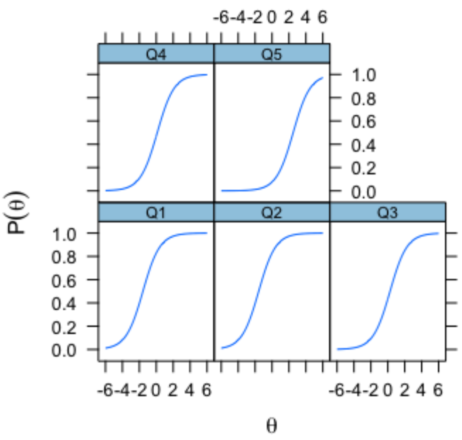


Item trace lines

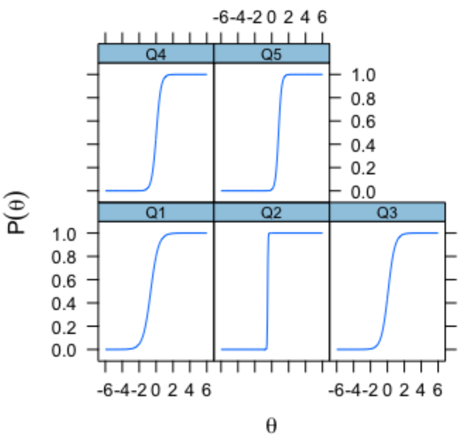


Period: 2017-1

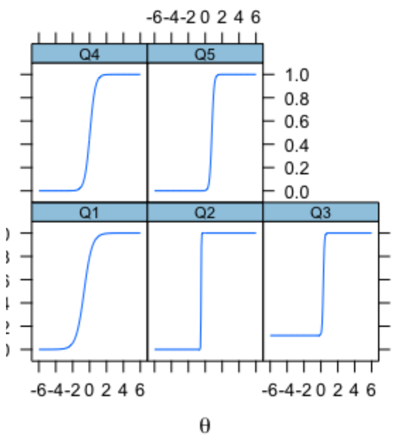
Item trace lines



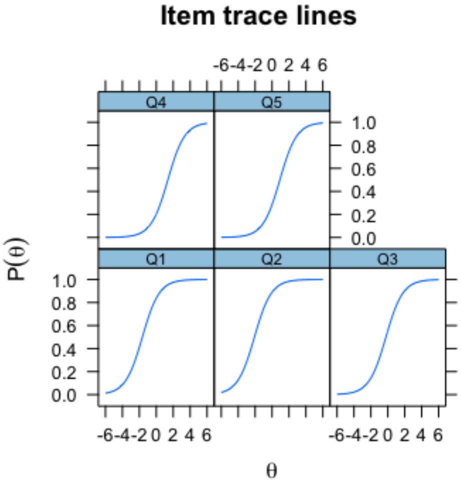
Item trace lines

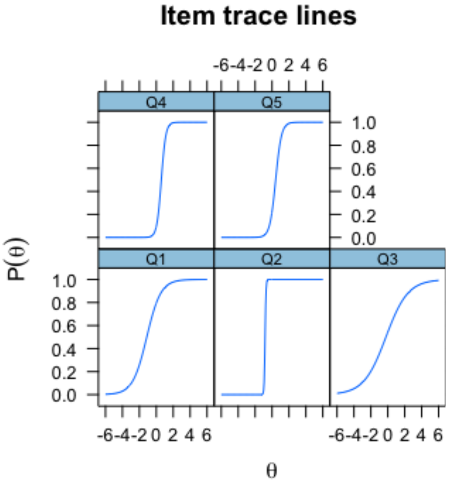


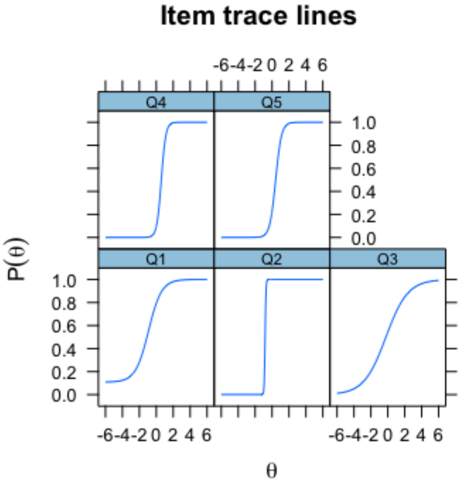
Item trace lines



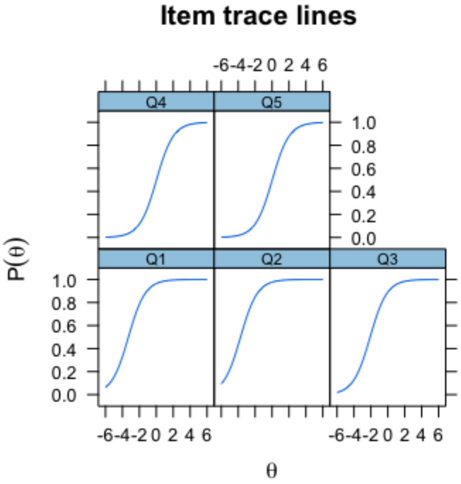
Period: 2017-2

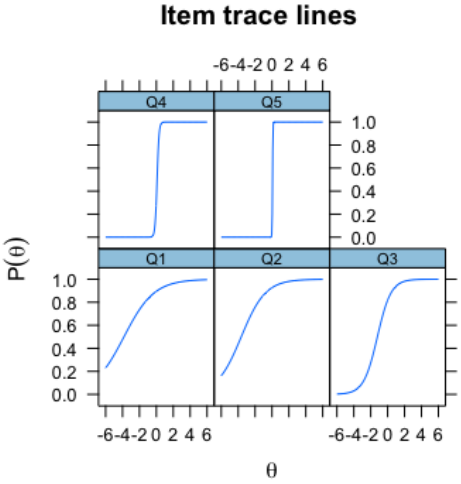






Period: 2018-1





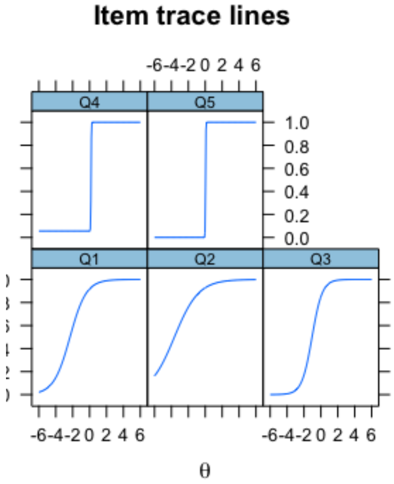
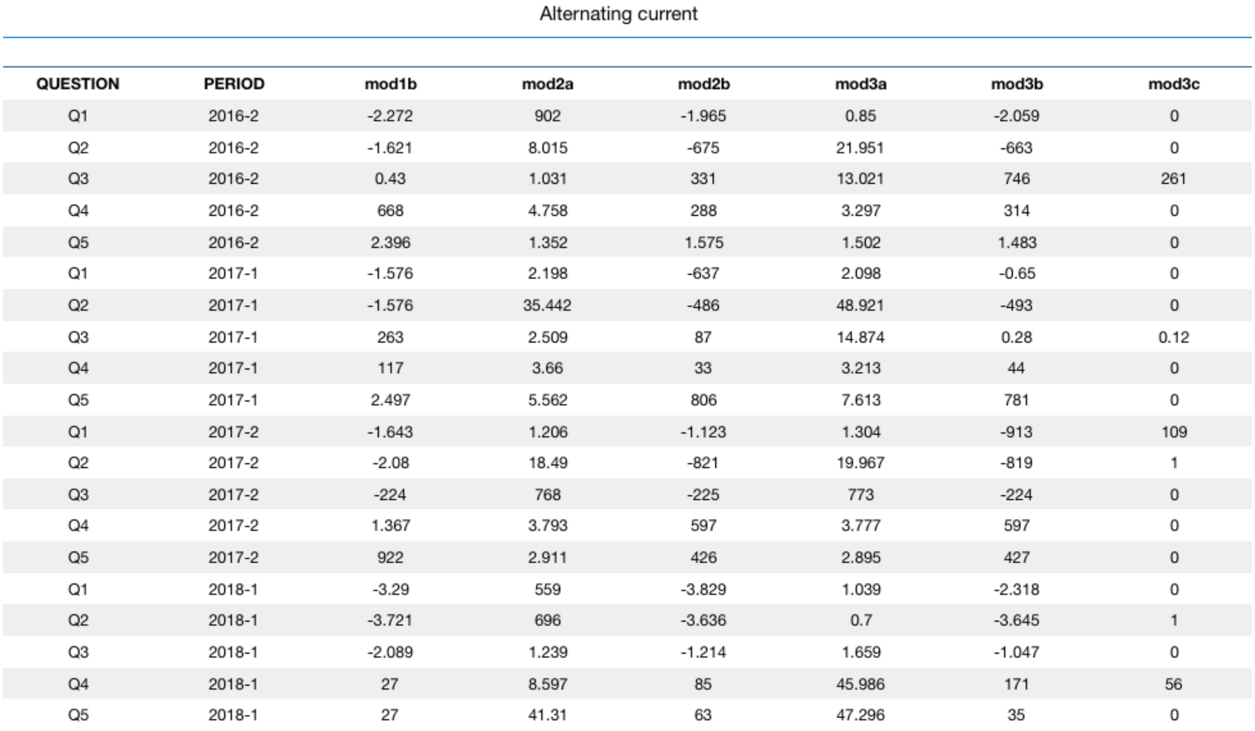


Table E: Coefficients of Logistic Models for Alternating current questionnaire

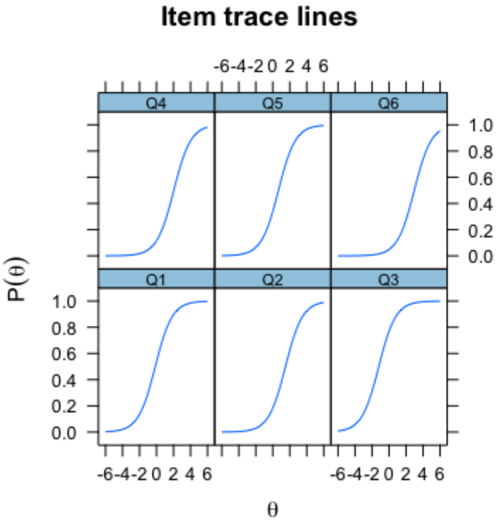


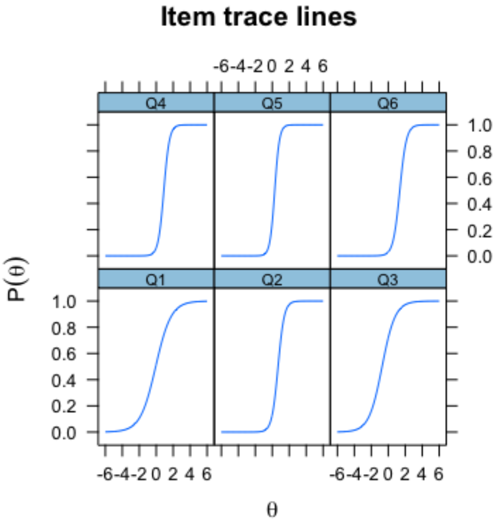
Alternating current

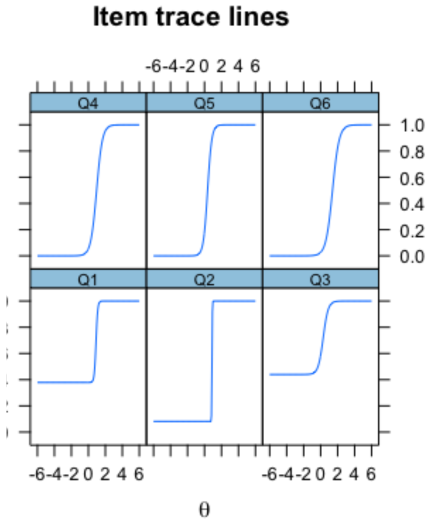
QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2016-2	-2.272	902	-1.965	0.85	-2.059	0
Q2	2016-2	-1.621	8.015	-675	21.951	-663	0
Q3	2016-2	0.43	1.031	331	13.021	746	261
Q4	2016-2	668	4.758	288	3.297	314	0
Q5	2016-2	2.396	1.352	1.575	1.502	1.483	0
Q1	2017-1	-1.576	2.198	-637	2.098	-0.65	0
Q2	2017-1	-1.576	35.442	-486	48.921	-493	0
Q3	2017-1	263	2.509	87	14.874	0.28	0.12
Q4	2017-1	117	3.66	33	3.213	44	0
Q5	2017-1	2.497	5.562	806	7.613	781	0
Q1	2017-2	-1.643	1.206	-1.123	1.304	-913	109
Q2	2017-2	-2.08	18.49	-821	19.967	-819	1
Q3	2017-2	-224	768	-225	773	-224	0
Q4	2017-2	1.367	3.793	597	3.777	597	0
Q5	2017-2	922	2.911	426	2.895	427	0
Q1	2018-1	-3.29	559	-3.829	1.039	-2.318	0
Q2	2018-1	-3.721	696	-3.636	0.7	-3.645	1
Q3	2018-1	-2.089	1.239	-1.214	1.659	-1.047	0
Q4	2018-1	27	8.597	85	45.986	171	56
Q5	2018-1	27	41.31	63	47.296	35	0

Graphical IRT Analysis: Questionnaire 6 - Effective Value

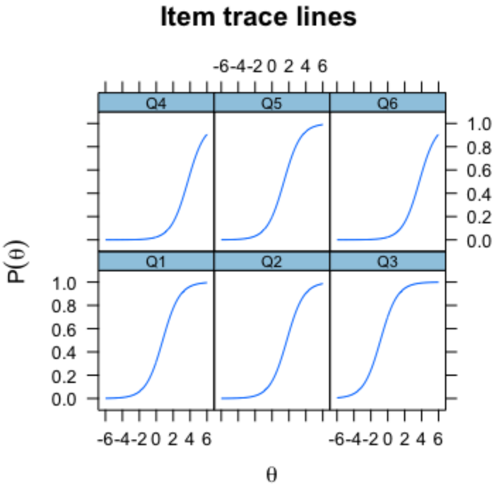
Period: 2016-2

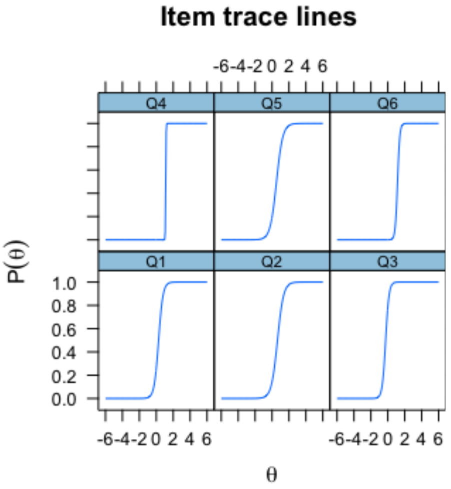


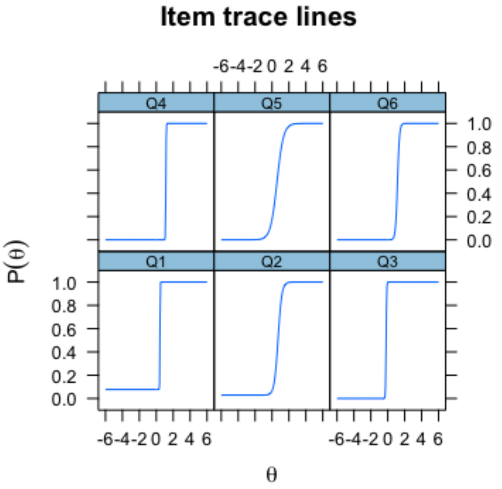




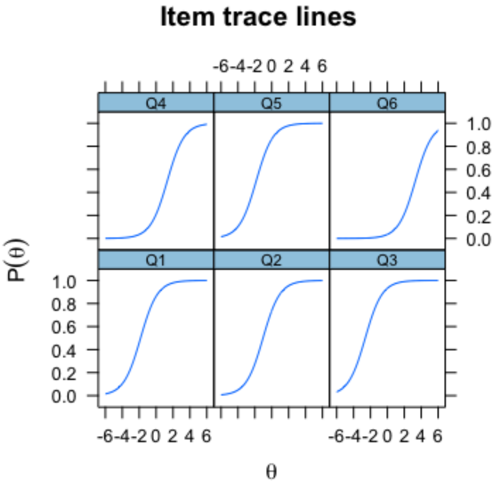
Period: 2017-1

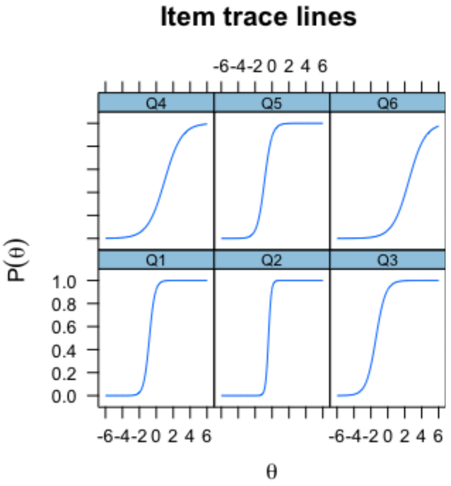


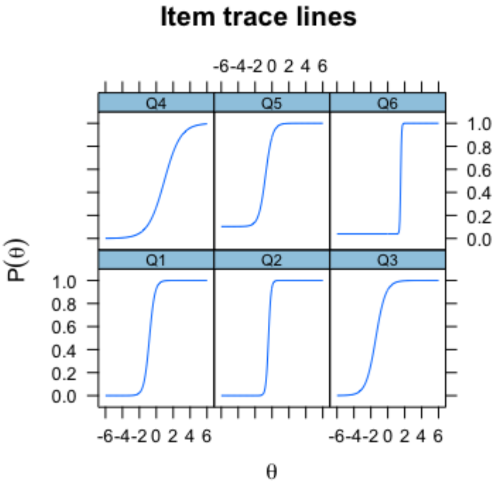




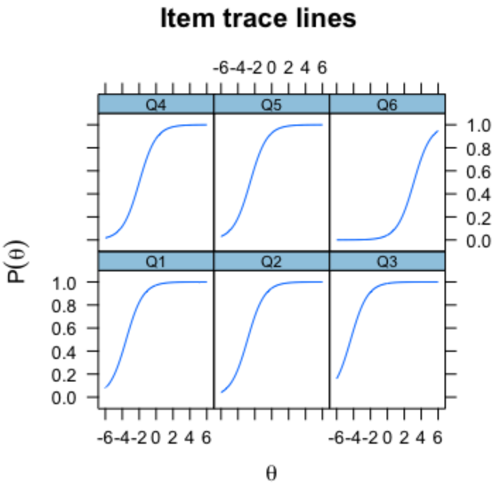
Period: 2017-2

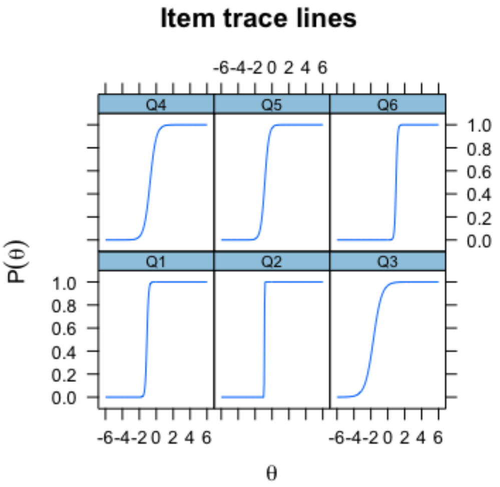






Period: 2018-1





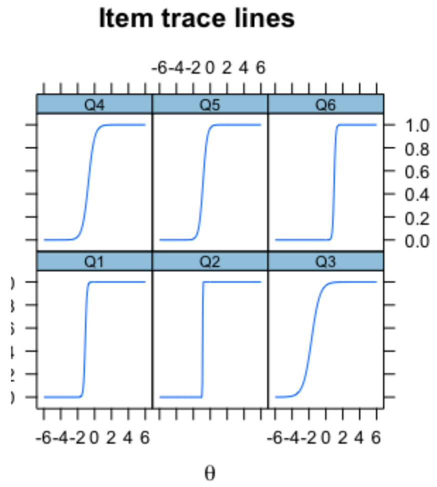
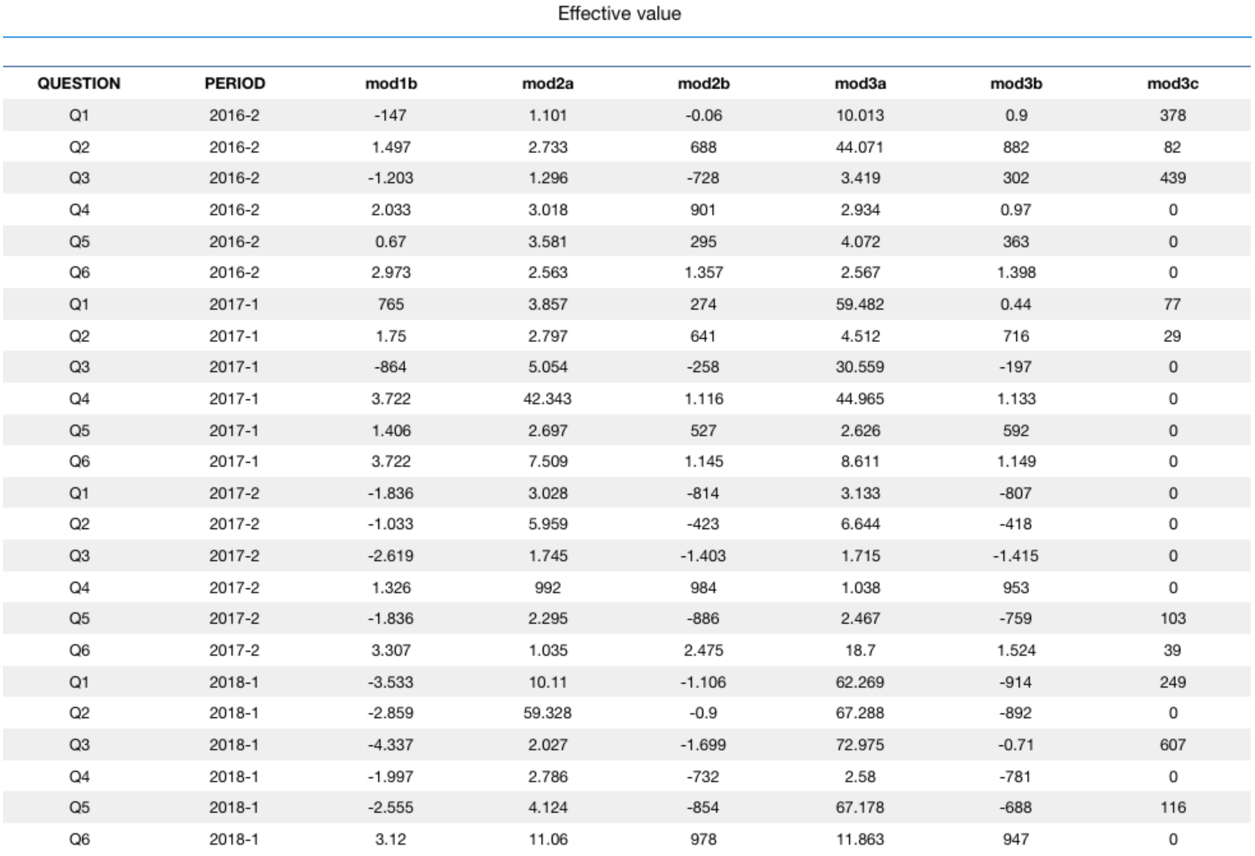


Table F: Coefficients of Logistic Models for Effective value questionnaire



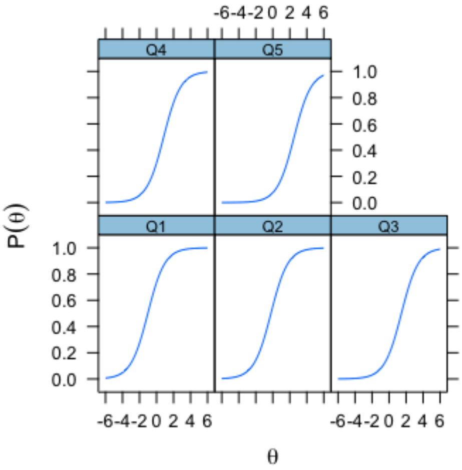
Effective value

QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2016-2	-147	1.101	-0.06	10.013	0.9	378
Q2	2016-2	1.497	2.733	688	44.071	882	82
Q3	2016-2	-1.203	1.296	-728	3.419	302	439
Q4	2016-2	2.033	3.018	901	2.934	0.97	0
Q5	2016-2	0.67	3.581	295	4.072	363	0
Q6	2016-2	2.973	2.563	1.357	2.567	1.398	0
Q1	2017-1	765	3.857	274	59.482	0.44	77
Q2	2017-1	1.75	2.797	641	4.512	716	29
Q3	2017-1	-864	5.054	-258	30.559	-197	0
Q4	2017-1	3.722	42.343	1.116	44.965	1.133	0
Q5	2017-1	1.406	2.697	527	2.626	592	0
Q6	2017-1	3.722	7.509	1.145	8.611	1.149	0
Q1	2017-2	-1.836	3.028	-814	3.133	-807	0
Q2	2017-2	-1.033	5.959	-423	6.644	-418	0
Q3	2017-2	-2.619	1.745	-1.403	1.715	-1.415	0
Q4	2017-2	1.326	992	984	1.038	953	0
Q5	2017-2	-1.836	2.295	-886	2.467	-759	103
Q6	2017-2	3.307	1.035	2.475	18.7	1.524	39
Q1	2018-1	-3.533	10.11	-1.106	62.269	-914	249
Q2	2018-1	-2.859	59.328	-0.9	67.288	-892	0
Q3	2018-1	-4.337	2.027	-1.699	72.975	-0.71	607
Q4	2018-1	-1.997	2.786	-732	2.58	-781	0
Q5	2018-1	-2.555	4.124	-854	67.178	-688	116
Q6	2018-1	3.12	11.06	978	11.863	947	0

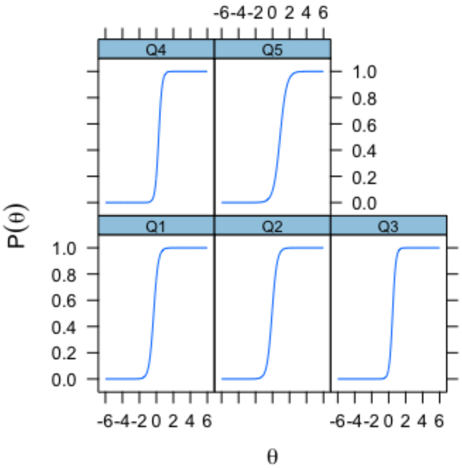
Graphical IRT Analysis: Questionnaire 7 - Phasors

Period: 2016-2

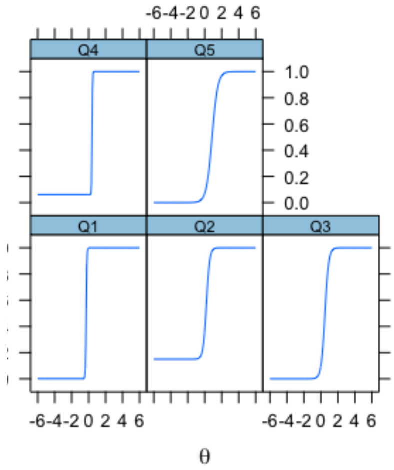
Item trace lines



Item trace lines

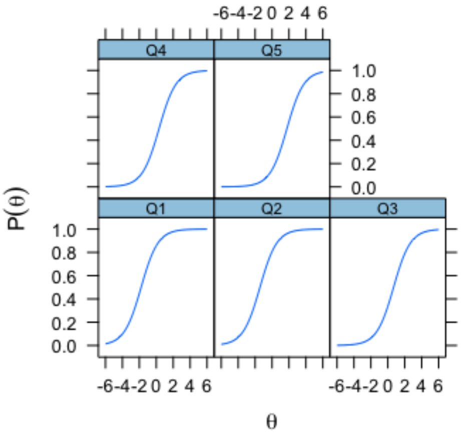


Item trace lines

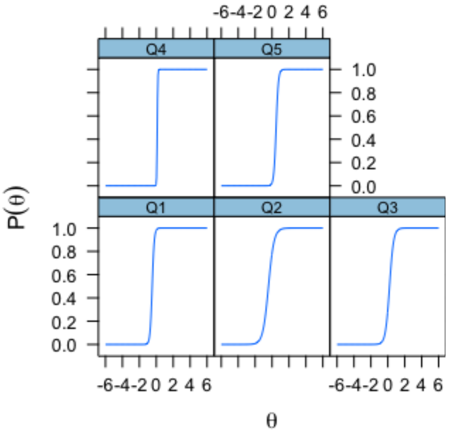


Period: 2017-1

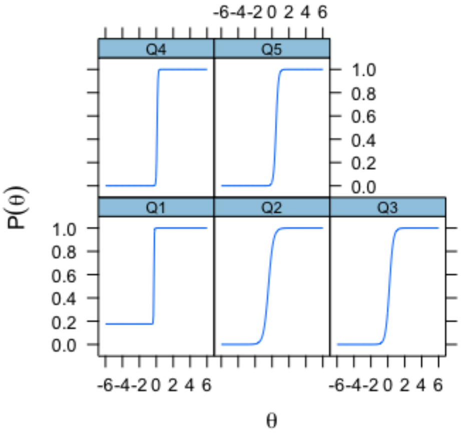
Item trace lines



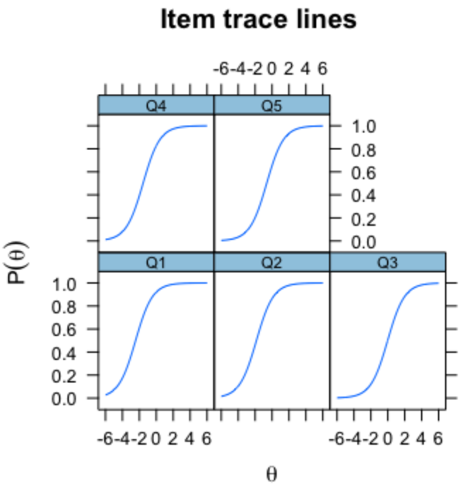
Item trace lines

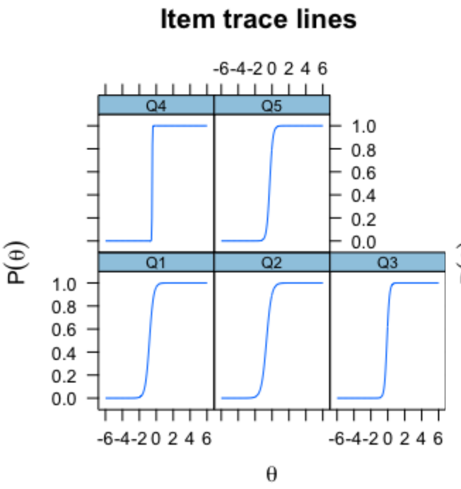


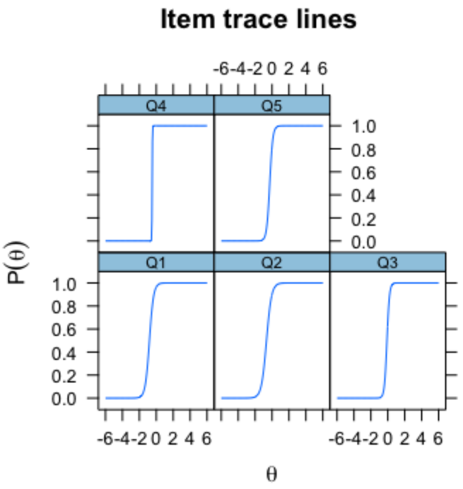
Item trace lines



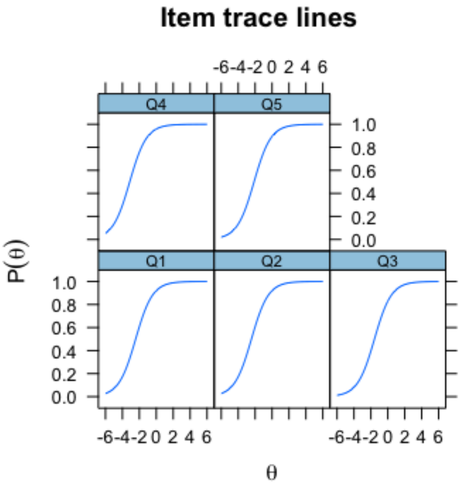
Period: 2017-2

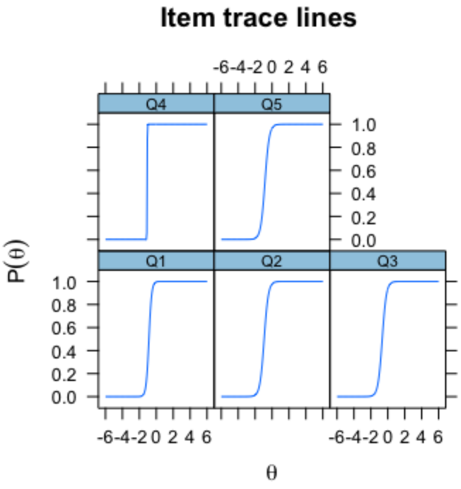






Period: 2018-1





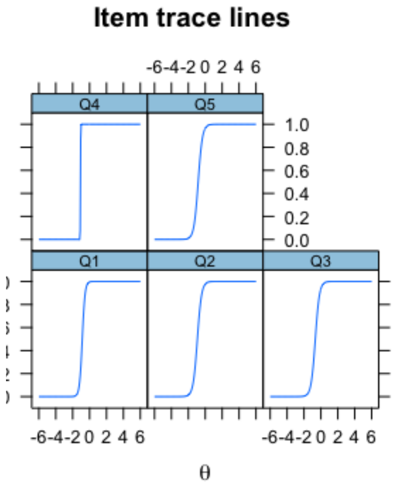
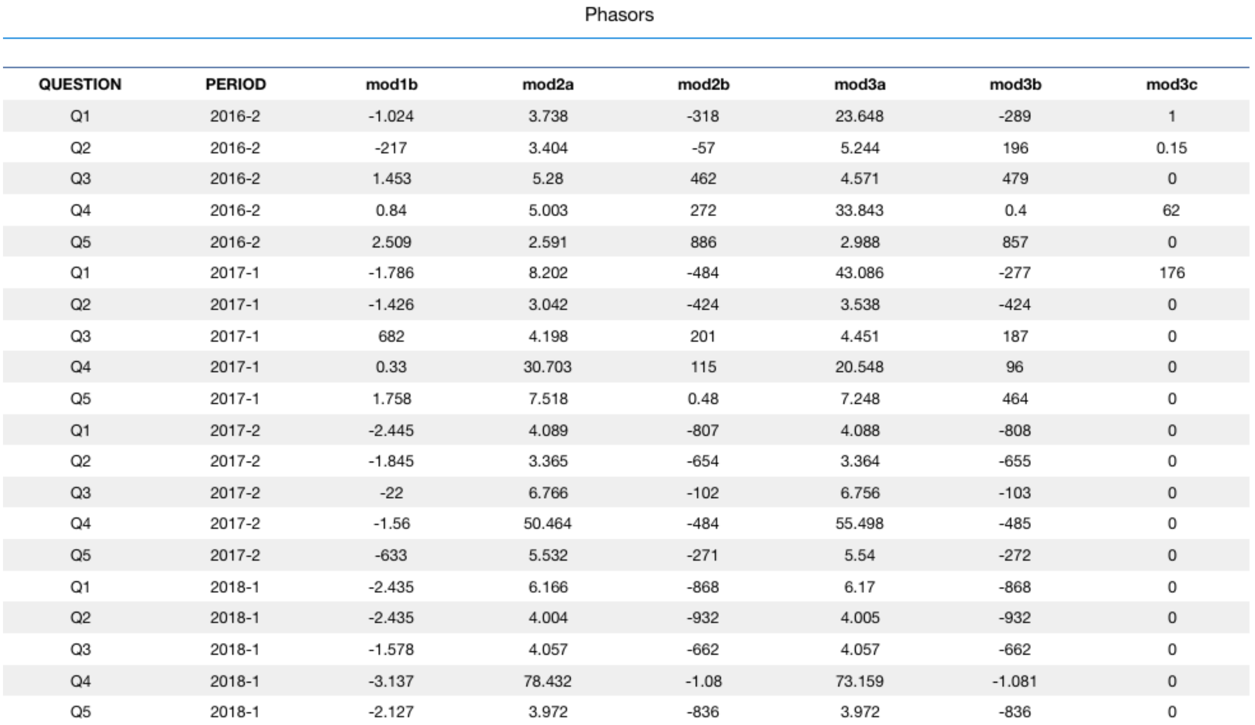


Table G: Coefficients of Logistic Models for Phasors questionnaire

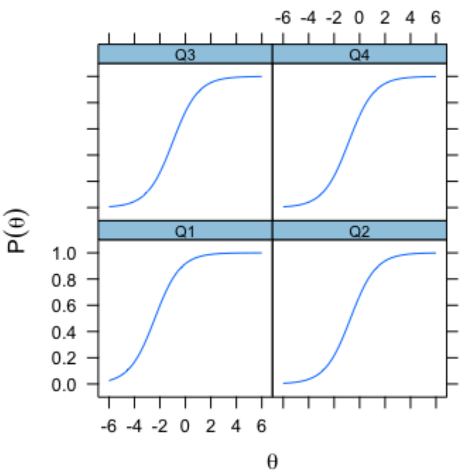


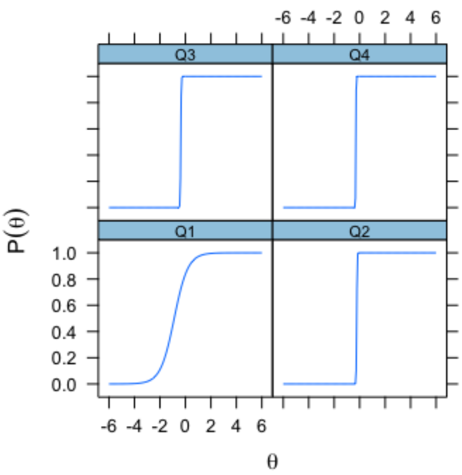
Phasors

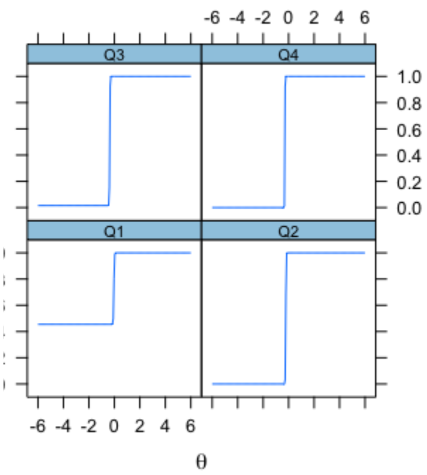
QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2016-2	-1.024	3.738	-318	23.648	-289	1
Q2	2016-2	-217	3.404	-57	5.244	196	0.15
Q3	2016-2	1.453	5.28	462	4.571	479	0
Q4	2016-2	0.84	5.003	272	33.843	0.4	62
Q5	2016-2	2.509	2.591	886	2.988	857	0
Q1	2017-1	-1.786	8.202	-484	43.086	-277	176
Q2	2017-1	-1.426	3.042	-424	3.538	-424	0
Q3	2017-1	682	4.198	201	4.451	187	0
Q4	2017-1	0.33	30.703	115	20.548	96	0
Q5	2017-1	1.758	7.518	0.48	7.248	464	0
Q1	2017-2	-2.445	4.089	-807	4.088	-808	0
Q2	2017-2	-1.845	3.365	-654	3.364	-655	0
Q3	2017-2	-22	6.766	-102	6.756	-103	0
Q4	2017-2	-1.56	50.464	-484	55.498	-485	0
Q5	2017-2	-633	5.532	-271	5.54	-272	0
Q1	2018-1	-2.435	6.166	-868	6.17	-868	0
Q2	2018-1	-2.435	4.004	-932	4.005	-932	0
Q3	2018-1	-1.578	4.057	-662	4.057	-662	0
Q4	2018-1	-3.137	78.432	-1.08	73.159	-1.081	0
Q5	2018-1	-2.127	3.972	-836	3.972	-836	0

Graphical IRT Analysis: Questionnaire 8 - Transformers

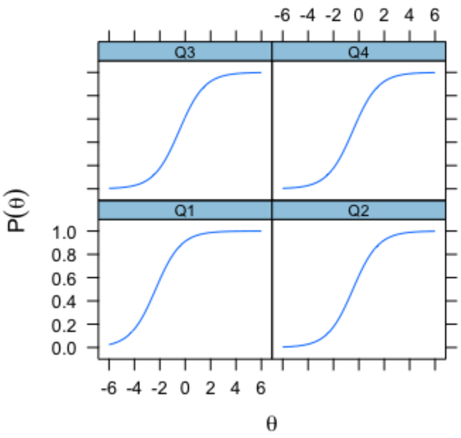
Period: 2016-2

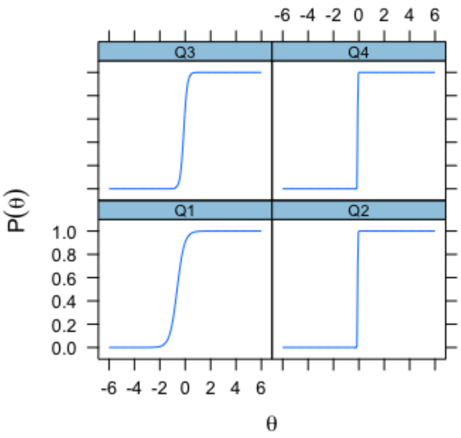

Item trace lines

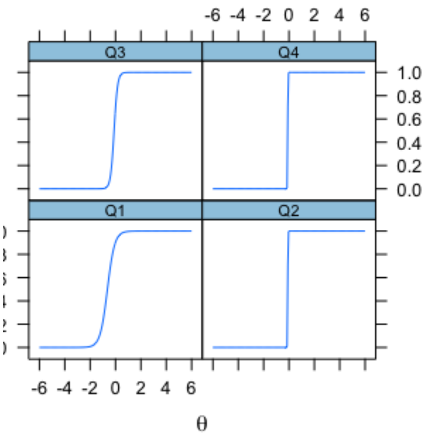

Item trace lines


Item trace lines

Period: 2017-1


Item trace lines


Item trace lines


Item trace lines

Period: 2017-2

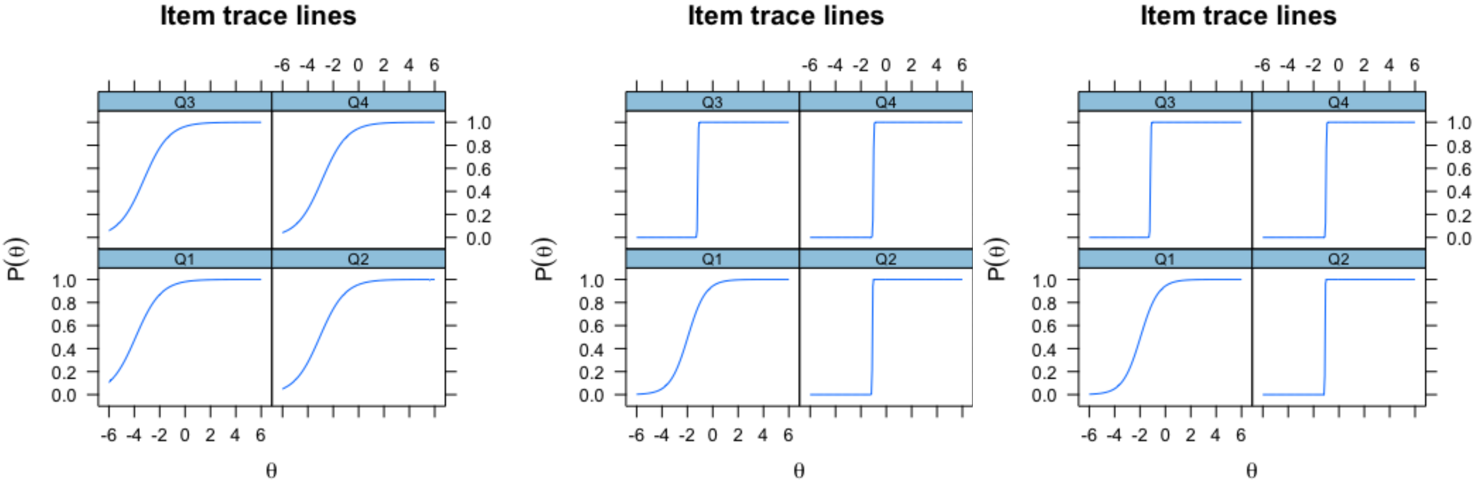
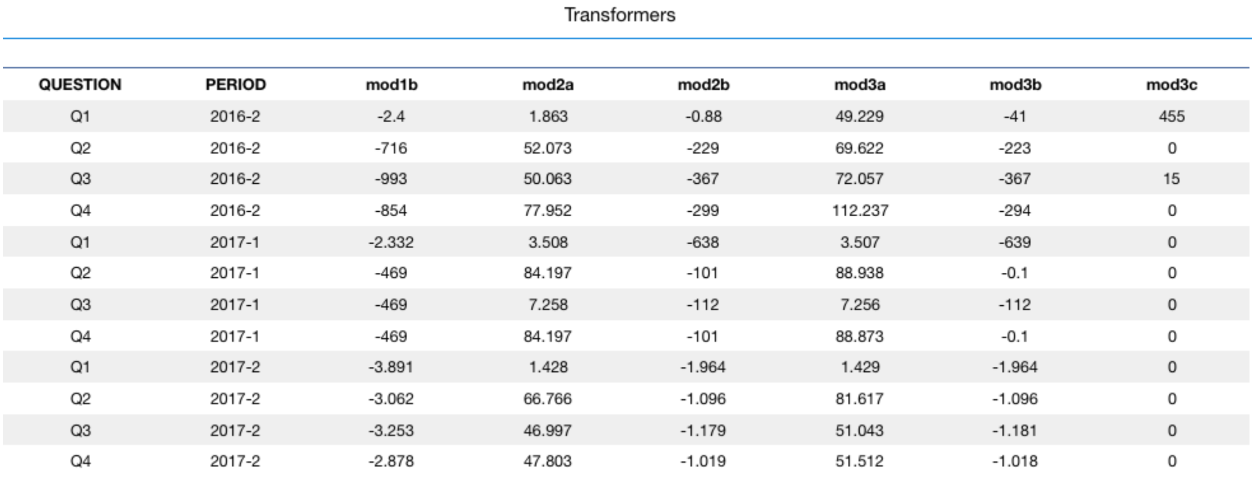


Table H: with coefficients of Logistic Models for Transformers questionnaire

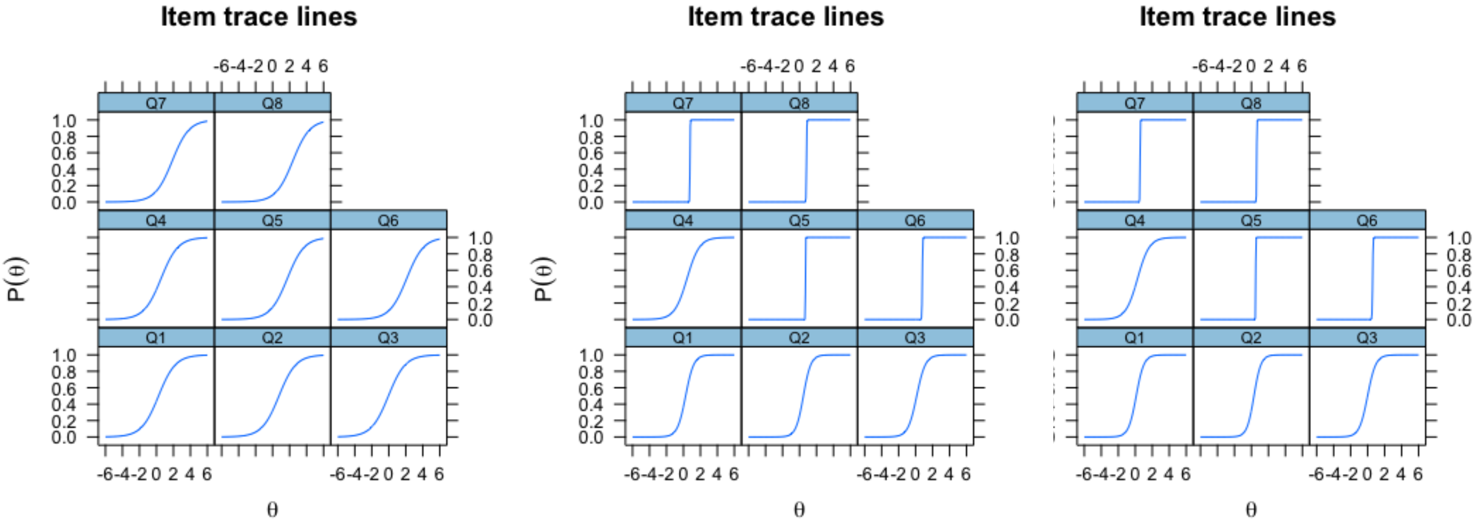


Transformers

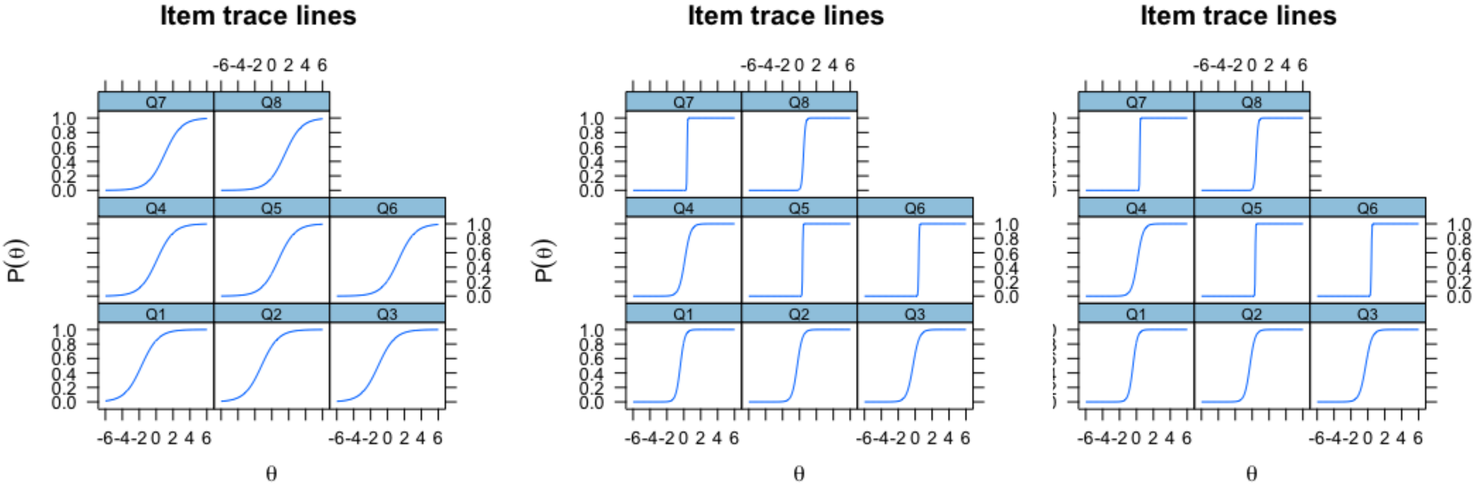
QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2016-2	-2.4	1.863	-0.88	49.229	-41	455
Q2	2016-2	-716	52.073	-229	69.622	-223	0
Q3	2016-2	-993	50.063	-367	72.057	-367	15
Q4	2016-2	-854	77.952	-299	112.237	-294	0
Q1	2017-1	-2.332	3.508	-638	3.507	-639	0
Q2	2017-1	-469	84.197	-101	88.938	-0.1	0
Q3	2017-1	-469	7.258	-112	7.256	-112	0
Q4	2017-1	-469	84.197	-101	88.873	-0.1	0
Q1	2017-2	-3.891	1.428	-1.964	1.429	-1.964	0
Q2	2017-2	-3.062	66.766	-1.096	81.617	-1.096	0
Q3	2017-2	-3.253	46.997	-1.179	51.043	-1.181	0
Q4	2017-2	-2.878	47.803	-1.019	51.512	-1.018	0

Graphical IRT Analysis: Questionnaire 9 - Multipole Alternator

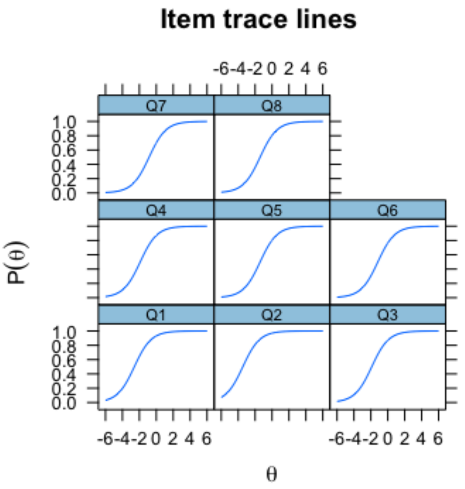
Period: 2016-2

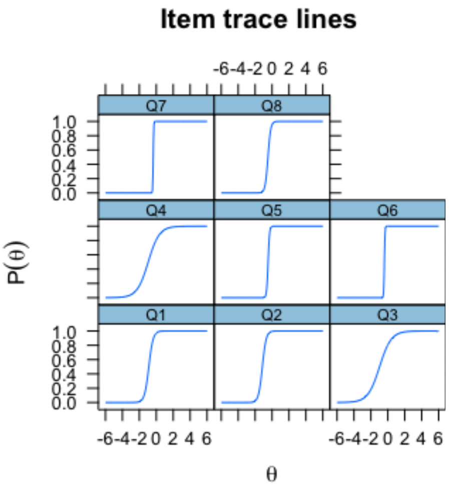


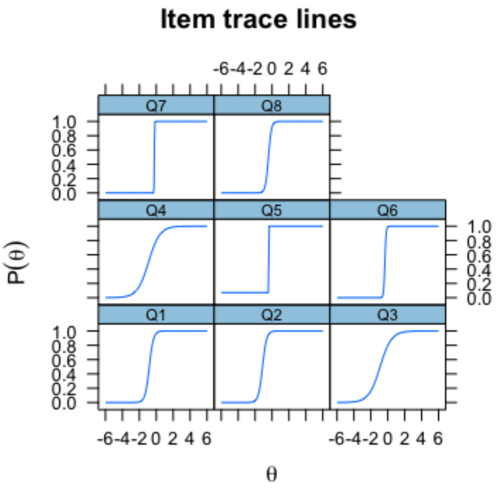
Period: 2017-1



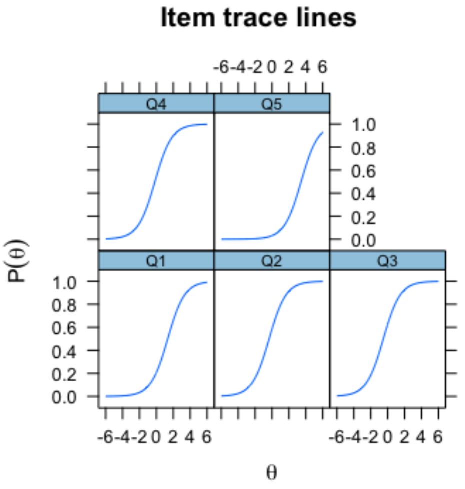
Period: 2017-2

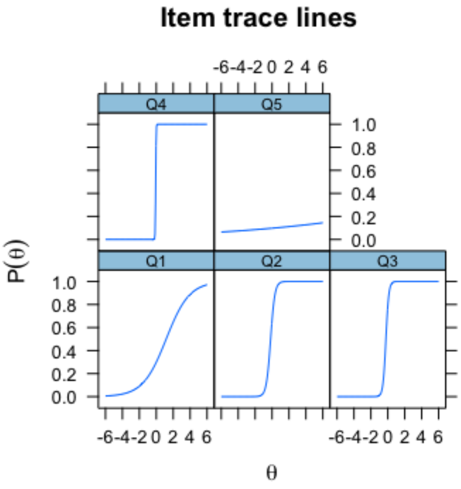






Period: 2018-1





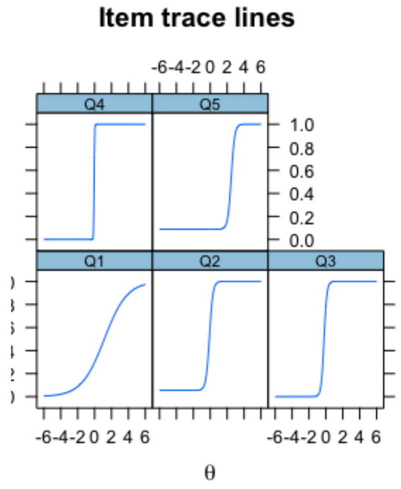
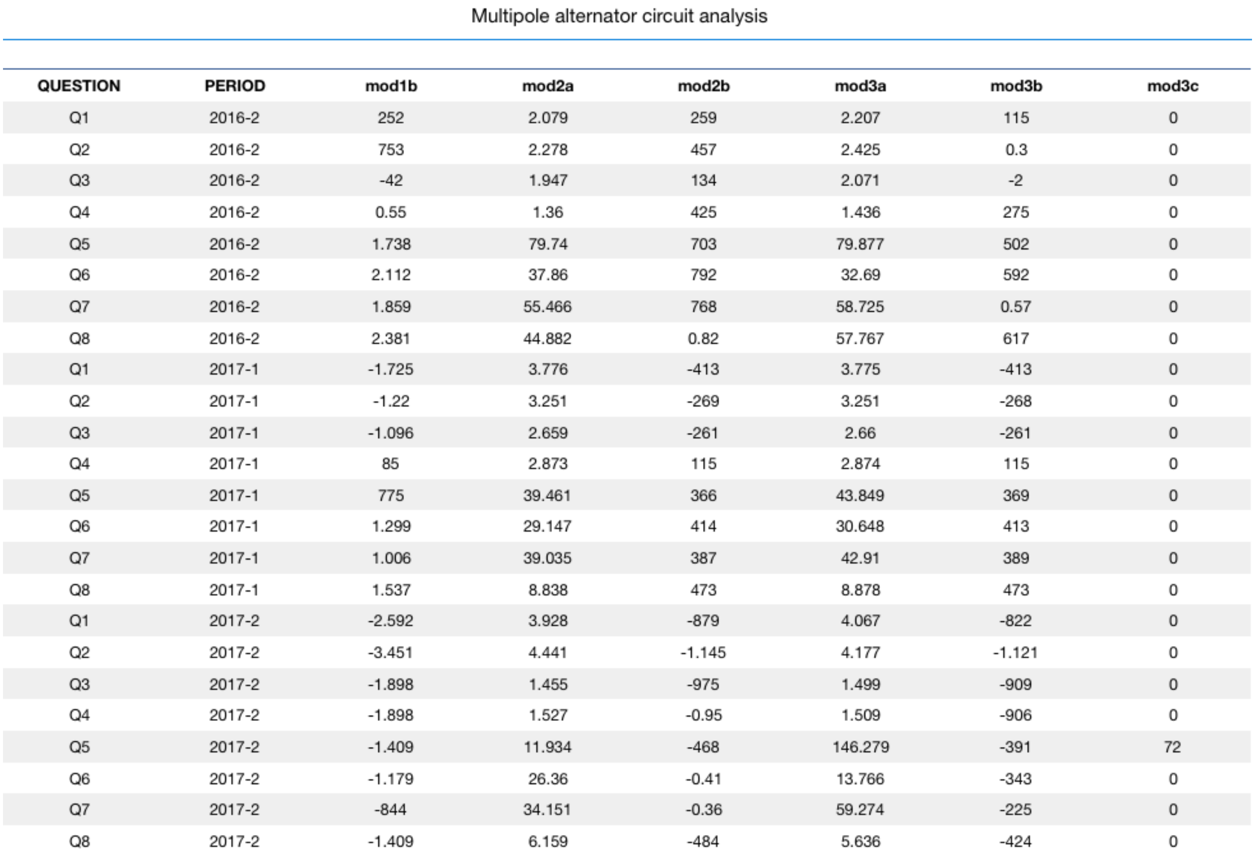


Table I: Coefficients of Logistic Models for Multi Pole alternator questionnaire

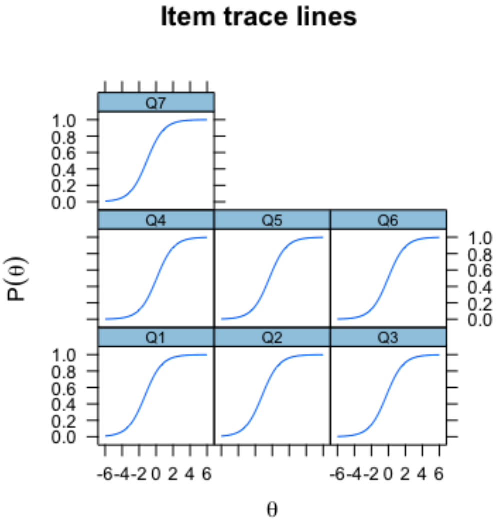


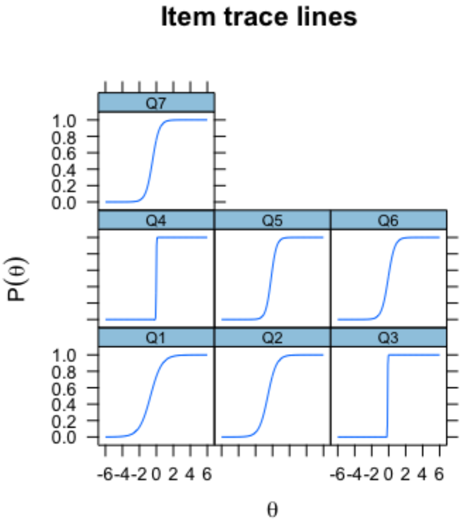
Multipole alternator circuit analysis

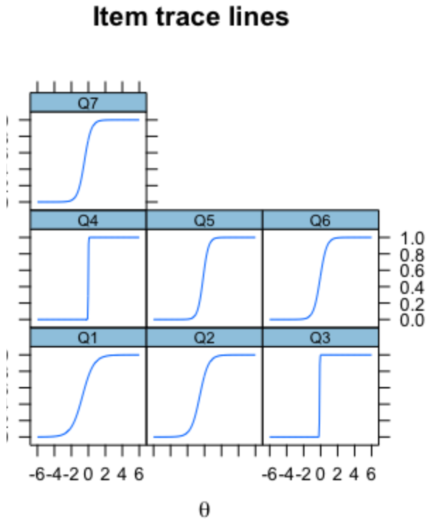
QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2016-2	252	2.079	259	2.207	115	0
Q2	2016-2	753	2.278	457	2.425	0.3	0
Q3	2016-2	-42	1.947	134	2.071	-2	0
Q4	2016-2	0.55	1.36	425	1.436	275	0
Q5	2016-2	1.738	79.74	703	79.877	502	0
Q6	2016-2	2.112	37.86	792	32.69	592	0
Q7	2016-2	1.859	55.466	768	58.725	0.57	0
Q8	2016-2	2.381	44.882	0.82	57.767	617	0
Q1	2017-1	-1.725	3.776	-413	3.775	-413	0
Q2	2017-1	-1.22	3.251	-269	3.251	-268	0
Q3	2017-1	-1.096	2.659	-261	2.66	-261	0
Q4	2017-1	85	2.873	115	2.874	115	0
Q5	2017-1	775	39.461	366	43.849	369	0
Q6	2017-1	1.299	29.147	414	30.648	413	0
Q7	2017-1	1.006	39.035	387	42.91	389	0
Q8	2017-1	1.537	8.838	473	8.878	473	0
Q1	2017-2	-2.592	3.928	-879	4.067	-822	0
Q2	2017-2	-3.451	4.441	-1.145	4.177	-1.121	0
Q3	2017-2	-1.898	1.455	-975	1.499	-909	0
Q4	2017-2	-1.898	1.527	-0.95	1.509	-906	0
Q5	2017-2	-1.409	11.934	-468	146.279	-391	72
Q6	2017-2	-1.179	26.36	-0.41	13.766	-343	0
Q7	2017-2	-844	34.151	-0.36	59.274	-225	0
Q8	2017-2	-1.409	6.159	-484	5.636	-424	0

Graphical IRT Analysis: Questionnaire 10 - Transformers II

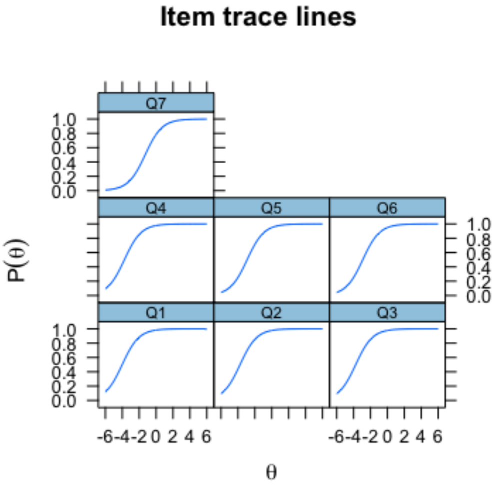
Period: 2016-2

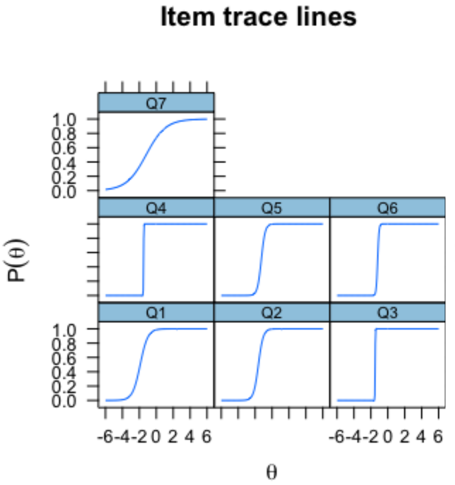


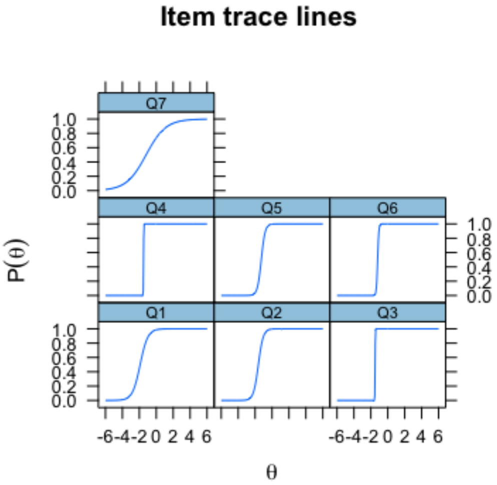




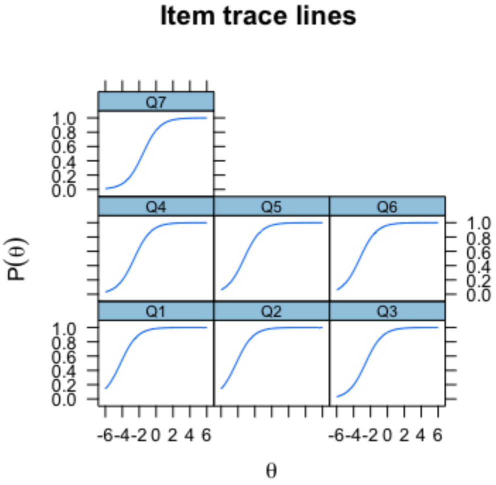
Period: 2017-2

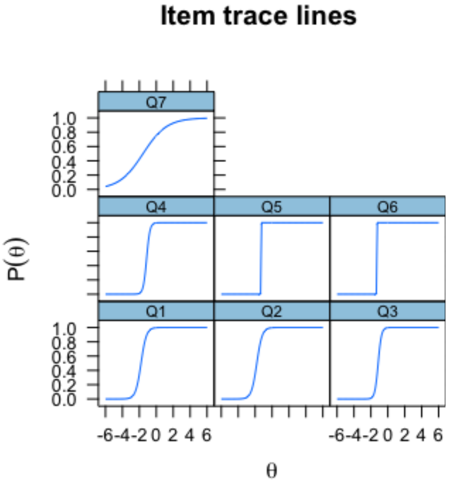






Period: 2018-1





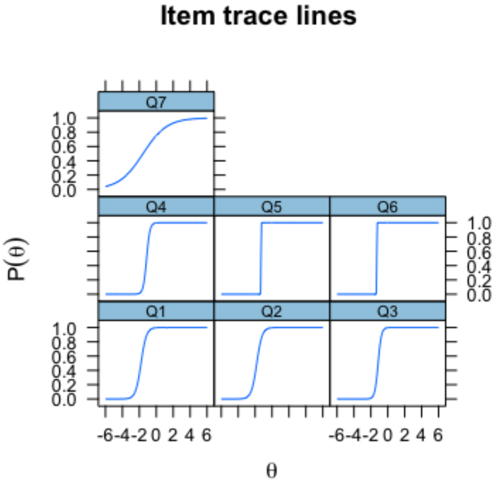
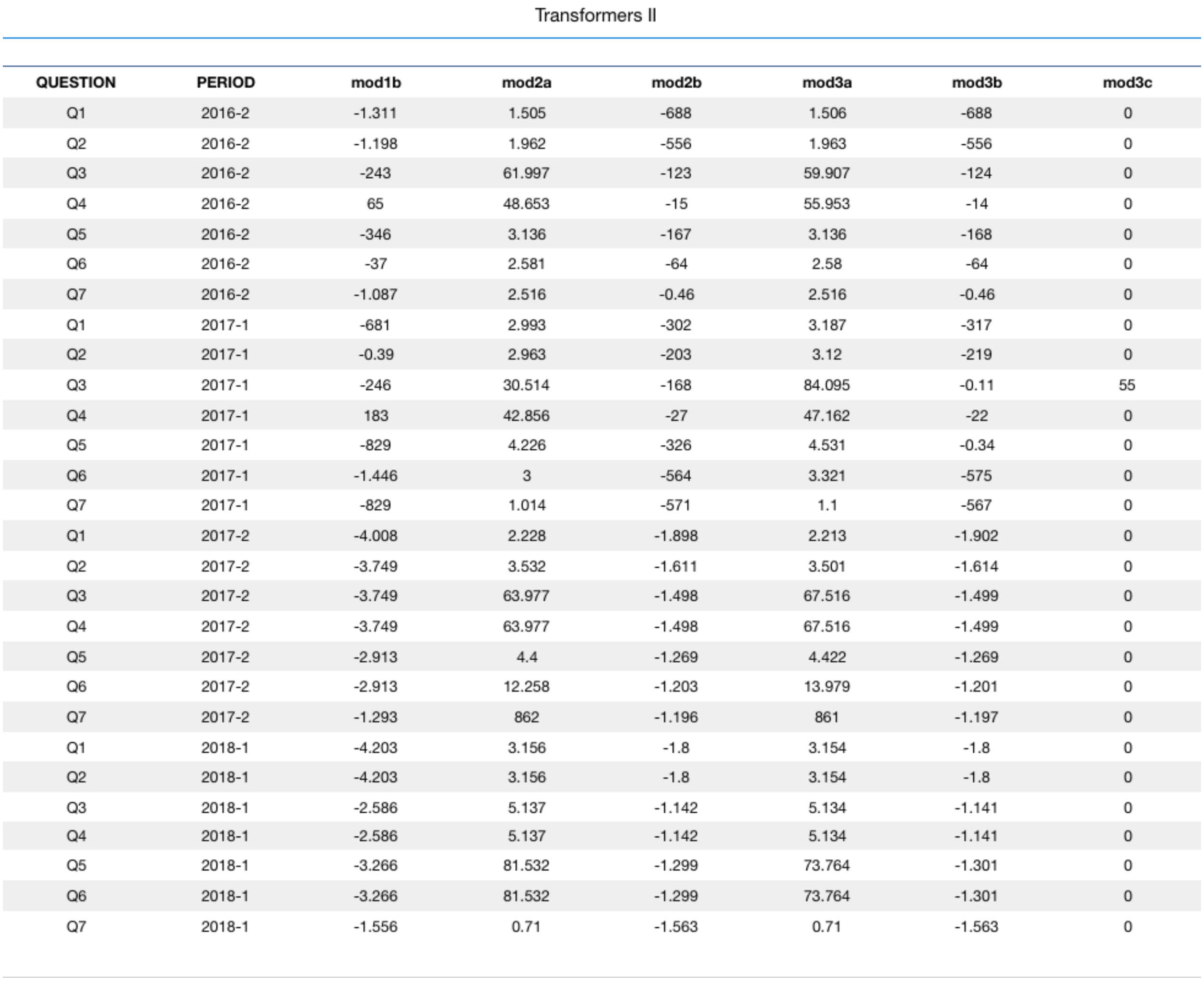


Table J: Coefficients of Logistic Models for Transformers II questionnaire

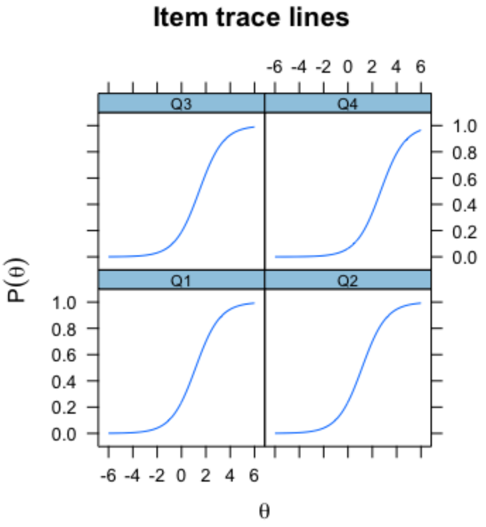


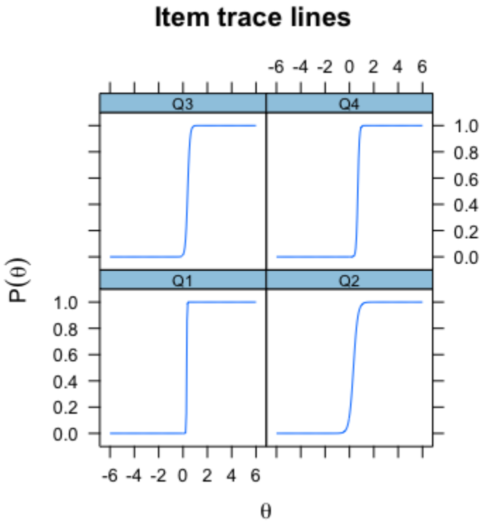
Transformers II

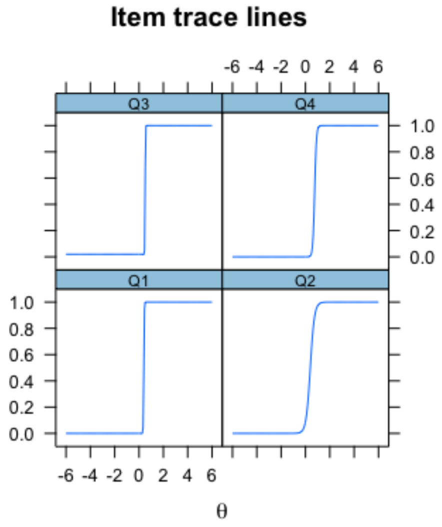
QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2016-2	-1.311	1.505	-688	1.506	-688	0
Q2	2016-2	-1.198	1.962	-556	1.963	-556	0
Q3	2016-2	-243	61.997	-123	59.907	-124	0
Q4	2016-2	65	48.653	-15	55.953	-14	0
Q5	2016-2	-346	3.136	-167	3.136	-168	0
Q6	2016-2	-37	2.581	-64	2.58	-64	0
Q7	2016-2	-1.087	2.516	-0.46	2.516	-0.46	0
Q1	2017-1	-681	2.993	-302	3.187	-317	0
Q2	2017-1	-0.39	2.963	-203	3.12	-219	0
Q3	2017-1	-246	30.514	-168	84.095	-0.11	55
Q4	2017-1	183	42.856	-27	47.162	-22	0
Q5	2017-1	-829	4.226	-326	4.531	-0.34	0
Q6	2017-1	-1.446	3	-564	3.321	-575	0
Q7	2017-1	-829	1.014	-571	1.1	-567	0
Q1	2017-2	-4.008	2.228	-1.898	2.213	-1.902	0
Q2	2017-2	-3.749	3.532	-1.611	3.501	-1.614	0
Q3	2017-2	-3.749	63.977	-1.498	67.516	-1.499	0
Q4	2017-2	-3.749	63.977	-1.498	67.516	-1.499	0
Q5	2017-2	-2.913	4.4	-1.269	4.422	-1.269	0
Q6	2017-2	-2.913	12.258	-1.203	13.979	-1.201	0
Q7	2017-2	-1.293	862	-1.196	861	-1.197	0
Q1	2018-1	-4.203	3.156	-1.8	3.154	-1.8	0
Q2	2018-1	-4.203	3.156	-1.8	3.154	-1.8	0
Q3	2018-1	-2.586	5.137	-1.142	5.134	-1.141	0
Q4	2018-1	-2.586	5.137	-1.142	5.134	-1.141	0
Q5	2018-1	-3.266	81.532	-1.299	73.764	-1.301	0
Q6	2018-1	-3.266	81.532	-1.299	73.764	-1.301	0
Q7	2018-1	-1.556	0.71	-1.563	0.71	-1.563	0

Graphical IRT Analysis: Questionnaire 11 - Three Phase Transformers

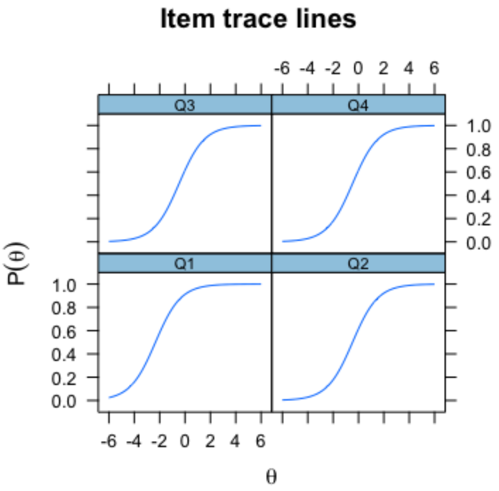
Period: 2016-2

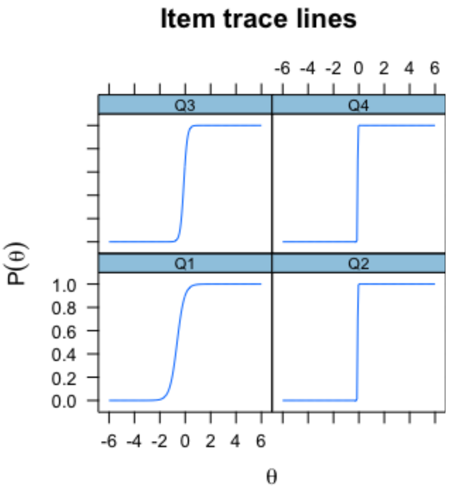


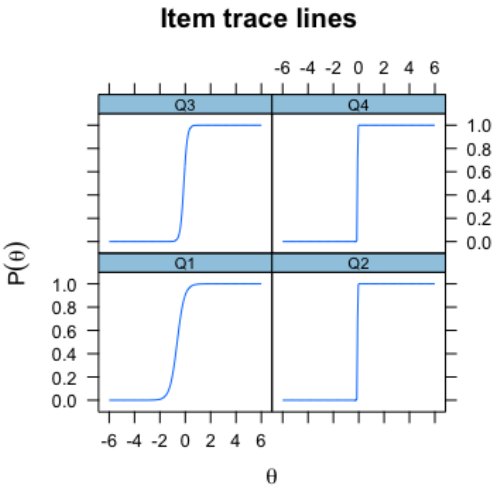




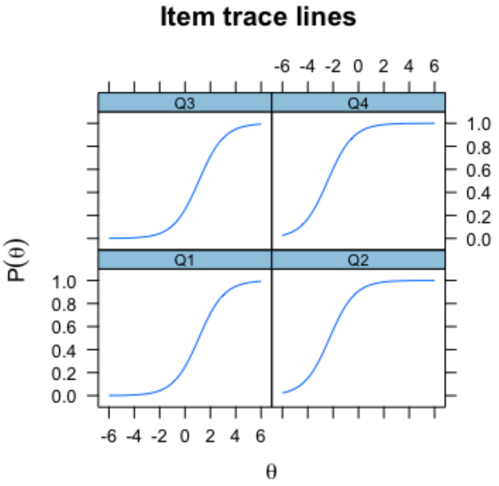
Period: 2017-1

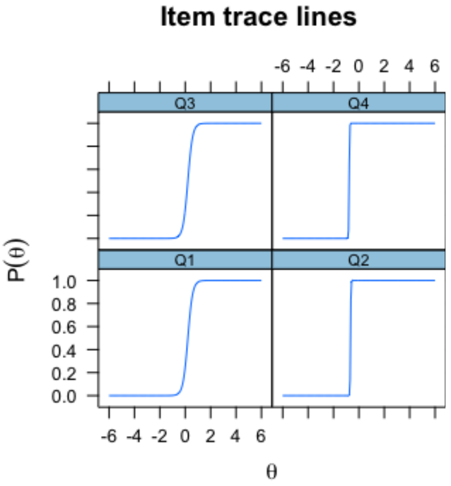


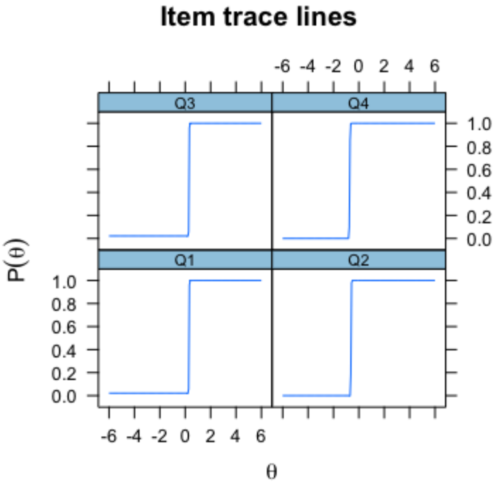




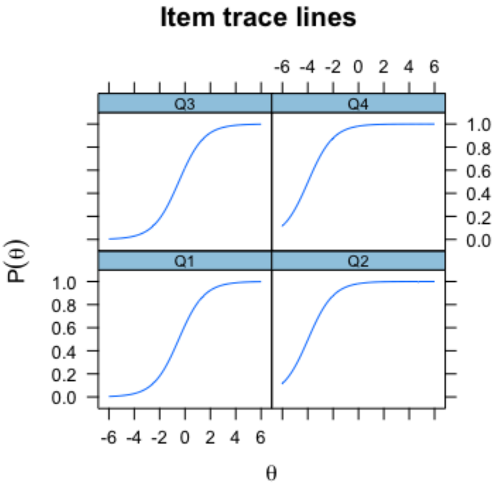
Period: 2017-2

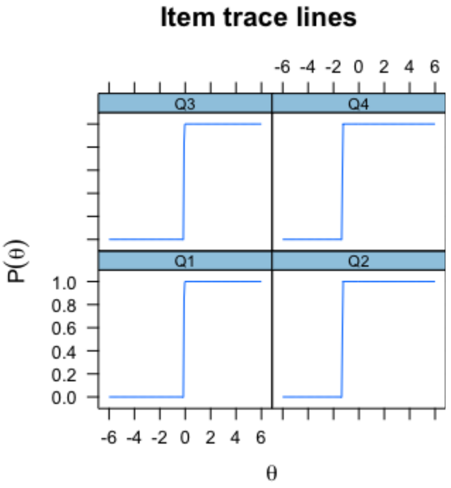






Period: 2018-1





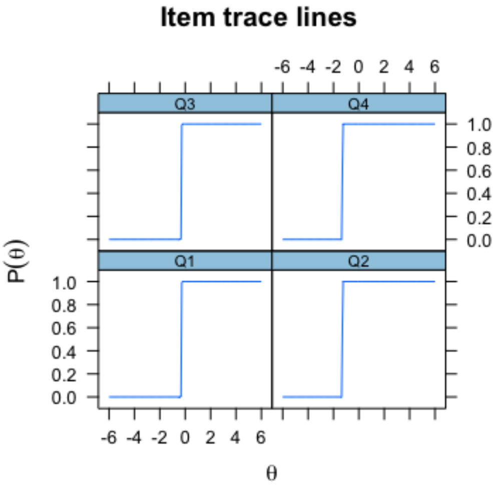
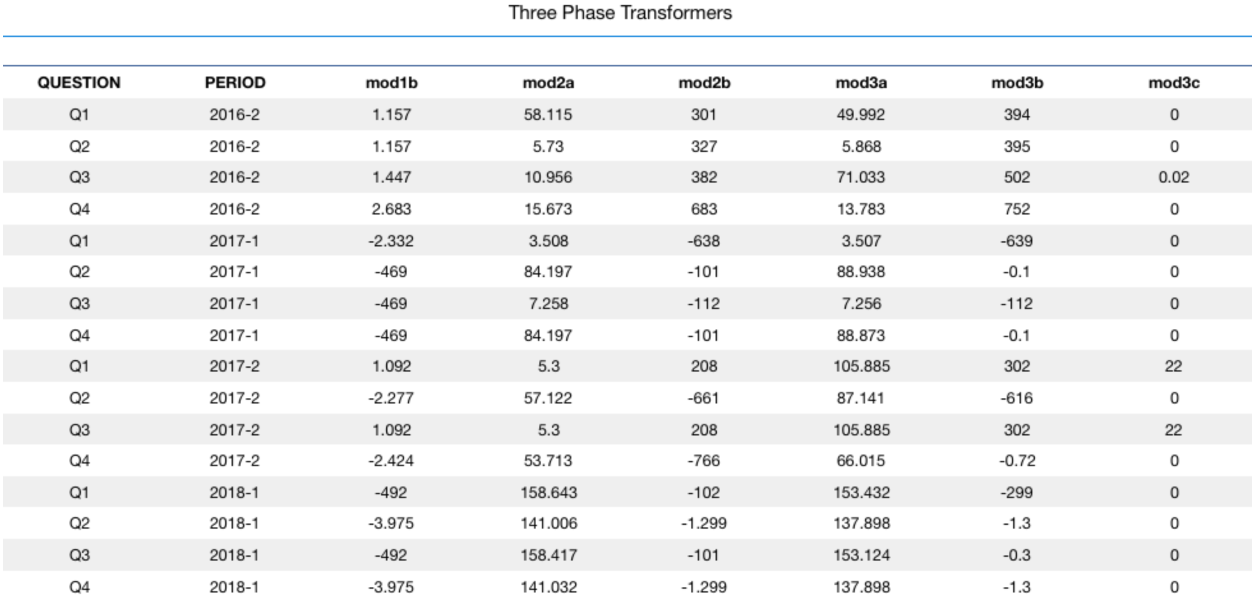


Table K: Coefficients of Logistic Models for Three phase transformers questionnaire

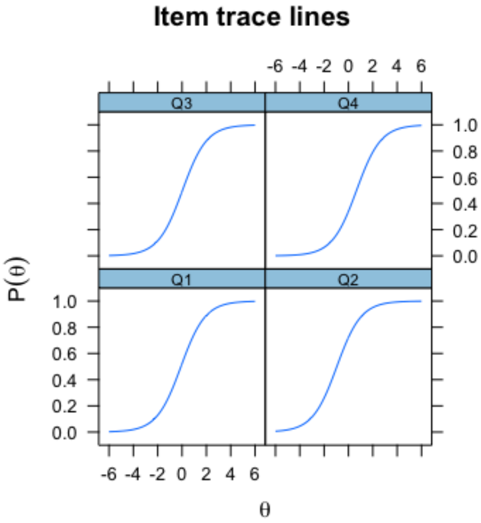


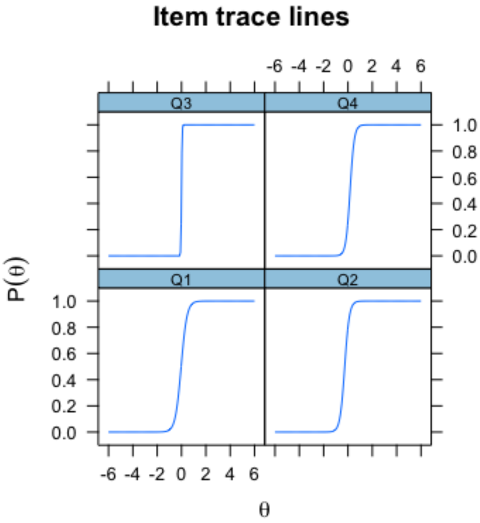
Three Phase Transformers

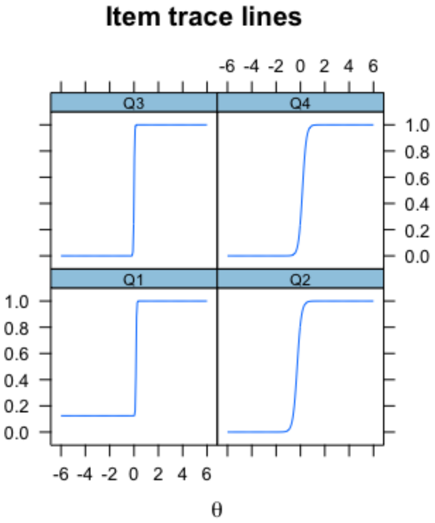
QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2016-2	1.157	58.115	301	49.992	394	0
Q2	2016-2	1.157	5.73	327	5.868	395	0
Q3	2016-2	1.447	10.956	382	71.033	502	0.02
Q4	2016-2	2.683	15.673	683	13.783	752	0
Q1	2017-1	-2.332	3.508	-638	3.507	-639	0
Q2	2017-1	-469	84.197	-101	88.938	-0.1	0
Q3	2017-1	-469	7.258	-112	7.256	-112	0
Q4	2017-1	-469	84.197	-101	88.873	-0.1	0
Q1	2017-2	1.092	5.3	208	105.885	302	22
Q2	2017-2	-2.277	57.122	-661	87.141	-616	0
Q3	2017-2	1.092	5.3	208	105.885	302	22
Q4	2017-2	-2.424	53.713	-766	66.015	-0.72	0
Q1	2018-1	-492	158.643	-102	153.432	-299	0
Q2	2018-1	-3.975	141.006	-1.299	137.898	-1.3	0
Q3	2018-1	-492	158.417	-101	153.124	-0.3	0
Q4	2018-1	-3.975	141.032	-1.299	137.898	-1.3	0

Graphical IRT Analysis: Questionnaire 12 - Power Factor

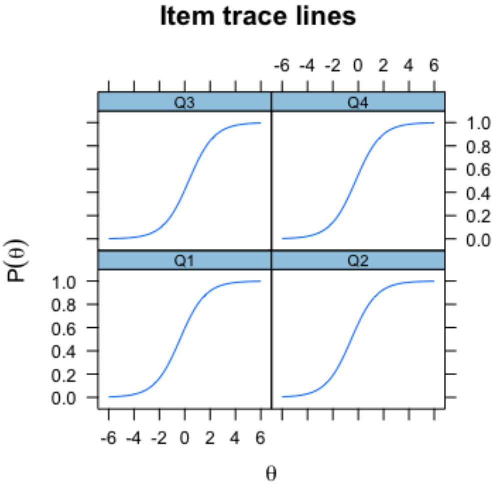
Period: 2016-2

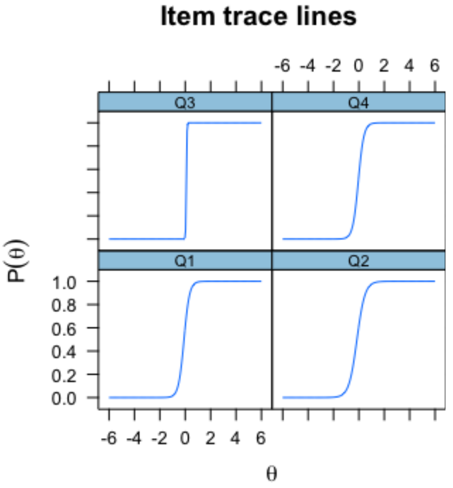

Item trace lines

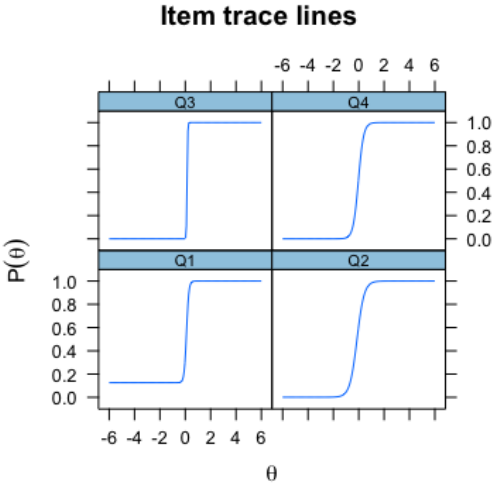

Item trace lines


Item trace lines

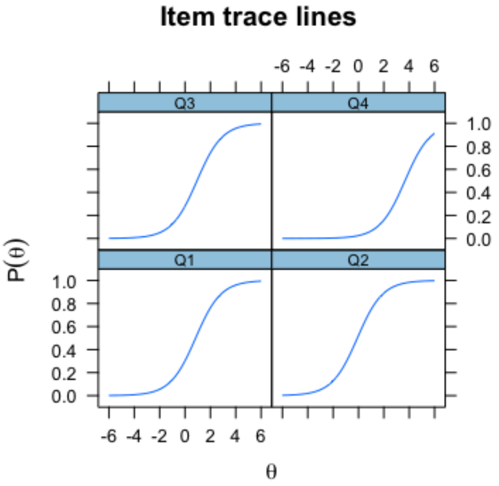
Period: 2017-1

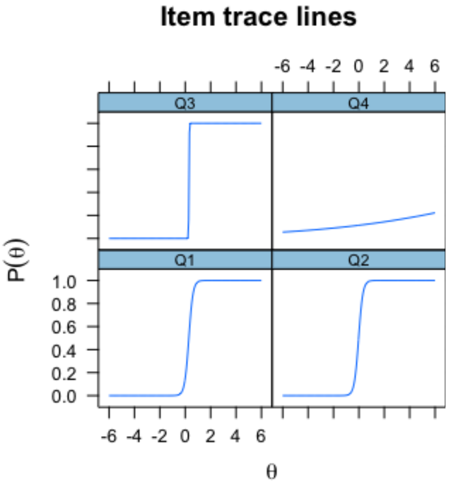

Item trace lines

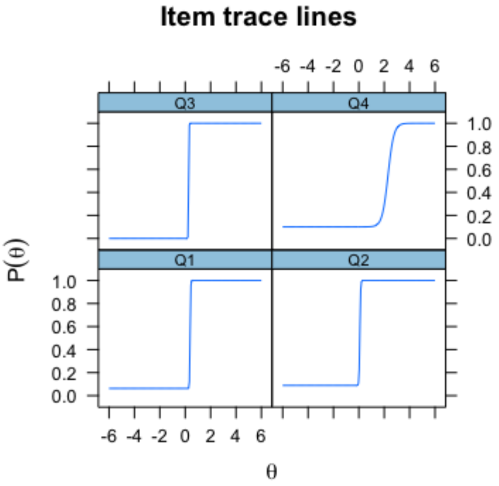

Item trace lines


Item trace lines

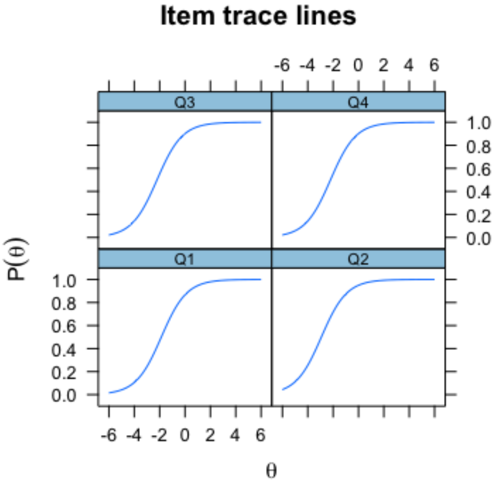
Period: 2017-2

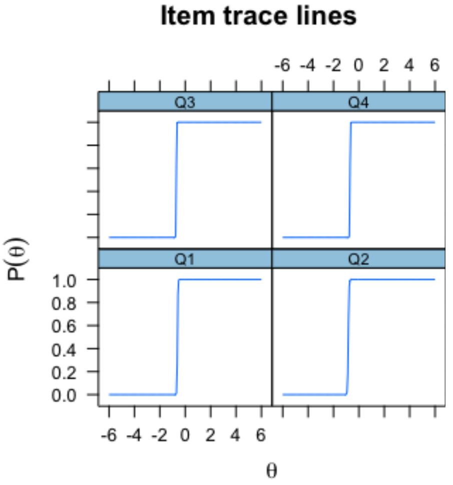






Period: 2018-1





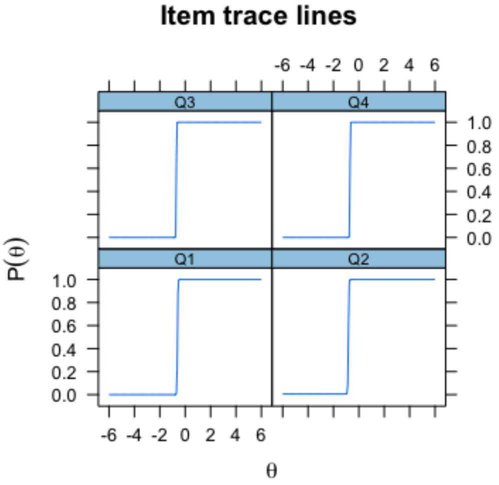
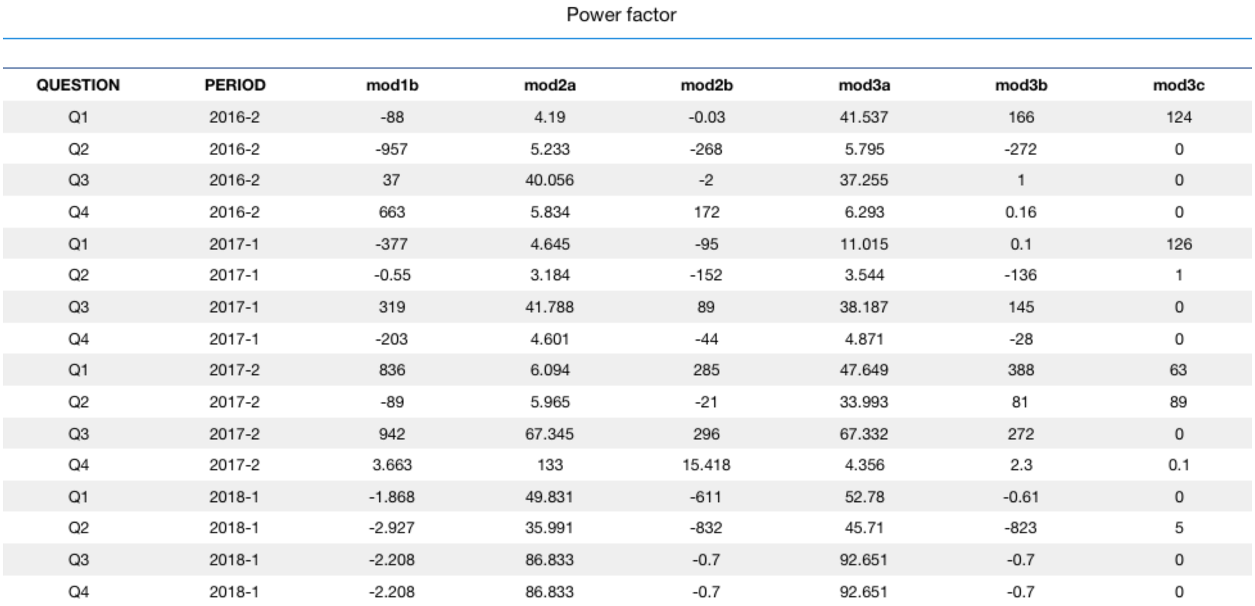


Table L: Coefficients of Logistic Models for Power factor questionnaire



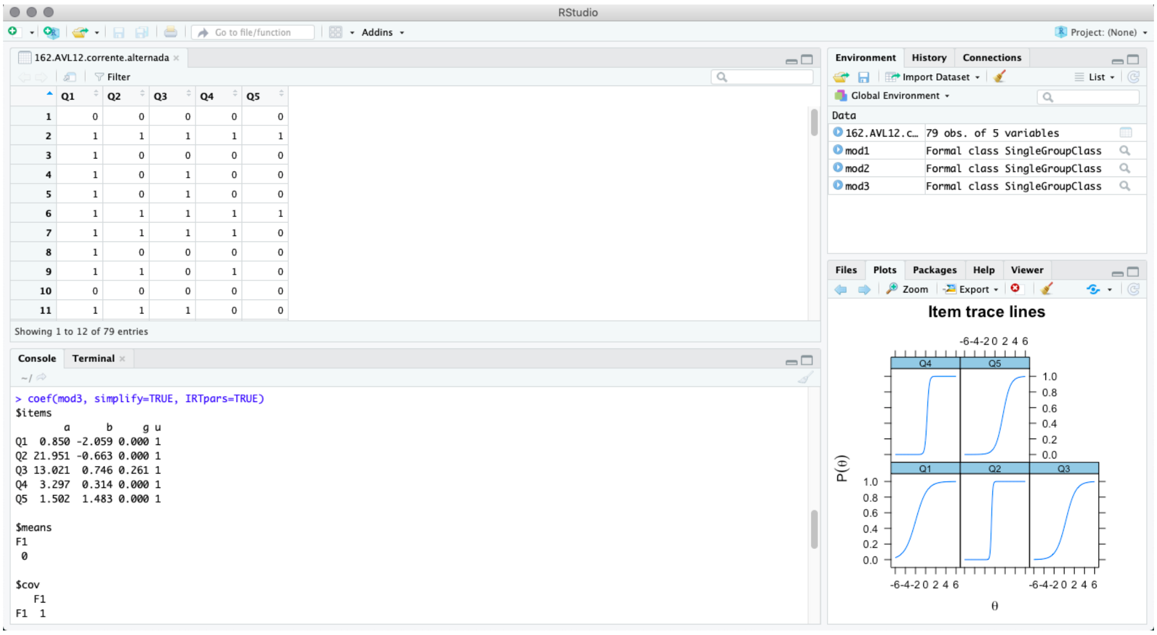
Power factor

QUESTION	PERIOD	mod1b	mod2a	mod2b	mod3a	mod3b	mod3c
Q1	2016-2	-88	4.19	-0.03	41.537	166	124
Q2	2016-2	-957	5.233	-268	5.795	-272	0
Q3	2016-2	37	40.056	-2	37.255	1	0
Q4	2016-2	663	5.834	172	6.293	0.16	0
Q1	2017-1	-377	4.645	-95	11.015	0.1	126
Q2	2017-1	-0.55	3.184	-152	3.544	-136	1
Q3	2017-1	319	41.788	89	38.187	145	0
Q4	2017-1	-203	4.601	-44	4.871	-28	0
Q1	2017-2	836	6.094	285	47.649	388	63
Q2	2017-2	-89	5.965	-21	33.993	81	89
Q3	2017-2	942	67.345	296	67.332	272	0
Q4	2017-2	3.663	133	15.418	4.356	2.3	0.1
Q1	2018-1	-1.868	49.831	-611	52.78	-0.61	0
Q2	2018-1	-2.927	35.991	-832	45.71	-823	5
Q3	2018-1	-2.208	86.833	-0.7	92.651	-0.7	0
Q4	2018-1	-2.208	86.833	-0.7	92.651	-0.7	0

APPENDIX B <Computational Resources>

Computational Resources

Using the R[4] language for the Implementation of the MIRT and LTM Libraries for the calculus of the of three logistic models, ANOVA test, Graphic Interface and coefficient matrix provided for the IRT Analysis, Figure 7 shows the results of the logistic Model of three parameters for the period of 2016-2. Describing the related figure, have a questionnaire "Alternate Current" from Electrical engineering Course, in the superior part are presented the binarized data imported from the .csv file, in the inferior left side are calculated de coefficients from the model and the right side shows the graphical results for each question analyzed.



	Q1	Q2	Q3	Q4	Q5
1	0	0	0	0	0
2	1	1	1	1	1
3	1	0	0	0	0
4	1	0	1	0	0
5	1	0	1	0	0
6	1	1	1	1	1
7	1	1	1	1	0
8	1	0	0	0	0
9	1	1	0	1	0
10	0	0	0	0	0
11	1	1	1	0	0

Showing 1 to 12 of 79 entries

```
> coef(mod3, simplify=TRUE, IRTpars=TRUE)
$items
        a      b     g u
Q1  0.850 -2.059 0.000 1
Q2 21.951 -0.663 0.000 1
Q3 13.021  0.746 0.261 1
Q4  3.297  0.314 0.000 1
Q5  1.502  1.483 0.000 1

$means
F1
 0

$cov
   F1
F1  1
```

Data	
162.AVL12.c…	79 obs. of 5 variables
mod1	Formal class SingleGroupClass
mod2	Formal class SingleGroupClass
mod3	Formal class SingleGroupClass

Figure 26: IRT Analysis in Rstudio Platform[5].

For helping teachers import and manipulate results, a Web application was built to allow the importation of .csv files using PHP[6] language. Figure 8 shows some layers used.

[4] https://www.r-project.org
[5] https://www.rstudio.com
[6] https://www.php.net PHP Version 7.0.15

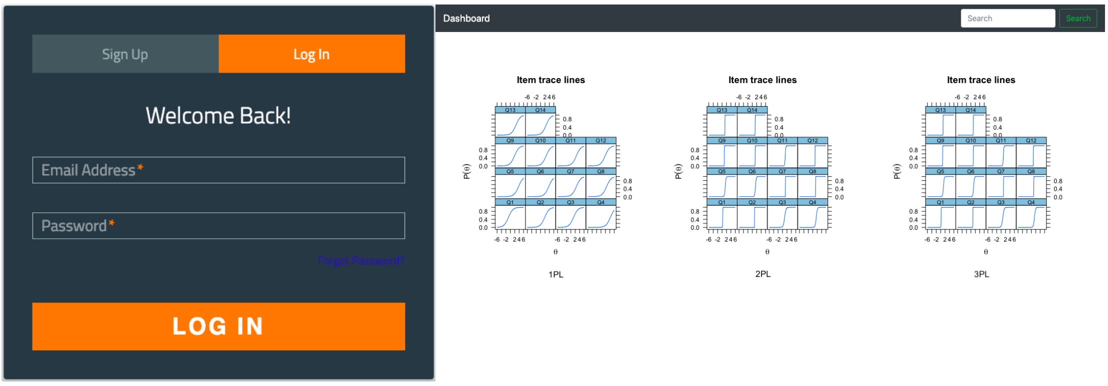
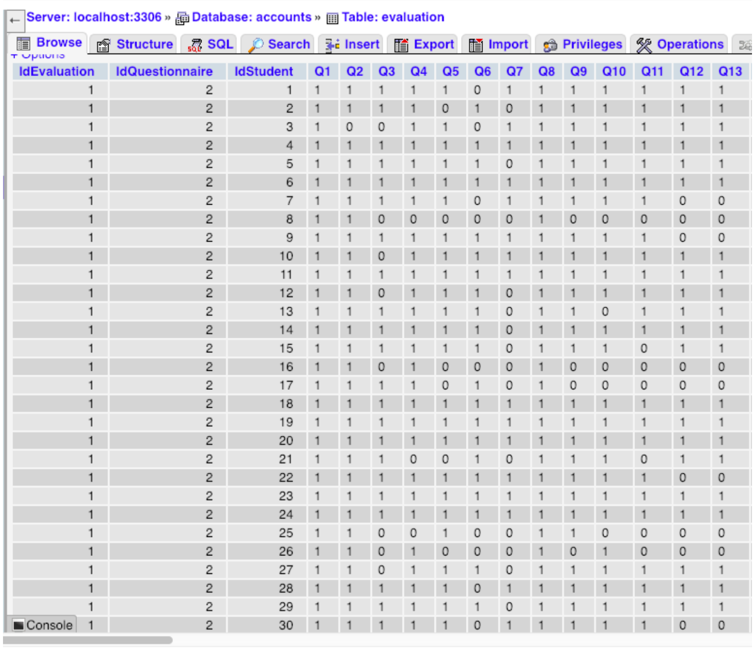


Figure 27: Interface with the login and dashboard

For the administration of the database were used Mysql (figure 28).



IdEvaluation	IdQuestionnaire	IdStudent	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
1	2	1	1	1	1	1	1	0	1	1	1	1	1	1	1
1	2	2	1	1	1	1	0	1	0	1	1	1	1	1	1
1	2	3	1	0	0	1	1	0	1	1	1	1	1	1	1
1	2	4	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	5	1	1	1	1	1	1	0	1	1	1	1	1	1
1	2	6	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	7	1	1	1	1	1	0	1	1	1	1	1	0	0
1	2	8	1	1	0	0	0	0	0	1	0	0	0	0	0
1	2	9	1	1	1	1	1	1	1	1	1	1	1	0	0
1	2	10	1	1	0	1	1	1	1	1	1	1	1	1	1
1	2	11	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	12	1	1	0	1	1	1	0	1	1	1	1	1	1
1	2	13	1	1	1	1	1	1	0	1	1	0	1	1	1
1	2	14	1	1	1	1	1	1	0	1	1	1	1	1	1
1	2	15	1	1	1	1	1	1	0	1	1	1	0	1	1
1	2	16	1	1	0	1	0	0	0	1	0	0	0	0	0
1	2	17	1	1	1	1	0	1	0	1	0	0	0	0	0
1	2	18	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	19	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	20	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	21	1	1	1	0	0	1	0	1	1	1	0	1	1
1	2	22	1	1	1	1	1	1	1	1	1	1	1	0	0
1	2	23	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	24	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	25	1	1	0	0	1	0	0	1	1	0	0	0	0
1	2	26	1	1	0	1	0	0	0	1	0	1	0	0	0
1	2	27	1	1	0	1	1	1	0	1	1	1	1	1	1
1	2	28	1	1	1	1	1	0	1	1	1	1	1	1	1
1	2	29	1	1	1	1	1	1	0	1	1	1	1	1	1
1	2	30	1	1	1	1	1	0	1	1	1	1	1	0	0

Figure 28: Binarized data - Kirchoff laws questionnaire 18-01 Period