University of Ioannina
Department of Primary School Education
Postgraduate Studies Program “Science in Education”

Calibration and Validation of Instruments measuring Academic Ability in Physics using Item Response Theory

Kapsalas Ioannis

Postgraduate Thesis

Supervisor: Mavridis Dimitrios

September 2017

Calibration and Validation of Instruments measuring Academic Ability in Physics using Item Response Theory

Καψάλας Ιωάννης

Μεταπτυχιακή Εργασία υποβληθείσα για την εκπλήρωση των προϋποθέσεων απονομής Μεταπτυχιακού Διπλώματος Ειδίκευσης στις Επιστήμες της Αγωγής του Π.Τ.Δ.Ε. του Πανεπιστημίου Ιωαννίνων

2017

Contents

Acknowledgments

Special thanks to my family, for the support through all the years of my studies, that made it possible for me to reach this level of academic education and also for the great help with the accomplishment of this thesis.

I would also like to thank my supervisor, Mr. Mavridis, for the support and guidance and Mr. Kotsis for generously providing me with the data of his research.

Finally, I would like to thank each one of the participants of my research for their willingness to take some of their personal time to fill up my questionnaires.

Abstract

Educational measurement typically aims at evaluating the abilities and knowledge of students in various fields such as mathematics, language, science, physics etc., using tests, questionnaires and other instruments. The aim of this thesis is to create valid instruments that measure the academic ability in mechanics through calibrating already used ones. An instrument that was used for similar purposes in other studies was analyzed separately for each category of the sample, by fitting the data into IRT models with the statistical software Stata 14. Based on the results of the analysis, a new instrument was created for each category, with only one exception, where this was not possible due to serious statistical problems. The four new instruments were distributed to a total of 489 subjects and subsequently they were analyzed in the same way, in order to examine their improvement. All of them appeared to be improved, each one at a different degree. In this way, it was once again highlighted how IRT can be of great importance at the development of instruments in the area of Physics research (and furthermore of educational research). The implementation of this theory can lead to more accurate measurement instruments and consequently to more accurate measurement and conclusions.

Περίληψη

Η εκπαιδευτική μέτρηση στοχεύει τυπικά στην αξιολόγηση των ικανοτήτων και των γνώσεων των μαθητών σε διάφορους τομείς όπως τα μαθηματικά, η γλώσσα, η φυσική κλπ., χρησιμοποιώντας τεστ, ερωτηματολόγια και άλλα όργανα. Ο σκοπός αυτής της εργασίας είναι να δημιουργήσει έγκυρα όργανα που μετρούν την ακαδημαϊκή ικανότητα στη Μηχανική μέσω της βαθμονόμησης ήδη χρησιμοποιημένων. Ένα όργανο που χρησιμοποιήθηκε για παρόμοιους σκοπούς σε άλλες μελέτες, αναλύθηκε ξεχωριστά για κάθε κατηγορία του δείγματος, προσαρμόζοντας τα δεδομένα σε μοντέλα της Θεωρίας Απόκρισης Ερωτήματος (ΘΕΑ) χρησιμοποιώντας το στατιστικό λογισμικό STATA 14. Με βάση τα αποτελέσματα της ανάλυσης, ένα νέο όργανο δημιουργήθηκε για κάθε κατηγορία, με μόνο μια εξαίρεση, όπου αυτό δεν κατέστη δυνατό λόγω σοβαρών στατιστικών προβλημάτων. Τα τέσσερα νέα ερωτηματολόγια μοιράσθηκαν σε ένα δείγμα 489 υποκειμένων συνολικά και στη συνέχεια αναλύθηκαν με τον ίδιο τρόπο, έτσι ώστε να εξεταστεί ο βαθμός βελτίωσής τους. Όλα τα όργανα παρουσίασαν βελτίωση, το καθένα σε διαφορετικό βαθμό. Μέσω αυτού, αναδεικνύεται η σημαντικότητα της ΘΕΑ στην κατασκευή οργάνων μέτρησης στον τομέα της έρευνας της Φυσικής (και κατ’ επέκταση της εκπαιδευτικής έρευνας). Η εφαρμογή αυτής της θεωρίας μπορεί να οδηγήσει σε πιο ακριβή όργανα μέτρησης και κατά συνέπεια σε πιο ακριβή μέτρηση και συμπεράσματα.

1. Introduction

Educational measurement typically aims at evaluating the abilities and knowledge of students in various fields such as mathematics, language, science, physics, etc. To this aim, tests, questionnaires and other instruments are constructed. Ideally, we would like to create valid and reliable instruments.

To achieve this, the classical test theory (CTT) and the item response theory (IRT) can be of great help. These theories are widely used in the area of psychometrics and have also a wide range of applications on the educational research as well. We typically consider ability not to be directly measurable. It is a latent variable and the observed items of the instrument serve as manifestations of ability. The IRT approach focuses more on the items and provides information about their relationship with the variable we would like to measure and the amount of information they provide regarding this variable. In this way, IRT gives researchers the opportunity to decide, based on the results, whether an item is actually worth including in an instrument.

Simsek (2016) argues that most teachers and trainers are still not capable of developing good achievement tests at any area of learning, which is also supported by his findings in the literature. A main reason for that seems to be the inadequate training (Hills, 1991; O'Sullivan & Chalnick, 1991; Zhang & Burry-Stock, 2003, as cited in Simsek, 2016). As a matter of fact, approximately 60% of the test items the teachers used, had mistakes that needed to be corrected or improved before administration (Simsek, 2016). Although IRT has a lot to offer in this direction, only a few physics education studies actually employ this theory (e.g. Ding & Beichner, 2009; Lee et al., 2008; Marshall, Hagedorn, & O'Connor, 2009; Pek & Poh, 2000; Planinic, Ivanjek, & Susac, 2010; Wang & Bao, 2010, as cited in Wallace & Bailey, 2010). Also, in the area of mathematics education in general, Callingham and Bond (2006, as cited in Long et al., 2011) observe that relatively few studies use statistical methods and tools, with the balance in favor of qualitative methods, which is somewhat surprising for a mathematics research community.

We found only a small number of studies written in Greek that implement IRT in the area of education. In these cases, the theory was applied to improve a self-evaluation tool for a Learning Management System (Φωτάρης, 2011) or for the personalized assessment of the learners for the development of an adaptive and intelligent web-based educational system (Hatzilygeroudis et al. 2006). Specifically, in the area of Physics education in Greece, we could not find any

studies that calibrate and validate instruments using this theory. For this reason, this study is focusing on providing new data in this field and highlighting the potential of IRT implementation in educational research, especially in the area of Mechanics.

In addition to the lack of relevant research, mechanics (as well as Physics in general) has some special and interesting attributes, that make research in this field very important. It is one of the oldest academic disciplines and it also relates to many others such as chemistry, engineering etc., as well as to the everyday life. The concepts Physics is dealing with, are mostly abstract and difficult to measure directly, thus making it challenging to evaluate the actual level of academic ability of the students. For these reasons, we consider any contribution to this direction of great importance.

Κώτσης (2011) explored how misconceptions in the area of physics, and more specifically in mechanics, change depending on the person's age. The field of mechanics was chosen because other fields of physics are using the concepts and the laws of mechanics to define themselves (Carson and Rowlands, 2005, as cited in Κώτσης 2011) and also concepts such as weight, force and mass are some of the most basic and widely known. This instrument was also used in past studies for the same reason (Κώτσης & Βέμης, 2002, Κώτσης & Κολοβός, 2002, Κώτσης, 2004, as cited in Κώτσης 2011) and was distributed to primary school, junior high school, senior high school and university students, as well as to primary school teachers. About 200 subjects from each of the aforementioned categories took part in the survey, leading to a total of 1032 subjects. The instrument was used to explore how misconceptions (or correct/false answers) vary across the different age groups.

Initially, it was not used to measure academic ability/knowledge. At the final part of the research, the percentage of right answers to each question was analyzed in relation to the age. Our aim is to use this questionnaire to measure the actual academic ability in the domain of mechanics by taking into account the scores achieved and to evaluate the reliability of the instrument and decide on the right subset of items that are needed. Since the author of this study agreed and provided the data, we fit IRT models using Stata 14.

In section 2, we present the theoretical context involved in this study. The aim and hypotheses of the study are defined in section 3 and the methodology followed is described in section 4. The results of the statistical analysis are presented in section 5, separately for each category of the sample. Finally, in

section 6, we sum up the results, extracted conclusions and make suggestions for further research.

2. Theoretical Context

2.1. Educational Measurement

Educational Measurement is used to gain insight into different kinds of abilities and knowledge that students possess, by obtaining and analyzing scores derived from educational assessments.

"Measurement is the assigning of numbers to individuals in a systematic way as a means of representing properties of individuals. Numbers are assigned to the individuals according to a carefully prescribed, repeatable procedure." (Allen & Yen, 1979)

The typical aim of educational measurement is to evaluate the abilities and knowledge of students in various fields such as mathematics, language, science, physics, etc. The means to achieve such thing are tests, questionnaires and other instruments that are constructed for this purpose. Most of the times, the outcomes of such instruments are total scores, which are analyzed and interpreted in order to assign characteristics to the students. A great amount of attention is focused on the reliability and validity of these instruments, so that they actually measure the respective attribute.

2.2. Classical Test Theory (CTT)

CTT is a psychometric theory that allows the prediction of outcomes of testing, such as ability of the test-takers and difficulty of items (Alagumalai & Curtis, 2005). Spearman's work in 1904 is considered to be the first work in Classical Test Theory (Traub, 1997).

Many constructs such as abilities and attitudes that are common in education, psychology, medicine and in the social sciences, are not directly observable and thus not directly measureable. For example, one cannot directly and objectively measure the pain level that a patient feels or the academic ability of a student in a specific subject (e.g. Physics). But also in the case of measureable variables, one can only get the results of the impact of a phenomenon on a measurement instrument, and not the phenomenon itself. CTT comprises a set of principles that allow us to determine how successful our proxy indicators are at estimating the unobservable variables of interest (DeVellis, 2006). It concerns using observable information (such as scores on a

questionnaire items) to gather insights into variables (such as patient satisfaction) that cannot be directly observed (DeVellis, 2006). The aim of CTT is to explain and improve the reliability of an instrument.

2.2.1. CTT Terminology

True score

The main principle that lies on the very foundations of CTT is that the observed score equals the actual state of the unobservable variable of interest plus error contributed by all other influences on the observable variable. The actual state of the unobserved variable is its hypothetical true score (DeVellis, 2006).

The total observed score $O(x)$ is equal to the true (latent) score $T(x)$ and the error E associated with the item.

$$O(x) = T(x) + E$$

For example, in physics, one cannot measure directly the academic ability of a student in this subject. Using a questionnaire with a series of questions, we expect them to reflect their knowledge on these questions.

Random Error

Under CTT all errors are assumed to be random and not correlated with the true score or the observed score. That means that these errors are assumed to be as likely to increase or decrease the observed score for this item and are also assumed to be independent from one another. The error associated with each item is unique to that item. Since errors are random, they have a zero mean value. That means that when all errors are combined, they should cancel each other out and have little or no effect on the item mean. But on the other hand, error will increase item variability.

Item Reliability

A reasonable question, when it comes to measuring instruments, is in which degree it is reliable. In other words, how accurate it is in measuring the true score. A good instrument should include items that provide us with scores close to the true score. This implies that the true score and the observed score

are correlated and that a suitable index of the association between the true score and the observed score would provide one key piece of information about how good an indicator the latter is (DeVellis, 2006). For this purpose, we conventionally use the square of the correlation coefficient as a means of representing the proportion of variance shared between the two variables (DeVellis, 2006). If we could correlate an item's score with the true score, and then square that correlation, we would have a very useful piece of information about how well the item served as a proxy for the true score; that is the proportion of the item's variation that was shared with the true score. Under CTT, that proportion is defined as the item's reliability.

Item Discrimination

Discrimination is the ability of a test item to differentiate individuals who rank high on the latent scale from those who rank low.

The fact that CTT is relying on inter-item correlations to establish item reliability, reveals that items more strongly correlated with each other are also more strongly correlated with the true score of the unobserved variable of interest and, thus, are fundamentally better items with greater discrimination (DeVellis, 2006). Discrimination is essentially an item's strength of association with other items and thus, presumably, with the true score (DeVellis, 2006). An item that correlates strongly with the set of unidimensional items as a whole can more sharply discriminate between those who score low and those who score high on the entire set of items (DeVellis, 2006). The item discrimination index is the correlation coefficient between the scores on the item and the scores on the total test (Alagumalai & Curtis, 2005). Theoretically, when one removes items that are not good discriminators, the test reliability should be increased.

Difficulty

This term and attribute of CTT is a result of the expansion of the theory in the area of educational testing. The difficulty of an item is quantified as the probability of an average subject ($\theta=0$) to endorse the item correctly. That means that the more difficult an item is, the fewer people can answer it correctly.

In the case of educational testing, an individual answering correctly a "difficult" knowledge question implies that this individual possesses a relatively higher degree of knowledge in this area. Important point regarding an item's

difficulty is that it is defined by how some group of people have answered that item in comparison to how they answered other items (DeVellis, 2006).

Scale reliability

Reliability is the property of a set of test scores that indicates the amount of measurement error associated with the scores (Frisbie, 1988). Having in mind that the observed score consists of the true score and an error term, which is randomly distributed with a mean close to zero, one can conclude that the more items a scale has the more reliable this scale is. This occurs because the error has a smaller effect on the average score, as the errors of more items are more likely to cancel each other out. One of the most well-known indicators of scale reliability is Cronbach's coefficient alpha and it is driven by the correlations among the items and the number of items (DeVellis, 2006). The greater the items correlate with each other the more they share something in common and thus the more they reflect a common true score. For a set of items, the coefficient alpha, under the assumptions of CTT, quantifies the proportion of variance that reflects the true score of the variable to which the items are related (DeVellis, 2006). Because coefficient alpha is determined both by the number of items and the strength of the correlations among those items, increasing either of those influences will be expected to increase reliability (DeVellis, 2006).

The improvement of a scale's reliability in CTT can be achieved with increasing the number of items, deleting items that do not discriminate or that are imprecise, identical test conditions for all examinees and explicitly stated, objective-type questions and heterogeneous group (Alagumalai & Curtis, 2005).

2.2.2. Advantages and Disadvantages of CTT

An important advantage of CTT, which makes it popular among researchers, is the familiarity with its basic concepts. Most of the scales used in measurement theory have been developed based on the principles of CTT. There are also many known and easy to use programs, for performing CTT.

Another advantage is that this model fits commonly used instruments very well. All items are supposed to be equally good at measuring the true score of a variable. Adding all these scores coming from each item, the effect of the

error weakens. CTT has been widely used in the social sciences because the data of interest often fit this pattern (DeVellis, 2006).

Another important advantage is that the items do not have to be perfect. Even items that do not relate that strongly to the latent variable can be still used to measure a score. This property of CTT is really useful, as one can create better measuring instruments just by adding more items. And of course, this is something much easier than improving the items or finding better ones.

Although the last mentioned property of CTT provides us with a great advantage, on the other hand it also generates some problems. The scales are usually long and items often seem quite similar and the effort to develop items that correlate strongly with each other can result in superficial similarities (DeVellis, 2006). If this happens then not only the variable of interest but also other characteristics may be common among the items. In this case the true score becomes a mixture of all these characteristics and the variable of interest. But in this way the scale does not accurately measure what is supposed to measure. CTT methods have difficulty differentiating between common themes across items that are important to the variable of interest and common themes of this more superficial type (DeVellis, 2006).

Another disadvantage is that parameter estimates under CTT depend on the sample of individuals studied (DeVellis, 2006). In other words, different samples with different variances will not give the same data. That means that CTT is probably not the best theory for comparing different populations.

One could say that CTT is not the most proper theory when it comes to educational measurement. As mentioned before, making comparisons between different tests given to different populations for the measurement of the same latent variable can lead to false conclusions. The item's difficulty and discrimination are dependent on the group that they are administered and because of that are not helping in making general conclusions about the items and for the scales. Another problem is that observed and true scores are test dependent, thus, rising and falling when there are changes in test difficulty (Alagumalai & Curtis, 2005). The assumption of equality of errors for all examinees is not realistic as the measurement is more precise for students with an average ability and less for those on high and low ability (Alagumalai & Curtis, 2005).

2.3. Item Response Theory (IRT)

IRT is a theory used in the design, analysis, scoring, and comparison of tests and similar instruments whose purpose is to measure unobservable characteristics of the respondents (StataCorp, 2015).

"Item response theory is a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test. Item responses can be discrete or continuous and can be dichotomously or polychotomously scored; item score categories can be ordered or unordered; there can be one ability or many abilities underlying test performance; and there are many ways (i.e., models) in which the relationship between item responses and the underlying ability or abilities can be specified." (Hambleton & Jones, 1993).

IRT models are used extensively in the study of cognitive and personality traits, health outcomes, and in the development of item banks and computerized adaptive testing (StataCorp, 2015). Some examples of applied work include measuring computer anxiety in grade school children (King and Bond 1996, as cited in StataCorp, 2015), assessing physical functioning in adults with HIV (Wu, Hays, Kelly, Malitz, and Bozzette 1997, as cited in StataCorp, 2015), and measuring the degree of public policy involvement of nutritional professionals (Boardley, Fox, and Robinson 1999, as cited in StataCorp, 2015).

IRT methods primarily appeared around the 50's (Tucker, 1946; Lord, 1952, as cited in Harvey & Hammer, 1999). The beginning of IRT is often traced to Lord and Novick's (1968, as cited in Embertson & Reise, 2000) classic textbook entitled 'Statistical Theories of Mental Scores'. But it is only during the last years that IRT started to become more popular. The main reason for that was that the models of this theory required many and difficult computations. But since computers started to make these computations relatively easy and accessible to almost every researcher, IRT's popularity rose up. This is also the main reason why CTT is more known and popular compared to IRT. As it will be explicitly reasoned in the following part of this chapter, IRT offers a variety of solutions to problems of CTT.

CTT offered the opportunity to analyze dichotomous items that could be converted in the form "right-wrong". The first IRT models were also focused on this direction but afterwards, new models have been developed to give theoretically the opportunity to analyze every type of items and assessment instruments. Especially in the area of educational research, the IRT models are popular and well known, and particularly the 1-parameter or Rasch model (Wright, 1977, as cited in Harvey & Hammer, 1999).

2.3.1. IRT Assumptions and Terminology

IRT models have been based on the assumption that the item pool being analyzed is effectively unidimensional (Harvey & Hammer, 1999). That means that the items are manifestations of a unique latent construct. This assumption is not necessarily problematic because a multidimensional instrument includes subsets, which can be separately analyzed using a unidimensional IRT model. In practice, no scale composed of a reasonable number of items will ever be perfectly unidimensional (Harvey & Hammer, 1999).

A term encountered in IRT is the **latent trait (ability)**, which is the unobserved characteristic that is presumed to be responsible for the observed responses that are made to the test's items and is denoted theta (θ) (Harvey & Hammer, 1999). One could say that theta is the corresponding true score of the CTT.

A **homogeneous subpopulation** (HSP) is simply a collection of individuals who are homogeneous with respect to their scores on the underlying construct (θ) being assessed (Harvey & Hammer, 1999).

The Probability of Item Endorsement (PIE) or Probability of a Correct Response (PCR) is defined as the proportion of respondents, in each HSP of interest, giving the correct response to the item (Harvey & Hammer, 1999).

Item Characteristic Curve (ICC)

In IRT one can find different models that describe the relationship between the latent variable (θ) and the response to each item of the test. The difference between these models is the causal relationship that is presumed to exist between (θ) and the observed item response (Harvey & Hammer, 1999). The Item Characteristic Curve (ICC) is a two-dimensional scatterplot of θ on the x-axis by item-response probability (PRC or PIE), depicting the item response that would be expected from an HSP located at any given point on the underlying construct (Harvey & Hammer, 1999). The ICC describes actually the probability people at different ability levels "succeed" on a given item (individual test question) (StataCorp, 2015).

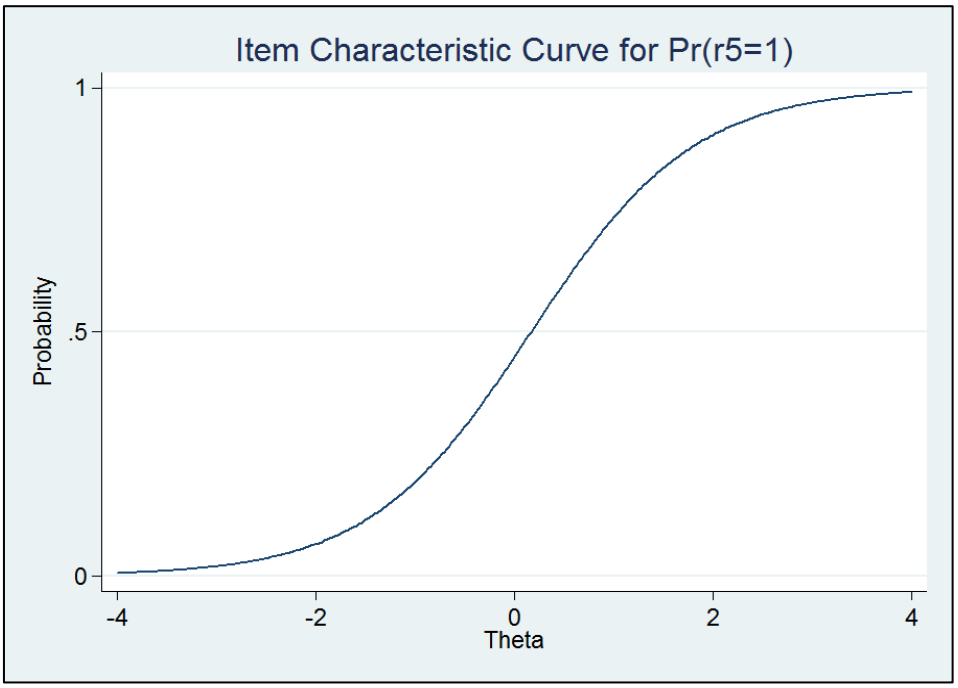


Figure 1: Example of an ICC

Item Information Function (IIF)

Another feature of IRT is the Item Information Function (IIF). IIFs indicate the range over θ where an item is best at discriminating among individuals (Edelen & Reeve, 2007). The term "information" refers to the **reliability** or precision of a whole instrument or (in the case of the IIF) one item (StataCorp, 2015). As a result, in IRT, the amount of information provided by an item or an instrument is corresponding to their level of reliability and precision. An item that provides a lot of information about the latent trait is of course an item with great reliability in measuring it. Items that are supposed to be more reliable measure the latent trait around the estimated difficulty parameter with greater precision (StataCorp, 2015). This feature can be of great use in item evaluation and furthermore in test development. It allows the construction of short forms or tailored assessments, ensuring that the selected subset of items provide adequate precision across the entire range of interest as well as maximizing precision along critical segments of the construct continuum (Edelen & Reeve, 2007).

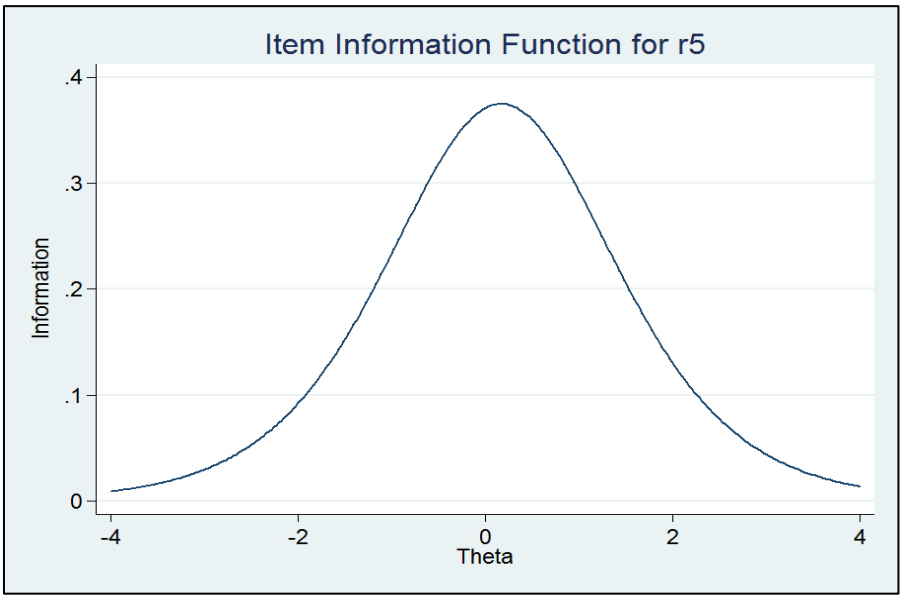


Figure 2: Example of an IIF

Test Characteristic Curve (TCC)

Although, in IRT, the main focus is on the items of a test, there are also features that refer to the whole test. One of them is the Test Characteristic Curve (TCC), which describes the expected score of a HSP of individuals for every θ value. The TCC is the sum of ICCs for the entire instrument and thus plots the expected score on the test along the latent trait continuum (StataCorp, 2015). This score has a minimum value of zero and a maximum equal to the number of the items that are included to the test. That is because every item can get a value of 0 or 1 (right or wrong), thus, making the cumulative test score vary between zero (when no item was endorsed correctly) and the number of the items (in the case that every item was answered correctly).

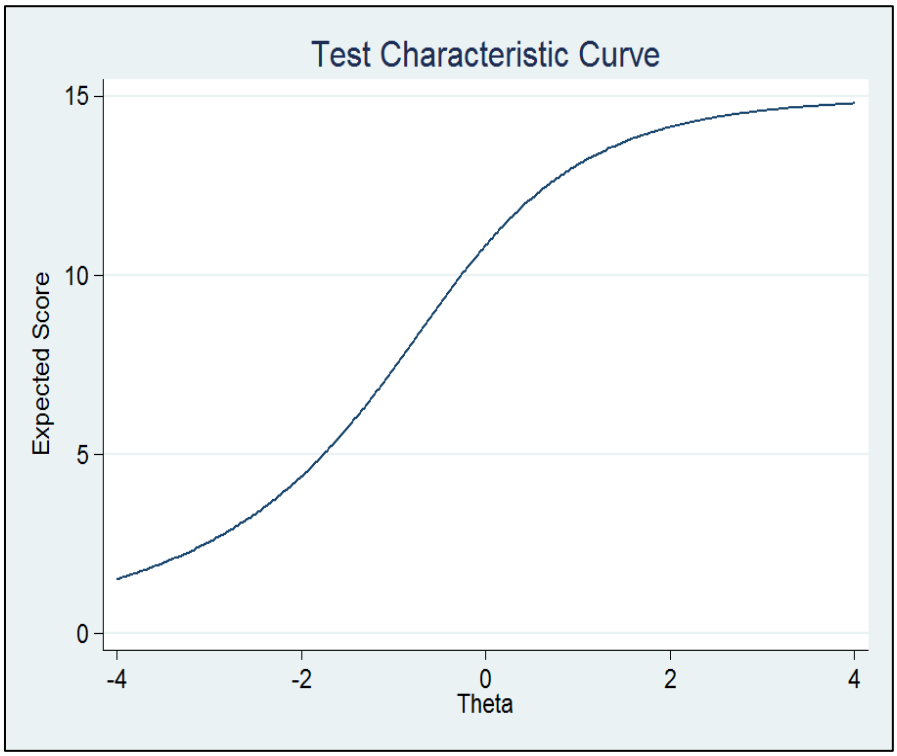


Figure 3: Example of a TCC

Test Information Function (TIF)

Every test item provides a certain amount of information about the latent trait at any ability level θ. The sum of the information from these items, for any given θ, is the amount of information the test provides for every θ and can be described by the Test Information Function (TIF). Through this feature, it is possible to evaluate how well the instrument measures ability across its whole range. The TIF is very useful in test development as it enables the researchers to evaluate how accurate the instrument is and how suitable it is for the purpose it is intended to. According to the definition, that is actually the reliability of the instrument. In Figure 4 the blue line represents the overall test information and the red line is the corresponding standard error. In this case, the test provides the greatest amount of information (with the minimum standard error) for the individuals belonging to the HSP between approximately -1,5 and 0.

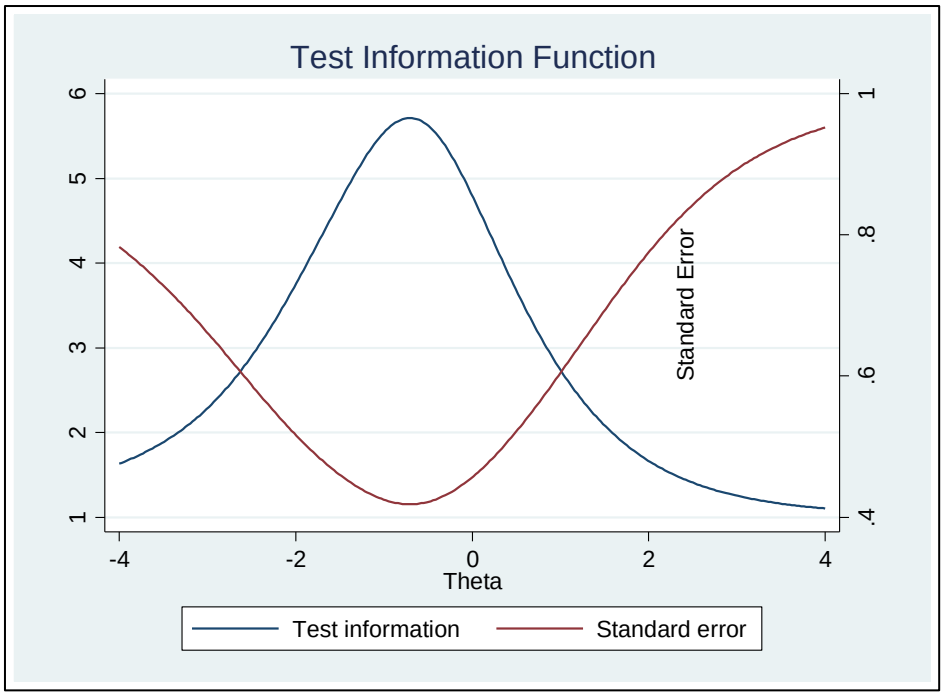


Figure 4: Example of a TIF

2.3.2. IRT Models

All models share common assumptions but they differ in the way they consider θ is affecting the item response. Actually, they mainly differ regarding the number of variables they use to describe the relationship between θ and items' response. As a result, there are several models with 1, 2 or more parameters, as well as models for non-binary items, hybrid models and so on. The most common models, and more appropriate for dichotomous items, are the 1-, 2- and 3- parameter logistic models (Edelen & Reeve, 2007) and we will describe them in the following part, as they are the ones used in this thesis.

1-parameter logistic model (1PL)

Also known as the Rasch model (Wright, 1997) as it was first published by the Danish mathematician Georg Rasch in the 1960s (Baker 2001). It is the simplest model of all, with only one parameter (as stated by the name itself) needed to determine the relationship between θ and the item response. This one parameter is named **difficulty (b)** and it is defined as the score on θ that is associated with a 50% likelihood of correct/endorsed item response (Harvey & Hammer, 1999). While CTT-based parameters lie on a different scale than that used to estimate each respondent's score on the trait in question, IRT-based parameters (among others the difficulty (b) as well) lie on the same scale with θ. And that is an important characteristic of the IRT models because they locate the person parameters (such as θ) and these of the item (such as b) on a common scale (Harvey & Hammer, 1999).

The probability of a person j providing a positive answer to item *i* is given by:

$$\Pr(Y_{ij} = 1|\theta_j) = \frac{\exp\{a(\theta_j - b_i)\}}{1 + \exp\{a(\theta_j - b_i)\}}, \qquad \theta_j \sim N(0,1)$$

where a represents the discrimination common to all items, b_i represents the difficulty of item *i*, and θ_j is the latent trait of person *j* (StataCorp, 2015).

Regarding the ICC's of the items of this model, the only difference is the left-right position of it on the horizontal axis. And this position is determined by the difficulty parameter (b). The form of the functional relationship between θ and the observed response is constant across items (Harvey & Hammer, 1999). Figure 5 shows the ICCs for three different items with b parameters of a value

of -1, 0 and 1. All other parameters (a and c) are the same across the items. For all HSP between the θ values of -3 and 3, the item with the greatest value of b (b=1) is the most "difficult", thus less unlikely to be correctly endorsed. As the b value gets lower (or greater) the less (or the more) difficult the item is.

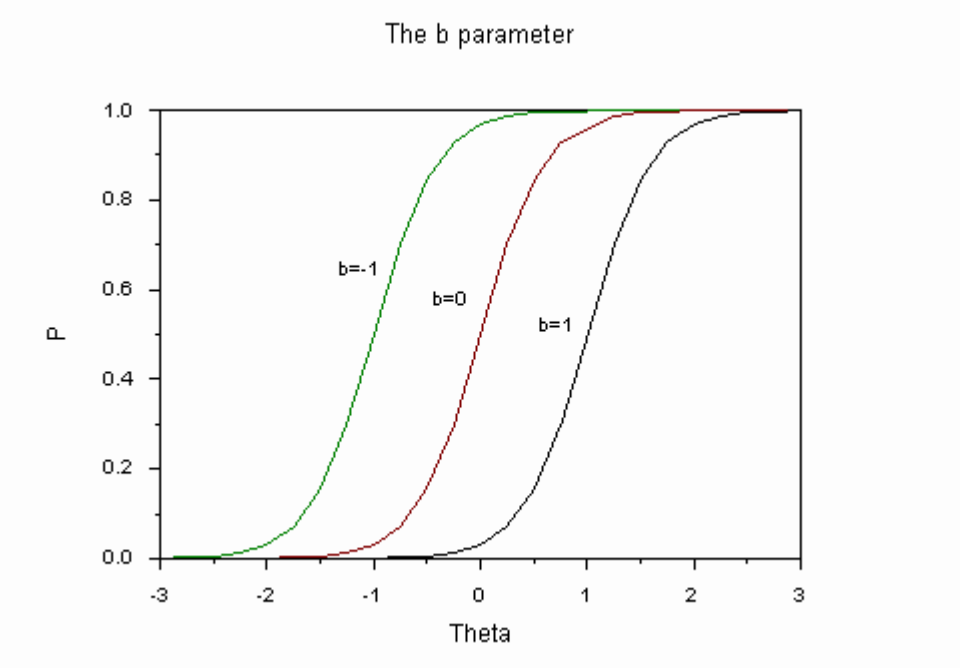


Figure 5: Example of ICCs according to the 1PL model

2-parameter logistic model (2PL)

The 2PL model takes additionally into account the parameter of **discrimination (α)**. This parameter represents the slope of the ICC at the value of the difficulty parameter and indicates the extent to which the item is related to the underlying construct (Edelen & Reeve, 2007). This fact covers a deficiency of the 1PL model, that of the assumption that all items have identically shaped ICCs. Although this might be applicable to an item pool that was really carefully selected from a much larger initial one, it would be though quite unusual in many applied assessment situations (Harvey & Hammer, 1999).

With the input of the parameter of discrimination we are able to obtain information about the strength of the relation between the item and the latent construct. Greater α value means greater relation to the latent variable, thus greater amount of information about the latent variable provided by the item, when all other factors are equal.

The probability of a person j providing a positive answer to item i is given by:

$$\Pr(Y_{ij} = 1|\theta_j) = \frac{\exp\{a_i(\theta_j - b_i)\}}{1 + \exp\{a_i(\theta_j - b_i)\}}, \qquad \theta_j \sim N(0,1)$$

where a_i represents the discrimination of item i, b_i represents the difficulty of item i, and θ_j is the latent trait of person j (StataCorp, 2015).

The ICCs of the items of a 2PL model can have different slopes depending on the discrimination power they have regarding the latent trait. The ICCs of three items with discrimination values of 2, 1 and 0.5 with all other parameters (b and c) remaining the same across them, are shown in Figure 6. For example, in the HSP with individuals that score at θ=1,5 (that means individuals with high level of ability regarding the latent trait), it is expected that the least discriminating item (a=0.5) will be correctly endorsed in a lower rate as the moderately discriminating item (a=1). The same applies for the moderately discriminating item, compared with the most discriminating one (a=2).

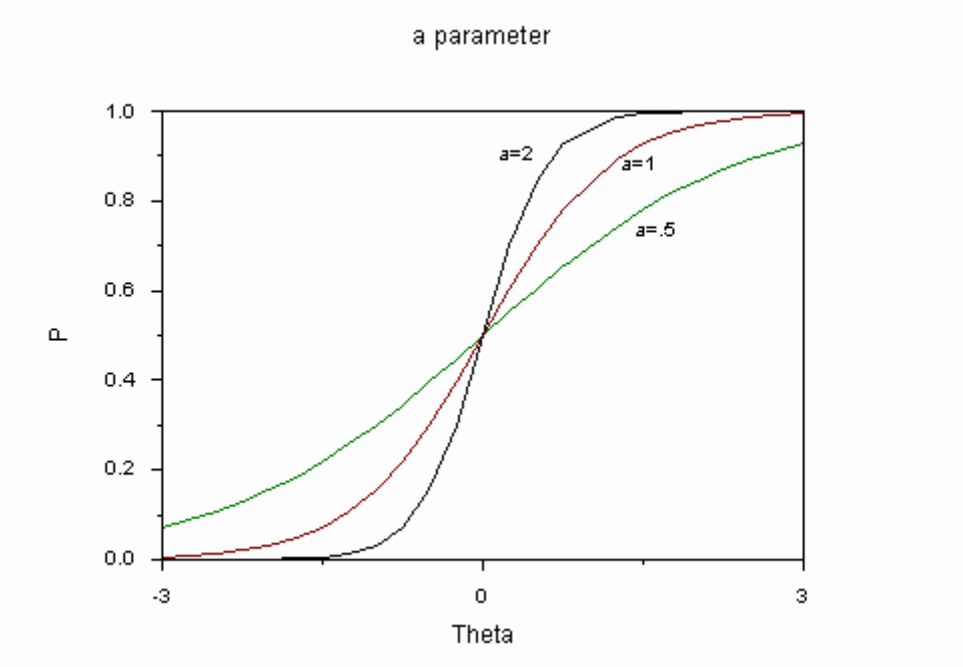


Figure 6: : Example of ICCs according to the 2PL model

3-parameter logistic model (3PL)

The 3PL model introduces one more parameter (c) to reflect the fact that the lower asymptote of the ICC may have non-zero minimum values. All HSP are expected to give non-zero values of correct answers, even to difficult items, because of guessing. So that is the problem that this model solves with this third parameter. It provides us with information regarding the probability of guessing the correct answer of an item.

The probability of a person j providing a positive answer to item *i* is given by:

$$\Pr(Y_{ij}=1|\theta_j) = c_i + (1-c_i)\frac{\exp\{a_i(\theta_j - b_i)\}}{1+\exp\{a_i(\theta_j - b_i)\}}, \qquad \theta_j \sim N(0,1)$$

where a_i represents the discrimination of item *i*, b_i represents the difficulty of item *i*, c_i represents the pseudo-guessing parameter, and θ_j is the latent trait of person *j* (StataCorp, 2015). By default, the c_i are constrained to be the same across all items (StataCorp, 2015).

The greater the c parameter is, the easier it is for a subject to answer this item correctly by guessing and the less informative this item is. Of course the opposite happens as the c parameter is decreased. This can also be seen in Figure 7 where items with different c parameters are presented, while all the other parameters (a and b) remain the same.

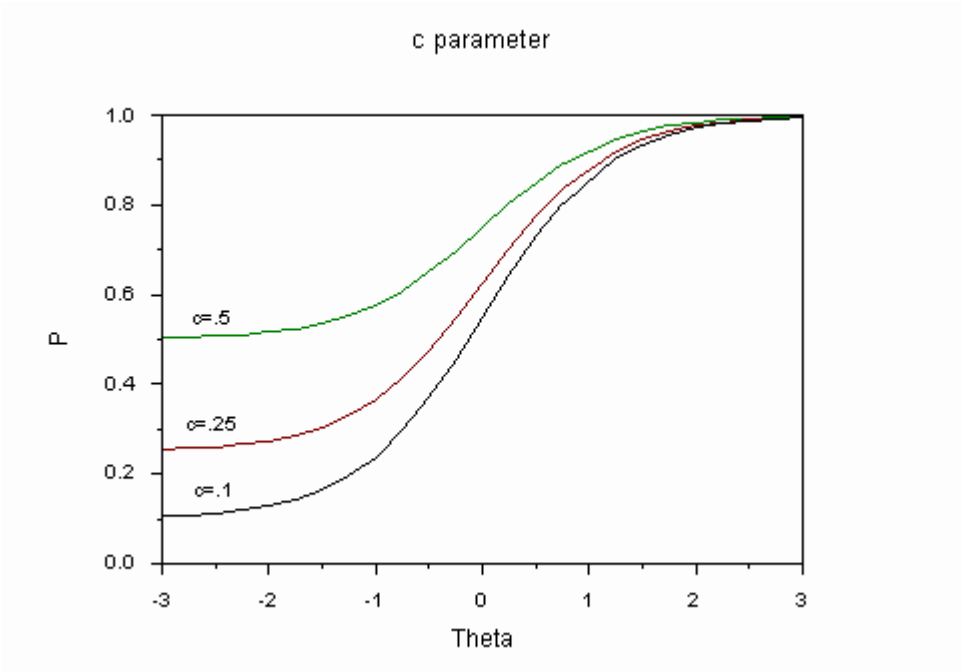


Figure 7: Example of ICCs according to the 3PL model

2.4. Comparing CTT and IRT

Item information

The main focus of IRT is on the items, especially with the ICC providing a great amount of information about each item regarding difficulty, discrimination and the probability of guessing (the three parameters of the models described previously). The CTT is a theory focusing on the test-level. Although CTT can quantify the sample difficulty or discrimination for an item, it lacks an effective means for simultaneously combining and presenting this information in an easily used format (Harvey & Hammer, 1999).

Another difference is related to the amount of information that items are assumed to provide. In the IRT approach, a greater amount of information for the latent construct is produced by items with higher discrimination values and smaller lower asymptote values. Inversely, the greater the standard error, the less the information provided about the θ score. The last fact is also valid for the CTT, where the standard error of measurement is inversely related to the reliability of the test. The difference lies on the fact that in CTT, it is assumed that all the items provide the same amount of information about the latent variable, whereas in IRT it is not necessary. Actually, this is extremely rare, because of different values of discrimination and lower asymptote among the test's items.

Test development and item selection

The IRT provides parameters that help evaluating the quality of an item. Using these parameters, it can be determined whether adding or deleting an item to or from the test can have a positive or negative effect on the instrument's reliability. That means that the improvement of an instrument's reliability can be increased by adding items, as well as by deleting others. This attribute gives to the researcher the advantage of creating better instruments without making them necessarily longer at the same time. But this is not the case in CTT. As it is already described in a previous paragraph, under the CTT, each item is considered to be equally related to the latent construct and providing the same amount of information. As a result, adding more items to an instrument raises its reliability. One disadvantage is that it leads to long scales with many items.

2.5. Test Calibration

Researchers developing instruments have a clear aim about what their instrument and its items should measure. What they do not know is how items and subjects relate to the latent trait that the instrument is supposed to measure. Under IRT, test calibration, is the task to determine the values of the item parameters and examinee abilities in a metric for the underlying trait (Baker, 2001).

2.6. Validity and Validation

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated." (AERA, APA & NCME, 1999, p. 9, as cited in Goodwin & Leech, 2003)

3. Aim and Hypotheses

The aim of this thesis is to create valid instruments that measure the academic ability in mechanics through calibrating already used ones. The instruments are calibrated to various levels of Greek students and primary school teachers.

It is expected that this instrument will be reliable in measuring the academic ability in mechanics and that the new, modified ones will be more accurate in measuring the latent variable of interest.

4. Method

The initial instrument chosen for this study included 28 questions regarding the measurement of physics ability (the latent variable) and some more about personal data of the subjects. The same questionnaire was distributed to all age groups. As mentioned in a previous paragraph, the instrument was also used in past studies for the same reason (Κώτσης & Βέμης, 2002, Κώτσης & Κολοβός, 2002, Κώτσης, 2004, as cited in Κώτσης 2011).

The data collected from the study of Κώτσης (2011), were initially encoded with the number that was corresponding to the answer given by the subjects and not in "right-wrong" format. In order to use the binary IRT models on this data it was necessary to convert the values of each variable. For the right answer the variable value was "1" and for any other wrong answer "0".

Subsequently, the instrument was analyzed separately for each category and a new instrument was created for each category, with only exception, the senior high school students. For this group, it was not able to calibrate the instrument with the available data. It was decided not to distribute the same initial instrument again for this group, because even if we did that, it would be still not possible (because of the difficulty to find more subjects) to calibrate and validate the new adjusted one using a new sample. And since the goal of this study is to calibrate and validate instruments, either ways, we would not be able to achieve this goal.

The statistical analysis of the data was executed with the statistical software "Stata 14 MP". In order to present some descriptive statistics in a better way, the statistical software "IBM SPSS Statistics 21" was implemented as well.

In order to adjust the instrument, it is necessary to define some cut-off criteria for the items. The most useful tool for that is the discrimination parameter. Baker (2001) proposed labels for the item discrimination values as shown in Table 1:

Verbal label	Range of values
None	0
Very low	0.01-0.34
Low	0.35-0.64
Moderate	0.65-1.34
High	1.35-1.69
Very high	>1.70
Perfect	+ infinity

Table 1: Labels for the item discrimination values

A discrimination value of 0.34 or more implies an acceptable level of discrimination power (low or more). It also allows a low or larger level of relation to the latent construct. In this study, we assume that items having a discrimination value of at least 0.34 provide sufficient information about the latent construct. Furthermore, taking into account the discrimination values of all the items, this cut-off value leads to instruments with a sufficient number of items. In this way, we have the chance to evaluate again the items with values close to the limit and determine if it is really worth including them to the instrument. Other greater cut-off values can be, of course, selected, leading to shorter instruments.

Regarding negative values, this is a sign indicating some sort of a problem concerning these items. Negative values give a negative slope to the ICC (Figure 7Figure 8), which actually means that the lower the ability one possesses, the more likely it is to answer the item correctly and vice versa. This is not desirable when seeking items that discriminate the subjects in a proper way. These items provide also only a little amount of information about the latent construct. In Figure 9 we can see the difference between the information provided by such kind of items (q28, q19, q24, q25 and q5) and the information provided by a highly discriminating item (q3). For this reason, items with negative values are the first to be excluded from the new instruments.

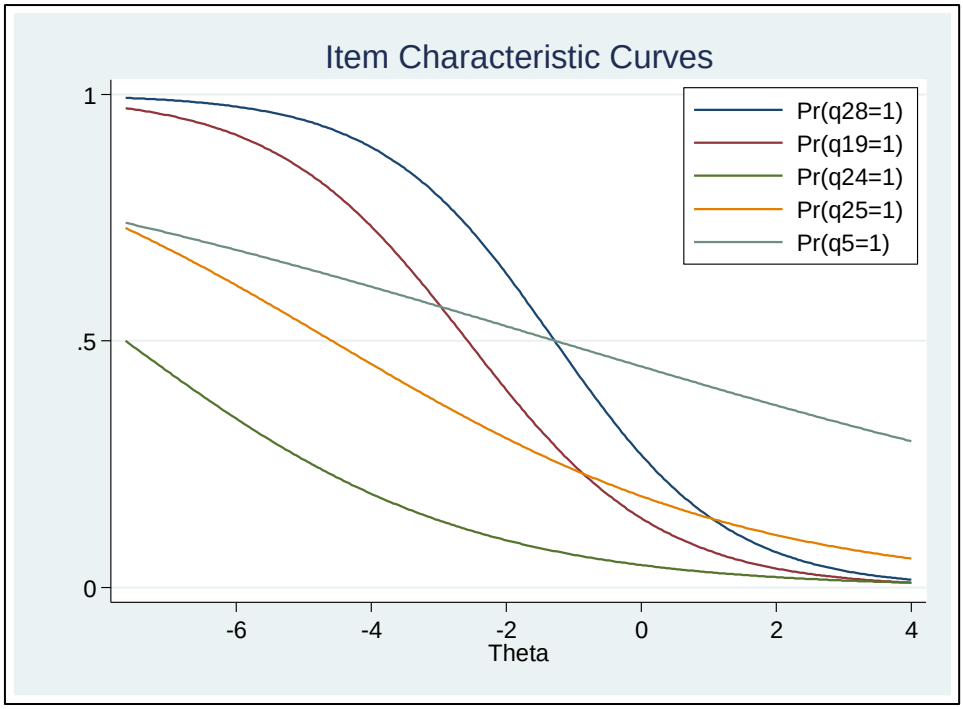


Figure 8: ICCs of items with negative discrimination values

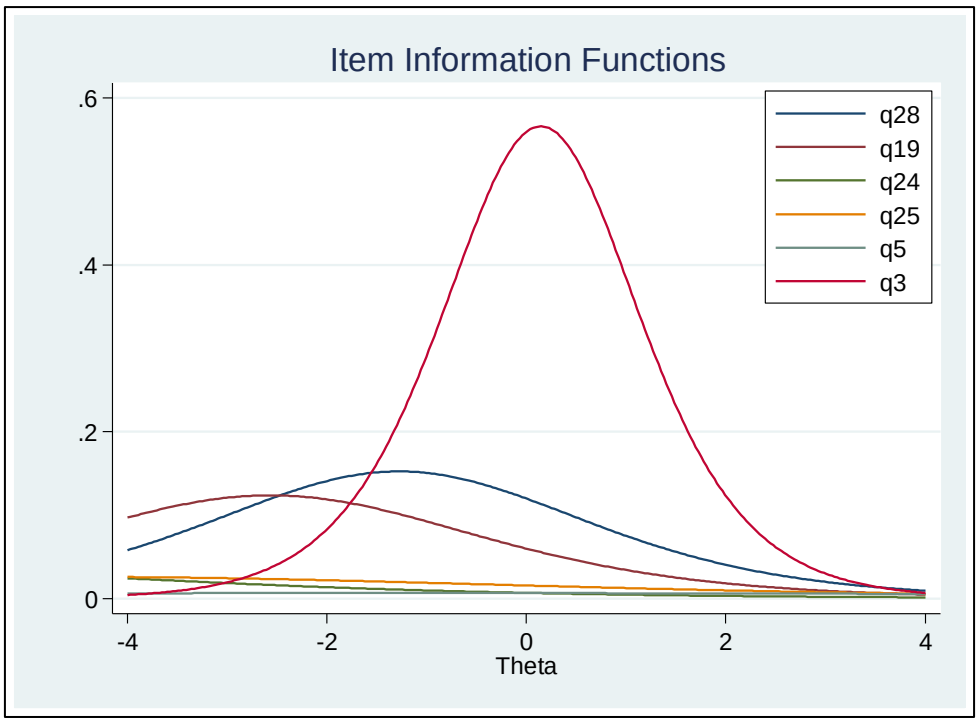


Figure 9: Example of IIFs with different discrimination values

There is no consensus on what is the minimum sample size needed to accurately estimate the parameters of the model. However, there are some guidelines proposed. Sample size should be increased depending on the complexity of the model, but on the other hand, the better the item response data meet the assumptions of IRT, the smaller sample size is needed (Edelen & Reeve, 2007). Some have suggested that 200 or fewer observations can be adequate for models with more than one parameter (Orlando & Marshall, 2002; Thissen et al., 1986, as cited in Edelen & Reeve, 2007). Considering the above, we assume that the observations from the study of Κώτσης (2011) are sufficient for the 2PL model analysis. As the new instruments show a stronger relation to the latent construct, thus supporting the assumption of unidimensionality, and additionally by reducing the number of the items, leading to a less complex model, we can assume that less observations are required.

The four adjusted instruments (presented in Appendix) were distributed to the subjects in different ways. Concerning the primary and junior high school students, the questionnaires were printed out and given to their teachers in order to be filled out in the classroom. Oral and written instructions about how to distribute and fill in the questionnaires were given to each teacher. Regarding the primary school education students, the instrument was assembled using the Google Forms and was distributed online to them. A mixture of online and printed out questionnaires were distributed to teachers of primary and secondary school education that are active in various schools of Greece. In this way, it was possible to collect a greater amount of data about this category. The data from both questionnaires were merged in one dataset and were analyzed regardless of the form (online and printed out).

The sample in this study consisted of a sum of 489 subjects in total, from which, 122 were primary school students, 116 were junior high school students, 155 were primary school education students and 96 were teachers. The instrument was distributed to 60 primary school students of the fifth class and 62 of the sixth, as they are the ones that have a physics class at school. They were all studying in public primary schools in the city of Arta, Greece. Equal number of students (58) from each of the second and third grade of junior high school took part in the study. They all came from public junior high schools of Arta. The students were coming from all the four obligatory years of study and a few were also in the fifth or greater year. The Departments of Primary School Education that they studied were these of the University of Ioannina, the Democritus University of Thrace and the University of Thessaly. From the 96

teachers who took part in the study, 73 were active in the primary school education level and the rest 23 in the secondary, also having various subject specializations. For detailed descriptive statistics of the sample see the Appendix.

5. Results

We analyzed the data from the instrument separately for each group of participants. We applied the 3PL model but the algorithm of the model failed to converge. It is likely that this is happening because the 3PL model is too complex and needs a much bigger sample.

5.1. Primary School

For the primary school students, we used the 2PL model we present the item discrimination and difficulty parameters and their uncertainty in Table 2. We sorted items in ascending order with regards to their discrimination value, which is also the most useful parameter in assessing them. In the upper half of Table 2, under the column “Coef.”, are the discrimination values and likewise in the second lower half the difficulty values. Taking a quick look at the discrimination values, one can see that they are relatively low.

Two-parameter logistic model — Number of obs = 205
Log likelihood = -3239.6297

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q28	-.7808284	.2569323	-3.04	0.002	-1.284406	-.2772503
q19	-.7037032	.294844	-2.39	0.017	-1.281587	-.1258195
q24	-.398033	.4294057	-0.93	0.354	-1.239653	.4435867
q25	-.3233875	.2383914	-1.36	0.175	-.790626	.143851
q5	-.1637992	.1791942	-0.91	0.361	-.5150134	.187415
q8	-.1250254	.2035234	-0.61	0.539	-.5239239	.2738731
q7	-.1132817	.1911899	-0.59	0.554	-.4880071	.2614436
q15	-.0779762	.1927291	-0.40	0.686	-.4557183	.2997659
q16	-.0696315	.193742	-0.36	0.719	-.4493587	.3100958
q2	-.018606	.1867899	-0.10	0.921	-.3847075	.3474955
q23	-.0026474	.1842566	-0.01	0.989	-.3637836	.3584888
q1	.1405604	.1868913	0.75	0.452	-.2257399	.5068606
q9	.2162747	.2169603	1.00	0.319	-.2089597	.6415092
q26	.2320812	.190826	1.22	0.224	-.141931	.6060934
q12	.3414713	.1897511	1.80	0.072	-.0304339	.7133766
q4	.3438353	.2189716	1.57	0.116	-.0853413	.7730118
q18	.5299825	.2219819	2.39	0.017	.094906	.9650589
q21	.5563853	.2469568	2.25	0.024	.0723589	1.040412
q10	.5623709	.4319756	1.30	0.193	-.2842857	1.409028
q6	.5692462	.2843999	2.00	0.045	.0118326	1.12666
q13	.6298698	.217885	2.89	0.004	.202823	1.056917
q22	.7096668	.2253524	3.15	0.002	.2679842	1.151349
q14	.795164	.3017659	2.64	0.008	.2037137	1.386614
q17	.9300987	.2829457	3.29	0.001	.3755353	1.484662
q20	1.036176	.3031325	3.42	0.001	.4420467	1.630304
q27	1.056041	.3297168	3.20	0.001	.4098076	1.702274
q11	1.177304	.3436471	3.43	0.001	.5037684	1.85084
q3	1.504651	.3855123	3.90	0.000	.7490603	2.260241
Diff						
q28	-1.279285	.3981982	-3.21	0.001	-2.059739	-.4988305
q19	-2.569529	.9473944	-2.71	0.007	-4.426388	-.7126702
q24	-7.638587	7.910215	-0.97	0.334	-23.14232	7.86515
q25	-4.577464	3.279963	-1.40	0.163	-11.00607	1.851146
q5	-1.263831	1.613708	-0.78	0.434	-4.426641	1.898978
q8	-5.801028	9.448599	-0.61	0.539	-24.31994	12.71788
q7	4.316595	7.3517	0.59	0.557	-10.09247	18.72566
q15	-9.853923	24.36106	-0.40	0.686	-57.60072	37.89288
q16	-10.38904	28.91842	-0.36	0.719	-67.0681	46.29002
q2	-25.0927	251.9865	-0.10	0.921	-518.9772	468.7918
q23	-130.2567	9065.883	-0.01	0.989	-17899.06	17638.55
q1	4.08252	5.476933	0.75	0.456	-6.652072	14.81711
q9	5.286117	5.249267	1.01	0.314	-5.002256	15.57449
q26	-1.859409	1.613971	-1.15	0.249	-5.022734	1.303915
q12	-.1481369	.4275827	-0.35	0.729	-.9861835	.6899097
q4	2.92193	1.827583	1.60	0.110	-.6600674	6.503927
q18	-1.496601	.6297241	-2.38	0.017	-2.730837	-.2623641
q21	-2.375765	.983697	-2.42	0.016	-4.303776	-.4477547
q10	-5.34576	3.786622	-1.41	0.158	-12.7674	2.075884
q6	-3.084695	1.409678	-2.19	0.029	-5.847613	-.3217761
q13	-.2235491	.2503385	-0.89	0.372	-.7142035	.2671053
q22	.5055624	.2593879	1.95	0.051	-.0028286	1.013953
q14	-2.27629	.7397832	-3.08	0.002	-3.726239	-.826342
q17	-.893846	.2707745	-3.30	0.001	-1.424554	-.3631378
q20	-.9065389	.251645	-3.60	0.000	-1.399754	-.4133237
q27	-1.490846	.3747514	-3.98	0.000	-2.225346	-.756347
q11	-1.243698	.2902487	-4.28	0.000	-1.812575	-.6748209
q3	.1453437	.134393	1.08	0.279	-.1180619	.4087492

Table 2: 2PL model results for primary school

The TCC for this test is shown in Figure 10. As expected, with relatively low item discrimination values, the TCC shows that the test is not so effective in discriminating the subjects. For an ability level of θ=0, the expected score is 14.1. Moving to the right, where the ability level rises, the expected score is rising as well. However, it does not reach great values. After the point of θ=4 the improvement of the expected score is minimal. As we are moving to the opposite side, we notice a greater differentiation but still it remains at a low level.

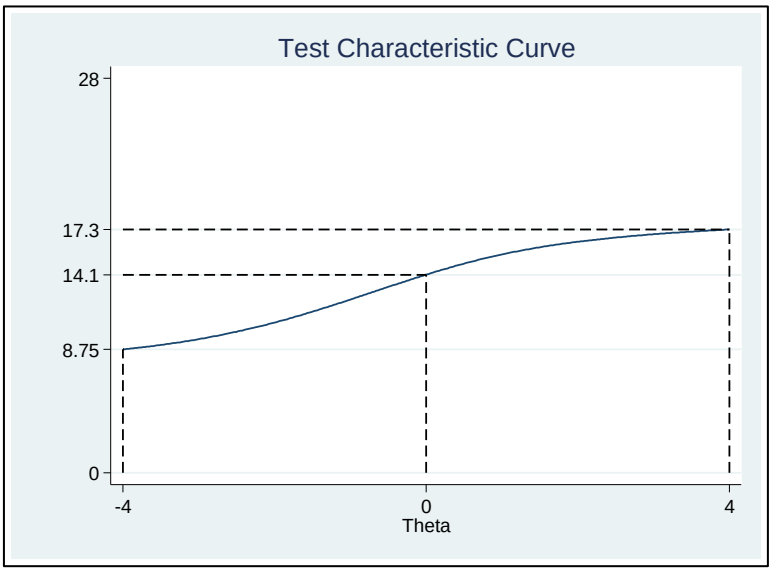


Figure 10: TCC for primary school

As for the TIF presented in Figure 11, we notice that the greatest amount of information is provided for subjects belonging to the HSPs with ability approximately between -1.5 and 0.

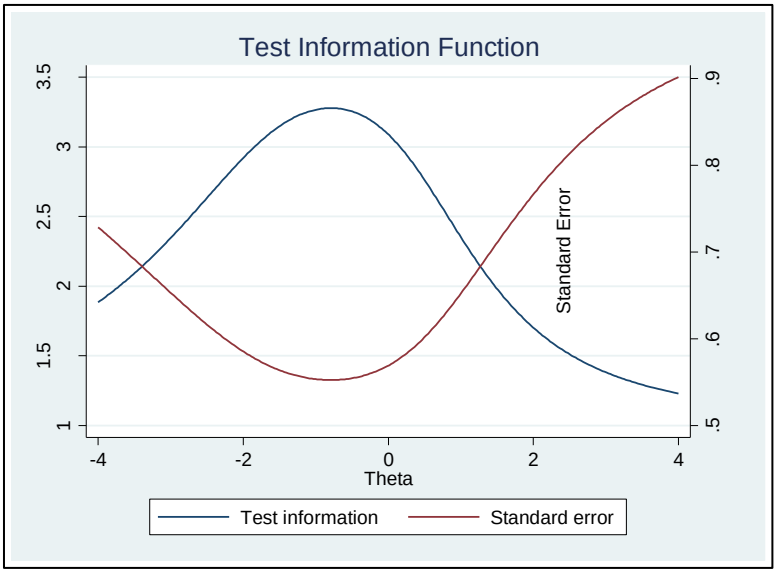


Figure 11: TIF for primary school

As we see, many of the items have negative discrimination values. These items are the first to be excluded from the new instrument. The next step is to evaluate which of the remaining items are appropriate for inclusion in the new instrument. As mentioned in Section 4, we excluded items with discrimination values lower than 0.34. In Figure 12, the ICCs of the first 6 items with discrimination value greater than 0.34 are shown. For items q12, q4 and q10, the z-test and the 95% CI (as shown in Table 2), reveal that their discrimination values are statistically equal to zero. For this reason, we decided to exclude them from the new instrument as well. This fact is also manifested in the ICC and IIF graphs in Figure 12 and Figure 13. The ICCs for these items show that their discrimination power is not enough to discriminate the individuals properly and the corresponding IIFs show that they do not provide sufficient information about the latent construct, especially in compare with items q18, q21 and q6.

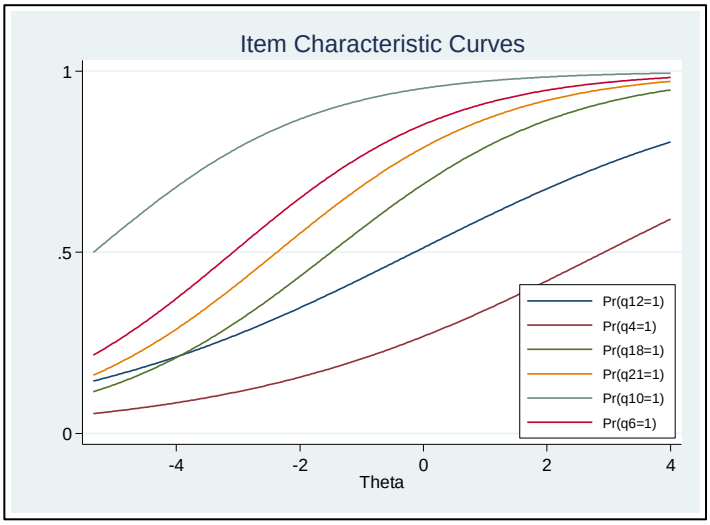


Figure 12: ICCs of items with discrimination close to the cut-off values for primary school

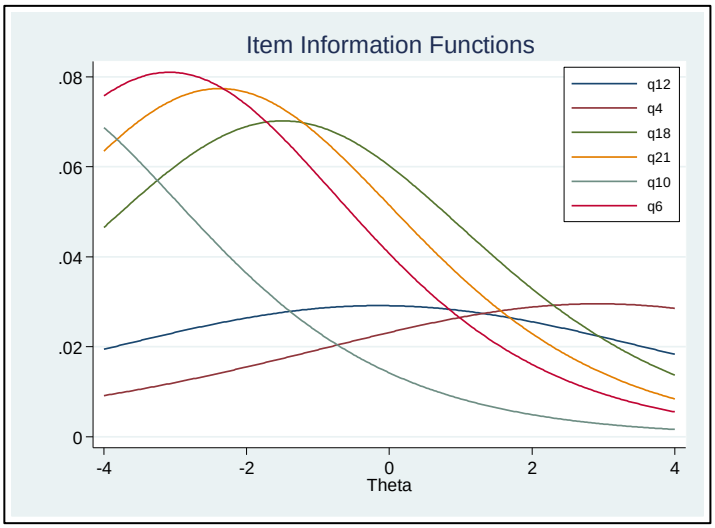


Figure 13: IIFs of items with discrimination close to the cut-off values for primary school

Considering all the above, we excluded the 17 items that have a discrimination value lower than 0.34 and/or their values are statistically equal to zero, from the new instrument for this group. In order to acquire some early information about the new questionnaire before it is distributed, the 2PL model was applied to the remaining 11 items. The results are shown in Table 3.

Two-parameter logistic model — Number of obs = 205
Log likelihood = -1273.7847

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q6	.5408922	.2879078	1.88	0.060	-.0233967	1.105181
q13	.5938382	.219423	2.71	0.007	.163777	1.0239
q18	.6043385	.2397228	2.52	0.012	.1344904	1.074187
q21	.640117	.2608227	2.45	0.014	.1289139	1.15132
q22	.7381882	.2348061	3.14	0.002	.2779767	1.1984
q14	.7545464	.3043833	2.48	0.013	.1579662	1.351127
q17	.8097416	.272536	2.97	0.003	.2755809	1.343902
q27	.903282	.3066492	2.95	0.003	.3022606	1.504303
q20	1.081477	.3174445	3.41	0.001	.459297	1.703657
q11	1.203391	.3691927	3.26	0.001	.4797864	1.926995
q3	1.557053	.4391096	3.55	0.000	.6964136	2.417692
Diff						
q6	-3.228389	1.58215	-2.04	0.041	-6.329345	-.127432
q13	-.2338479	.2649076	-0.88	0.377	-.7530573	.2853616
q18	-1.334736	.5258284	-2.54	0.011	-2.365341	-.3041313
q21	-2.10378	.7826393	-2.69	0.007	-3.637725	-.5698348
q22	.4902143	.2502779	1.96	0.050	-.0003213	.9807499
q14	-2.374901	.8286082	-2.87	0.004	-3.998943	-.7508589
q17	-.9897167	.3305077	-2.99	0.003	-1.6375	-.3419335
q27	-1.667415	.4766457	-3.50	0.000	-2.601623	-.7332062
q20	-.8796411	.2420091	-3.63	0.000	-1.35397	-.4053119
q11	-1.225652	.2928185	-4.19	0.000	-1.799565	-.6517381
q3	.1449005	.1321318	1.10	0.273	-.1140731	.4038742

Table 3: 2PL model results for primary school before distribution

We can see now that the instrument has relatively greater discrimination values. The TCC's slope (Figure 14) is now steeper, implying a greater discrimination ability of the test and also covers a much greater range of the expected score.

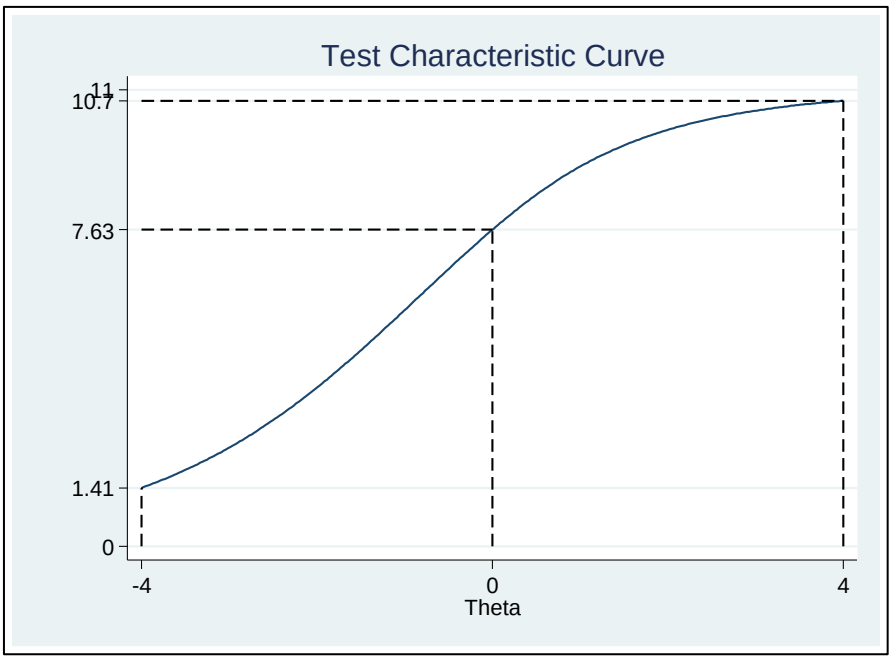


Figure 14: TCC for primary school before distribution

The TIF (Figure 15) is showing only a small differentiation regarding the HSPs where the test provides the most information.

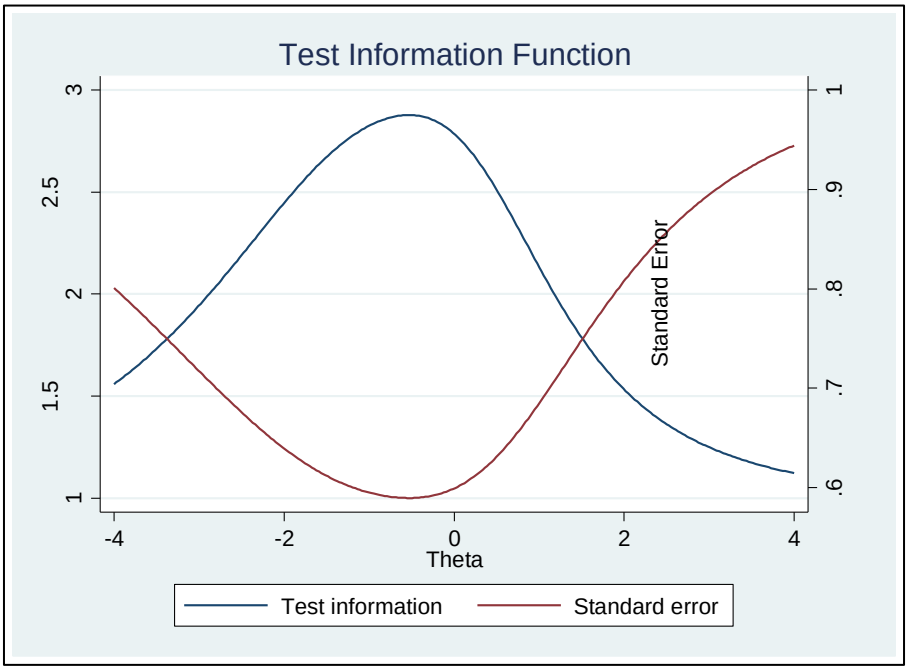


Figure 15: TIF for primary school before distribution

The data collected with the new instrument were analyzed with the same way as before, applying the 2PL model. The results are presented in Table 4.

Two-parameter logistic model — Number of obs = 122
Log likelihood = -673.1182

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q22	-1.170702	.3683766	-3.18	0.001	-1.892706	-.4486967
q11	-.4217926	.2504411	-1.68	0.092	-.9126481	.0690628
q17	1.285276	.4308308	2.98	0.003	.4408634	2.129689
q13	1.316791	.3902181	3.37	0.001	.5519777	2.081605
q3	1.392204	.3629149	3.84	0.000	.6809042	2.103504
q18	1.663953	.5063845	3.29	0.001	.6714579	2.656449
q6	1.727281	.4972936	3.47	0.001	.7526035	2.701958
q20	1.820847	.4678224	3.89	0.000	.9039316	2.737762
q14	2.327351	.6992389	3.33	0.001	.9568677	3.697834
q21	2.787752	1.089273	2.56	0.010	.6528157	4.922689
q27	3.283159	1.223094	2.68	0.007	.885939	5.680379
Diff						
q22	-1.128021	.3090654	-3.65	0.000	-1.733778	-.5222637
q11	2.652601	1.549129	1.71	0.087	-.3836354	5.688838
q17	-1.418504	.3649014	-3.89	0.000	-2.133698	-.7033105
q13	-.8607149	.2380467	-3.62	0.000	-1.327278	-.394152
q3	-.3945068	.1829546	-2.16	0.031	-.7530912	-.0359223
q18	-1.116439	.2446732	-4.56	0.000	-1.59599	-.6368886
q6	-.8563047	.1985017	-4.31	0.000	-1.245361	-.4672485
q20	-.3930182	.1577766	-2.49	0.013	-.7022548	-.0837817
q14	-.8899078	.1733685	-5.13	0.000	-1.229704	-.5501118
q21	-1.218047	.2166311	-5.62	0.000	-1.642637	-.7934584
q27	-.9015287	.1539278	-5.86	0.000	-1.203222	-.5998358

Table 4: 2PL model results of the new instrument for primary school

It is obvious that the discrimination values are now significantly greater, thus making the items more informative (as an example see IIFs in Figure 16).

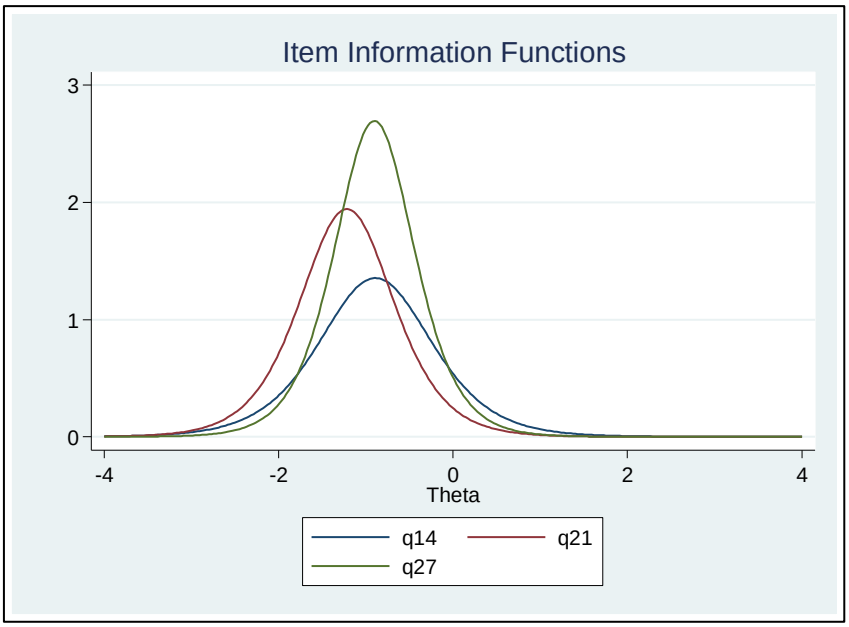


Figure 16: IIFs of the new instrument for primary school

This is also confirmed by the TCC (Figure 17), which is now much steeper than expected, and the TIF (Figure 18), which reveals that the test provides a much greater amount of information about the latent construct.

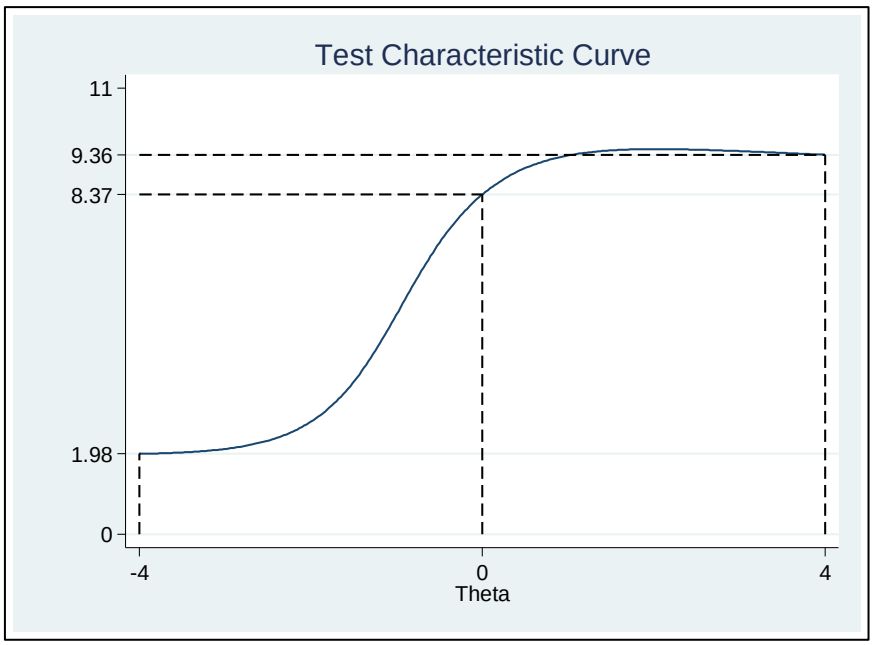


Figure 17: TCC of the new instrument for primary school

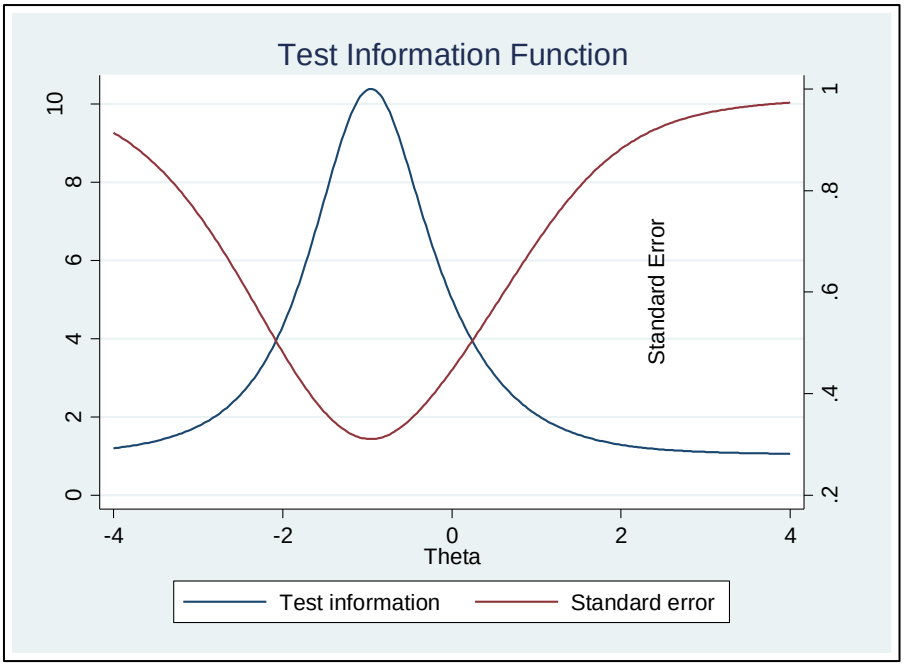


Figure 18: TIF of the new instrument for primary school

In order to observe the variations of the discrimination values through the different phases of the analysis, the following table was created (Table 5Table 14). Values marked with red are the ones that are statistically equal to zero. For phases 1 and 2, the discrimination values are more or less the same. For phase 3, these values are now different, probably because they come from a different set of data. All items seem to work better (by being more discriminating and informative) in the new instrument, except for items q11 and q22. Their discrimination values are now negative and for q11 also statistically equal to zero. Nevertheless, as discussed before, the overall image of the test is improved to a greater level as expected.

	Coef.		
Discrim	Phase 1	Phase 2	Phase 3
q3	1.504651	1.557053	1.392204
q6	.5692462	.5408922	1.727281
q11	1.177304	1.203391	-.4217926
q13	.6298698	.5938382	1.316791
q14	.795164	.7545464	2.327351
q17	.9300987	.8097416	1.285276
q18	.5299825	.6043385	1.663953
q20	1.036176	1.081477	1.820847
q21	.5563853	.640117	2.787752
q22	.7096668	.7381882	-1.170702
q27	1.056041	.903282	3.283159

Table 5: Comparing table for primary school

5.2. Junior High School

The 2PL IRT model failed to converge for junior high school students. Taking a look at the summarizing table of the items (Table 6), question 10 (q10) seems to be correctly endorsed at an almost 98% with the lowest Standard Deviation value of all the others (≈0.14). That means that this item does not variate enough among the individuals of the sample and could possibly cause the problem to the 2PL model analysis. Indeed, executing a 2PL model analysis and excluding this item, we get the results shown in Table 7.

Variable	Obs	Mean	Std. Dev.	Min	Max
q1	196	.6020408	.4907304	0	1
q2	196	.5714286	.4961389	0	1
q3	196	.75	.4341216	0	1
q4	196	.8010204	.400255	0	1
q5	196	.5357143	.5	0	1
q6	196	.9183673	.2745054	0	1
q7	196	.8469388	.3609685	0	1
q8	196	.7346939	.4426267	0	1
q9	196	.9081633	.2895349	0	1
q10	196	.9795918	.141754	0	1
q11	196	.9285714	.2581989	0	1
q12	196	.6836735	.4662329	0	1
q13	196	.8367347	.3705541	0	1
q14	196	.8877551	.3164759	0	1
q15	196	.4489796	.4986638	0	1
q16	196	.6377551	.4818799	0	1
q17	196	.7346939	.4426267	0	1
q18	196	.8061224	.3963465	0	1
q19	196	.3877551	.488486	0	1
q20	196	.6632653	.4738035	0	1
q21	196	.8163265	.3882093	0	1
q22	196	.4489796	.4986638	0	1
q23	196	.6938776	.4620615	0	1
q24	196	.5102041	.501176	0	1
q25	196	.1530612	.3609685	0	1
q26	196	.6530612	.477215	0	1
q27	196	.8163265	.3882093	0	1
q28	196	.5357143	.5	0	1

Table 6: Summarizing table of the items for junior high school

Two-parameter logistic model — Number of obs = 196
Log likelihood = -2807.2808

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q2	-.088726	.187931	-0.47	0.637	-.4570641	.2796121
q27	.0361365	.2460396	0.15	0.883	-.4460922	.5183653
q18	.0446636	.2337296	0.19	0.848	-.413438	.5027653
q19	.1804043	.198086	0.91	0.362	-.207837	.5686457
q17	.1881973	.212991	0.88	0.377	-.2292573	.6056519
q25	.2336194	.2536083	0.92	0.357	-.2634436	.7306825
q5	.2952585	.187735	1.57	0.116	-.0726954	.6632124
q22	.3014983	.1985282	1.52	0.129	-.0876099	.6906065
q21	.3138806	.2346057	1.34	0.181	-.1459381	.7736994
q13	.3342775	.2472495	1.35	0.176	-.1503226	.8188776
q23	.432647	.2173224	1.99	0.047	.0067028	.8585911
q9	.4500853	.3180526	1.42	0.157	-.1732864	1.073457
q20	.4708711	.2137831	2.20	0.028	.0518639	.8898782
q28	.5239202	.2035171	2.57	0.010	.1250341	.9228064
q26	.6933699	.249072	2.78	0.005	.2051977	1.181542
q3	.7076103	.255025	2.77	0.006	.2077704	1.20745
q1	.7532809	.2519033	2.99	0.003	.2595596	1.247002
q14	.8364951	.3167742	2.64	0.008	.215629	1.457361
q15	.843096	.2553719	3.30	0.001	.3425764	1.343616
q12	.9566294	.2922433	3.27	0.001	.3838432	1.529416
q16	.9685704	.2647902	3.66	0.000	.4495911	1.48755
q24	.9792999	.2793112	3.51	0.000	.4318601	1.52674
q4	1.049394	.3161783	3.32	0.001	.4296962	1.669092
q6	1.096883	.3946607	2.78	0.005	.3233619	1.870403
q8	1.521562	.4312733	3.53	0.000	.6762823	2.366843
q11	1.822839	.6627723	2.75	0.006	.5238288	3.121849
q7	2.642052	1.011992	2.61	0.009	.6585843	4.625519
Diff						
q2	3.248713	7.045478	0.46	0.645	-10.56017	17.0576
q27	-41.28974	281.017	-0.15	0.883	-592.073	509.4935
q18	-31.91898	166.9417	-0.19	0.848	-359.1187	295.2807
q19	2.551955	2.877478	0.89	0.375	-3.087798	8.191709
q17	-5.456007	6.138034	-0.89	0.374	-17.48633	6.574318
q25	7.403398	7.910286	0.94	0.349	-8.100478	22.90727
q5	-.494922	.5801087	-0.85	0.394	-1.631914	.6420701
q22	.694266	.6548131	1.06	0.289	-.5891442	1.977676
q21	-4.849843	3.532983	-1.37	0.170	-11.77436	2.074677
q13	-4.999085	3.586089	-1.39	0.163	-12.02769	2.02952
q23	-1.971642	.9864655	-2.00	0.046	-3.905078	-.0382048
q9	-5.271363	3.521049	-1.50	0.134	-12.17249	1.629767
q20	-1.512535	.7091008	-2.13	0.033	-2.902347	-.1227228
q28	-.2908576	.307359	-0.95	0.344	-.8932701	.311555
q26	-1.008346	.3825193	-2.64	0.008	-1.75807	-.2586215
q3	-1.71348	.5720653	-3.00	0.003	-2.834708	-.5922527
q1	-.6184566	.2727331	-2.27	0.023	-1.153004	-.0839095
q14	-2.771388	.8933911	-3.10	0.002	-4.522402	-1.020374
q15	.2785033	.207362	1.34	0.179	-.1279187	.6849254
q12	-.9556553	.283581	-3.37	0.001	-1.511464	-.3998468
q16	-.6977178	.2275345	-3.07	0.002	-1.143677	-.2517584
q24	-.0524558	.1754154	-0.30	0.765	-.3962636	.291352
q4	-1.595803	.3960683	-4.03	0.000	-2.372082	-.819523
q6	-2.618415	.7318872	-3.58	0.000	-4.052888	-1.183943
q8	-.9377589	.204836	-4.58	0.000	-1.33923	-.5362877
q11	-2.019473	.4294356	-4.70	0.000	-2.861151	-1.177795
q7	-1.209978	.1991761	-6.07	0.000	-1.600356	-.8196

Table 7: 2PL model results for junior high school

The discrimination values for this group are higher than the ones for the primary school students. This is an indication that the instrument is more appropriate for junior high school students than for primary school students. This fact is supported from the TCC (Figure 19), which is steeper than the one for the primary school, as well as from the TIF (Figure 20), which is providing generally a greater amount of information and a lower standard error value.

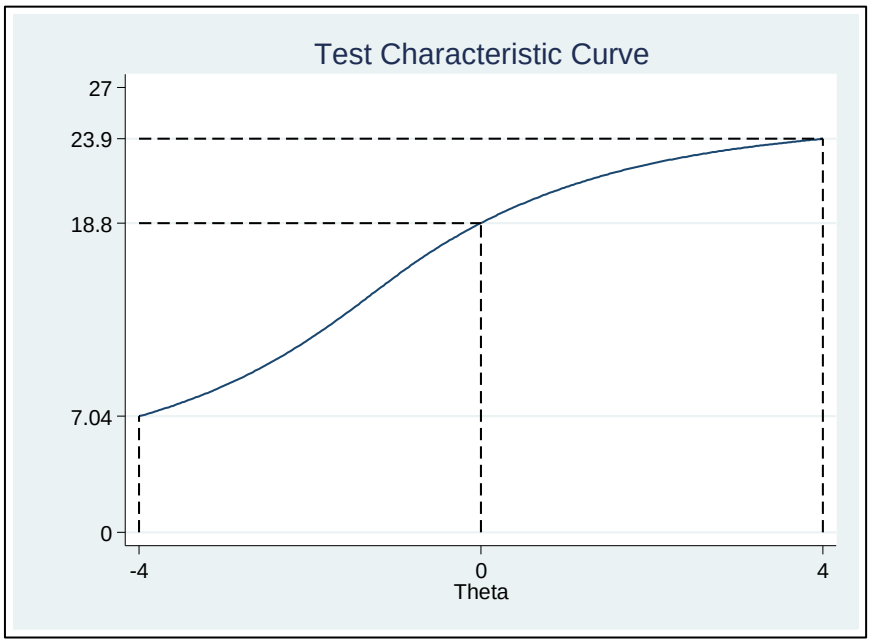


Figure 19: TCC for junior high school

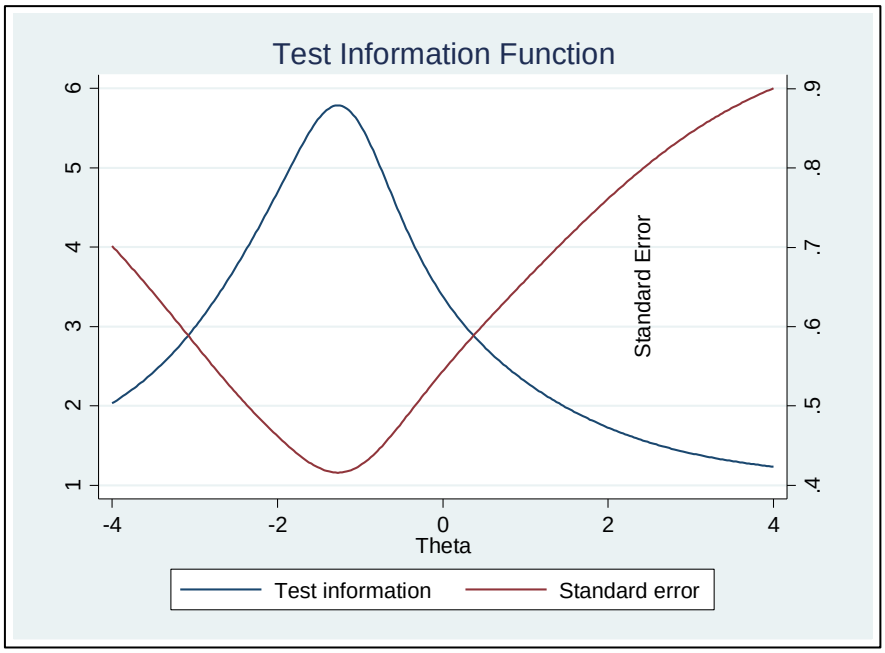


Figure 20: TIF for junior high school

The ICCs and IIFs for items with discrimination values closer to the cut-off value are presented in Figure 21 and Figure 22. It is clear, that items with values lower than the cut-off value (q22, q21 and q13) provide less information as items with greater ones (q23 and q20). The same applies for items whose value is greater than our limit, but also statistically equal to zero (q9). These items may sometimes appear to provide a relatively sufficient amount of information for some HSPs (e.g., q9 provides information for HSPs with $\theta \leq -4$), but since the statistical test are not satisfying we cannot take them into account.

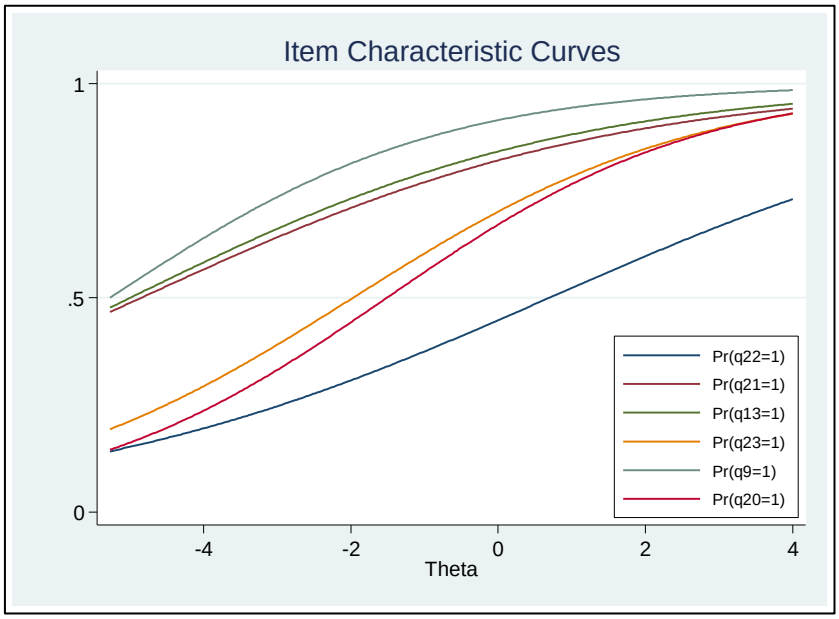


Figure 21: ICCs for items with discrimination values closer to the cut-off value

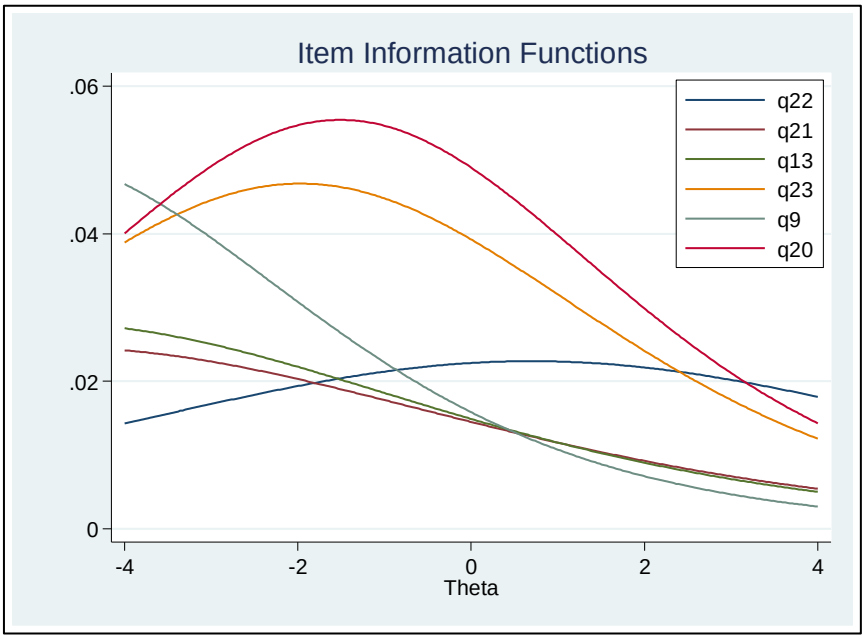


Figure 22: IIFs for items with discrimination values closer to the cut-off value

For the remaining 16 items, the 2PL model was applied. The discrimination values and the TCC are again improved (Table 8 and Figure 23).

Two-parameter logistic model — Number of obs = 196
Log likelihood = -1651.7073

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q23	.3680199	.2114548	1.74	0.082	-.0464239	.7824637
q20	.4078033	.2038378	2.00	0.045	.0082886	.8073181
q28	.5301519	.2036334	2.60	0.009	.1310377	.929266
q26	.6935888	.2310632	3.00	0.003	.2407132	1.146464
q1	.7244879	.2317056	3.13	0.002	.2703534	1.178623
q3	.7658868	.2556089	3.00	0.003	.2649026	1.266871
q14	.787162	.3097417	2.54	0.011	.1800794	1.394245
q24	.8464293	.2481162	3.41	0.001	.3601305	1.332728
q15	.8661604	.2503799	3.46	0.001	.3754249	1.356896
q6	.9801241	.3685978	2.66	0.008	.2576857	1.702563
q16	.9929612	.263316	3.77	0.000	.4768713	1.509051
q12	1.047694	.2963805	3.53	0.000	.4667992	1.628589
q4	1.11271	.3218983	3.46	0.001	.4818014	1.743619
q8	1.585394	.4465834	3.55	0.000	.7101063	2.460681
q11	1.8243	.6560534	2.78	0.005	.5384595	3.110141
q7	2.301246	.762464	3.02	0.003	.806844	3.795648
Diff						
q23	-2.292509	1.314229	-1.74	0.081	-4.86835	.2833315
q20	-1.726012	.8891699	-1.94	0.052	-3.468753	.0167292
q28	-.2871393	.3039509	-0.94	0.345	-.8828721	.3085935
q26	-1.007726	.3660477	-2.75	0.006	-1.725166	-.2902855
q1	-.6372292	.2789486	-2.28	0.022	-1.183958	-.0905
q3	-1.607023	.494284	-3.25	0.001	-2.575802	-.6382442
q14	-2.913407	.9883927	-2.95	0.003	-4.850621	-.9761928
q24	-.0554151	.1954857	-0.28	0.777	-.4385599	.3277298
q15	.2749782	.2017407	1.36	0.173	-.1204264	.6703828
q6	-2.84915	.8640491	-3.30	0.001	-4.542655	-1.155645
q16	-.6846488	.2208134	-3.10	0.002	-1.117435	-.2518626
q12	-.8962409	.2491877	-3.60	0.000	-1.38464	-.4078421
q4	-1.534846	.3601083	-4.26	0.000	-2.240646	-.8290472
q8	-.9190556	.1965037	-4.68	0.000	-1.304196	-.5339154
q11	-2.030578	.4204636	-4.83	0.000	-2.854671	-1.206484
q7	-1.276218	.2110477	-6.05	0.000	-1.689864	-.8625719

Table 8: 2PL model results for junior high school before distribution

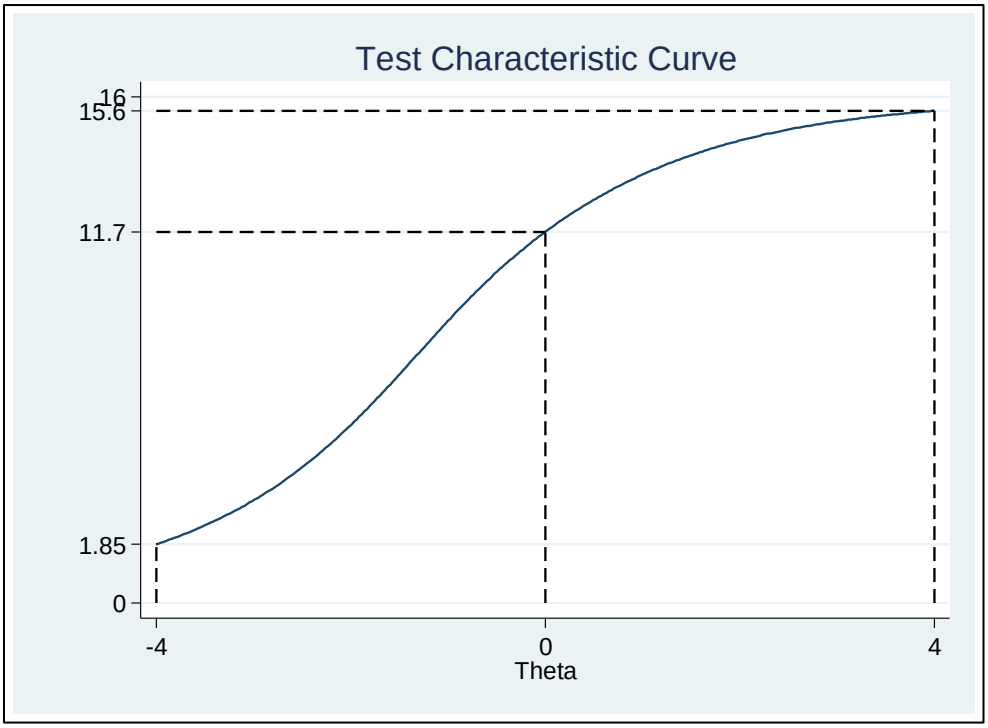


Figure 23: TCC for junior high school before distribution

The TIF (Figure 24) is more or less the same but the indications from this analysis are showing an overall improvement to the test.

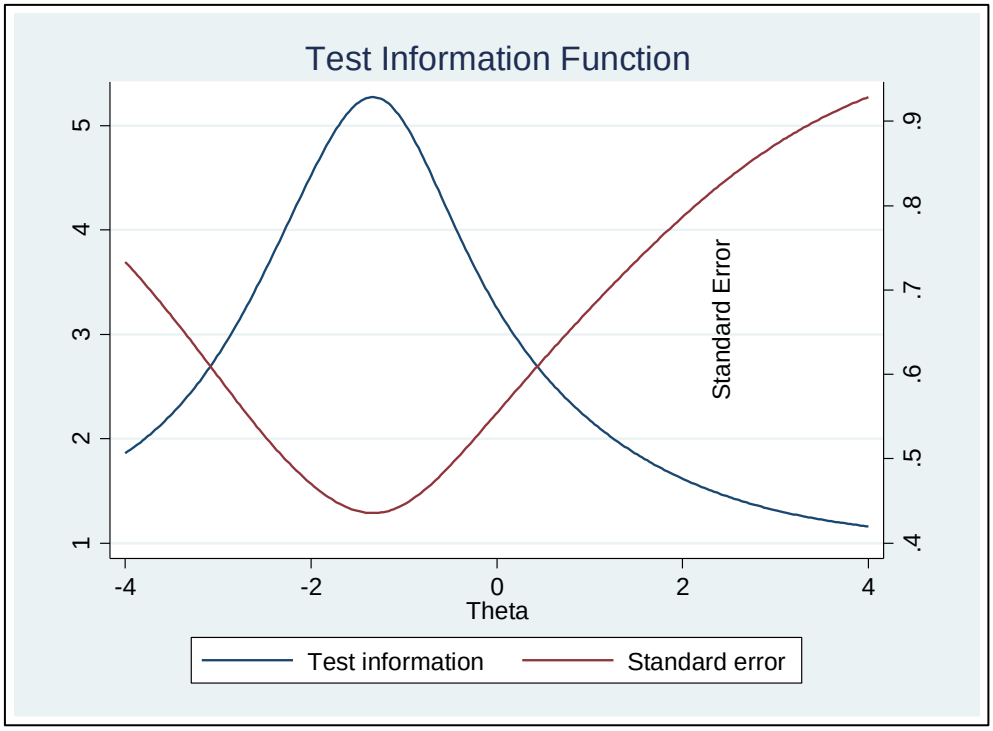


Figure 24: TIF for junior high school before distribution

The descriptive statistics for the data collected from the new instrument for this group can be found in the Appendix. In Table 9 the results of the 2PL model analysis is presented. In this case, the discrimination values are lower than expected, but at least they show an improvement, comparing with these of the initial instrument. The first 7 discrimination values of Table 9 are statistically equal to zero, and most of them at a relatively great significance level. The results from this sample show that these items do not provide enough information about the late construct.

Two-parameter logistic model — Number of obs = 116
Log likelihood = -1080.2378

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q26	.0214732	.2361054	0.09	0.928	-.4412849	.4842312
q23	.1514199	.2306261	0.66	0.511	-.3005989	.6034386
q3	.2418781	.2379733	1.02	0.309	-.224541	.7082972
q14	.2895951	.3109488	0.93	0.352	-.3198534	.8990436
q12	.3905426	.2614737	1.49	0.135	-.1219365	.9030218
q20	.4540692	.2518284	1.80	0.071	-.0395055	.9476438
q6	.5495159	.4073029	1.35	0.177	-.2487832	1.347815
q8	.8433384	.3244152	2.60	0.009	.2074963	1.47918
q28	.8517828	.3089006	2.76	0.006	.2463488	1.457217
q7	.8891533	.3522252	2.52	0.012	.1988047	1.579502
q16	.9265515	.3276608	2.83	0.005	.2843482	1.568755
q24	.9472937	.3437738	2.76	0.006	.2735094	1.621078
q11	1.297576	.5204887	2.49	0.013	.277437	2.317715
q1	1.394766	.4247522	3.28	0.001	.5622671	2.227265
q4	1.805679	.545838	3.31	0.001	.7358565	2.875502
q15	1.96843	.6679049	2.95	0.003	.65936	3.277499
Diff						
q26	-29.89432	328.752	-0.09	0.928	-674.2364	614.4477
q23	.917203	1.854951	0.49	0.621	-2.718434	4.55284
q3	-1.312856	1.480568	-0.89	0.375	-4.214716	1.589003
q14	-5.509936	5.783316	-0.95	0.341	-16.84503	5.825156
q12	-1.908018	1.300653	-1.47	0.142	-4.457251	.6412162
q20	1.304886	.7999052	1.63	0.103	-.2628993	2.872671
q6	-3.970033	2.711129	-1.46	0.143	-9.283749	1.343683
q8	-.0556547	.2545135	-0.22	0.827	-.5544919	.4431826
q28	.5621151	.3044254	1.85	0.065	-.0345477	1.158778
q7	-.7994665	.3435493	-2.33	0.020	-1.472811	-.1261222
q16	1.328489	.4421387	3.00	0.003	.4619125	2.195065
q24	1.538509	.4964402	3.10	0.002	.5655037	2.511514
q11	-1.298648	.386155	-3.36	0.001	-2.055498	-.541798
q1	-.0163665	.1820424	-0.09	0.928	-.3731632	.3404301
q4	.5974554	.1915045	3.12	0.002	.2221134	.9727973
q15	.0341593	.1547344	0.22	0.825	-.2691145	.3374331

Table 9: 2PL model results of the new instrument for junior high school

The TCC (Figure 25) is not as steep as expected, but still steeper and more discriminative than the one of the initial instrument.

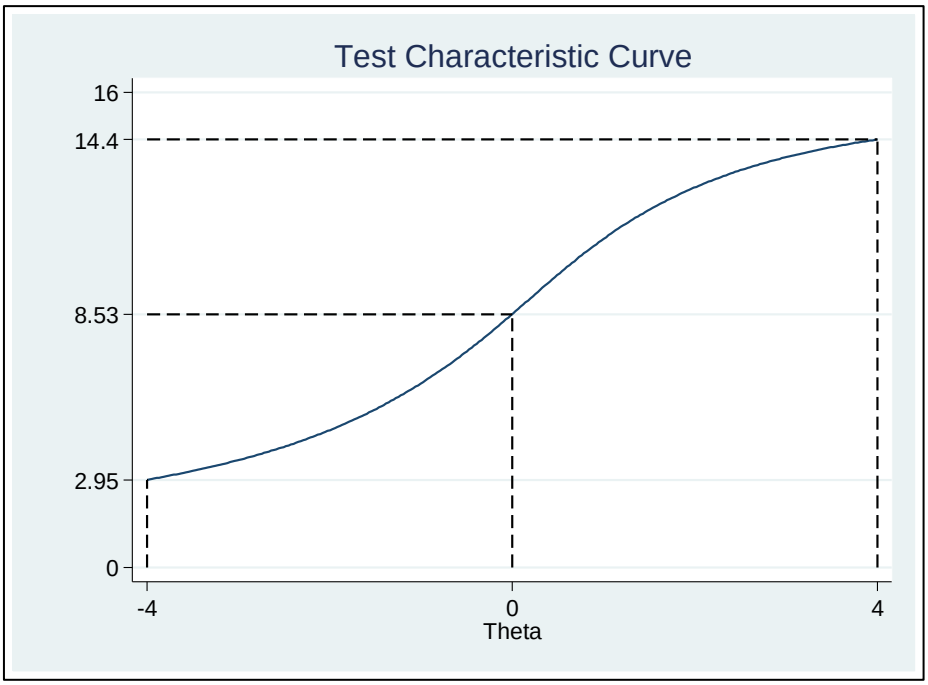


Figure 25: TCC of the new instrument for junior high school

As for the TIF (Figure 26), it is observed that this instrument is now providing more information for HSP with a θ-value between 0 and 0.5.

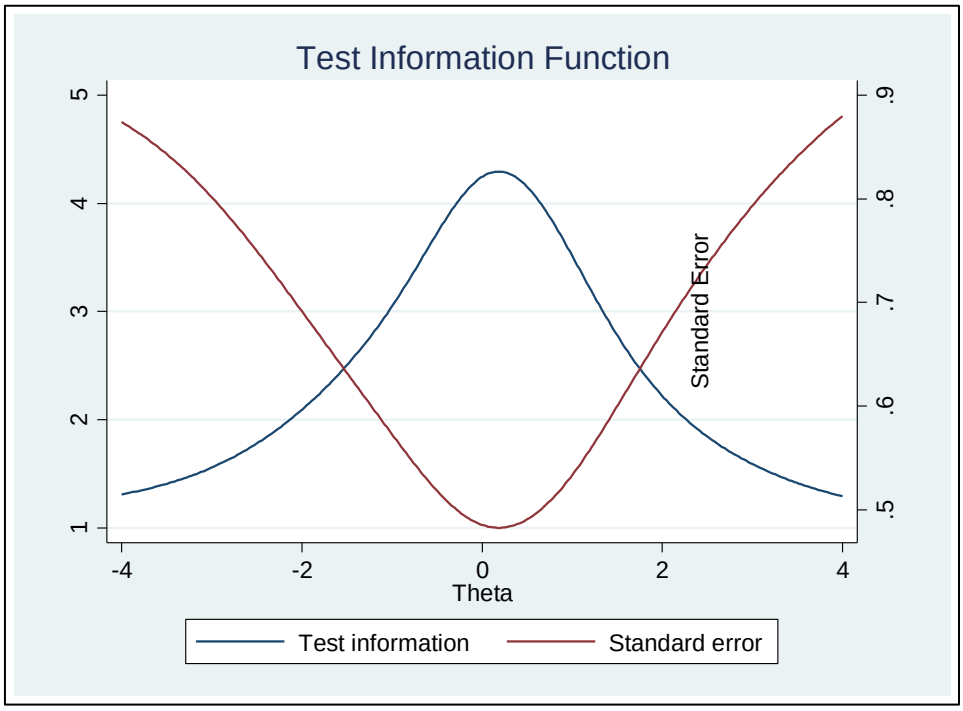


Figure 26: TIF of the new instrument for junior high school

Table 10 presents the discrimination values for the items of the test through the different phases. Again, the discrimination values are more or less the same in Phases 1 and 2, but a lot different in Phase 3. Specifically, items q20, q23, q26 and q3, that were the ones with the lowest discrimination values included in the new instrument, did not manage to prove that they can still be useful in measuring sufficiently the latent construct for this group. On the other hand, item q28 provided this time more information by having a greater discrimination value. There were also items that initially provided great values, but in the new instrument they did not seem to work at all (q6, q12 and q14) or at least as well as before (q7 and q8).

	Coef.		
Discrim	Phase 1	Phase 2	Phase 3
q1	.7532809	.7244879	1.394766
q3	.7076103	.7658868	.2418781
q4	1.049394	1.11271	1.805679
q6	1.096883	.9801241	.5495159
q7	2.642052	2.301246	.8891533
q8	1.521562	1.585394	.8433384
q11	1.822839	1.8243	1.297576
q12	.9566294	1.047694	.3905426
q14	.8364951	.787162	.2895951
q15	.843096	.8661604	1.96843
q16	.9685704	.9929612	.9265515
q20	.4708711	.4078033	.4540692
q23	.432647	.3680199	.1514199
q24	.9792999	.8464293	.9472937
q26	.6933699	.6935888	.0214732
q28	.5239202	.5301519	.8517828

Table 10: Comparing table for junior high school

5.3. Senior High School

We encountered serious statistical problems when trying to analyze data for senior high school students. The 2PL model failed to converge and was iterating indefinitely. We attempted to find items that caused this problem but, unfortunately, the 2PL model analysis was not completed, even when excluding many items and focusing on subsets of the instrument which appear to be less problematic. The most probable cause could be the violation of the unidimensionality of the instrument. There should be a lot more than one latent constructs that the items are related to, thus, even by excluding various combinations of items from the analysis, it was still not possible to complete the analysis.

The goal of this thesis is to calibrate and validate all instruments by calibrating the original one, analyzing the data and then tailoring it by removing items to give the best fit to the hypothesized models. Since it was not possible to calibrate this instrument with the available data, it was decided not to distribute the same initial instrument again for this group. If we had done so, then the new adjusted instrument should be distributed to a new sample, in order to be calibrated and validated as well. As already mentioned in Section 4, due to the difficulty of finding two different samples, we would not be able to achieve the goal of this study, which is to calibrate and validate the new instruments as well.

5.4. Students

The results of the 2PL model for this group (Table 11) are again better than those for the previous one.

Two-parameter logistic model — Number of obs = 219
Log likelihood = -2901.1379

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q13	-.265615	.2482319	-1.07	0.285	-.7521404	.2209105
q5	-.1902752	.1879669	-1.01	0.311	-.5586835	.1781331
q27	-.1731759	.189642	-0.91	0.361	-.5448673	.1985156
q23	-.0513794	.2102834	-0.24	0.807	-.4635273	.3607686
q10	.0490332	.6263084	0.08	0.938	-1.178509	1.276575
q15	.2278275	.1744317	1.31	0.192	-.1140522	.5697073
q22	.2443143	.1881498	1.30	0.194	-.1244526	.6130811
q19	.2665167	.1761315	1.51	0.130	-.0786947	.611728
q20	.2787843	.1784399	1.56	0.118	-.0709514	.62852
q14	.2789189	.4029267	0.69	0.489	-.510803	1.068641
q6	.3490572	.2540826	1.37	0.170	-.1489355	.8470499
q21	.4290073	.2174811	1.97	0.049	.0027522	.8552625
q17	.5113757	.2277004	2.25	0.025	.0650911	.9576603
q25	.5250382	.224257	2.34	0.019	.0855026	.9645738
q26	.5283342	.2070387	2.55	0.011	.1225458	.9341225
q16	.7287725	.2192364	3.32	0.001	.2990771	1.158468
q8	.750907	.2537177	2.96	0.003	.2536293	1.248185
q12	.7990266	.2478756	3.22	0.001	.3131993	1.284854
q3	.8520338	.3934602	2.17	0.030	.0808661	1.623202
q7	.922574	.3177858	2.90	0.004	.2997254	1.545423
q18	.9360947	.2891175	3.24	0.001	.3694348	1.502755
q24	1.054561	.2514925	4.19	0.000	.5616443	1.547477
q1	1.062129	.2924906	3.63	0.000	.4888582	1.635401
q28	1.087678	.2647386	4.11	0.000	.5688003	1.606557
q9	1.284475	.4260466	3.01	0.003	.4494387	2.119511
q11	1.538411	.7005267	2.20	0.028	.1654041	2.911418
q4	2.094593	.6628863	3.16	0.002	.79536	3.393826
q2	2.722854	.8286844	3.29	0.001	1.098663	4.347046
Diff						
q13	7.025219	6.431309	1.09	0.275	-5.579915	19.63035
q5	4.004015	3.964549	1.01	0.313	-3.766359	11.77439
q27	4.639023	5.08135	0.91	0.361	-5.32024	14.59829
q23	26.33669	107.7158	0.24	0.807	-184.7824	237.4558
q10	-81.28178	1037.673	-0.08	0.938	-2115.083	1952.519
q15	-.0412163	.601506	-0.07	0.945	-1.220146	1.137714
q22	-2.963317	2.297389	-1.29	0.197	-7.466116	1.539483
q19	.6632904	.669826	0.99	0.322	-.6495445	1.976125
q20	-1.876209	1.263765	-1.48	0.138	-4.353143	.6007247
q14	-11.02349	15.60994	-0.71	0.480	-41.61841	19.57144
q6	-5.397317	3.794621	-1.42	0.155	-12.83464	2.040004
q21	-3.078747	1.498229	-2.05	0.040	-6.015222	-.1422723
q17	-2.359788	.9965976	-2.37	0.018	-4.313083	-.4064924
q25	2.7241	1.09514	2.49	0.013	.5776639	4.870535
q26	-1.480155	.5893724	-2.51	0.012	-2.635304	-.3250066
q16	-1.223618	.3706757	-3.30	0.001	-1.950129	-.497107
q8	-1.773571	.5372829	-3.30	0.001	-2.826626	-.7205158
q12	-1.688137	.469261	-3.60	0.000	-2.607871	-.7684018
q3	-3.497727	1.358612	-2.57	0.010	-6.160557	-.8348968
q7	-2.299444	.6468397	-3.55	0.000	-3.567227	-1.031661
q18	-2.060384	.5261466	-3.92	0.000	-3.091612	-1.029155
q24	-.5366366	.1813072	-2.96	0.003	-.8919921	-.1812811
q1	-1.486287	.3365718	-4.42	0.000	-2.145955	-.8266182
q28	-.1006904	.1550994	-0.65	0.516	-.4046797	.2032989
q9	-2.294411	.5398843	-4.25	0.000	-3.352565	-1.236257
q11	-2.958822	.8828	-3.35	0.001	-4.689078	-1.228566
q4	-1.738669	.2811236	-6.18	0.000	-2.289661	-1.187677
q2	-.5882278	.1159601	-5.07	0.000	-.8155053	-.3609502

Table 11: 2PL model results for students

The difficulty values, as well as the TCC (Figure 27), show that the test was relatively easy for students. In this group, the test seems to be providing the most information for HSPs between -1 and 0 (Figure 28).

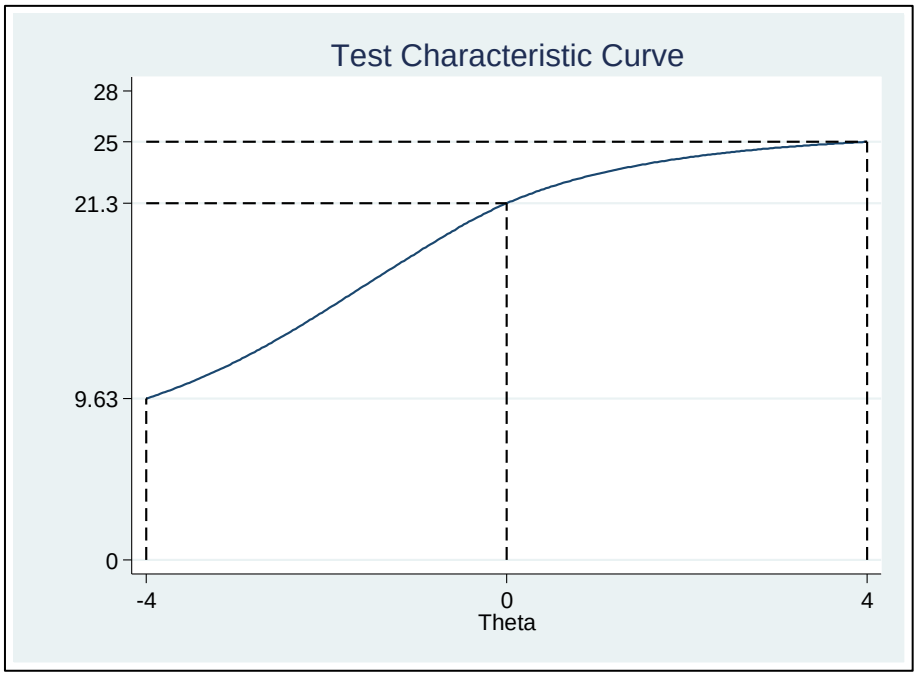


Figure 27: TCC for students

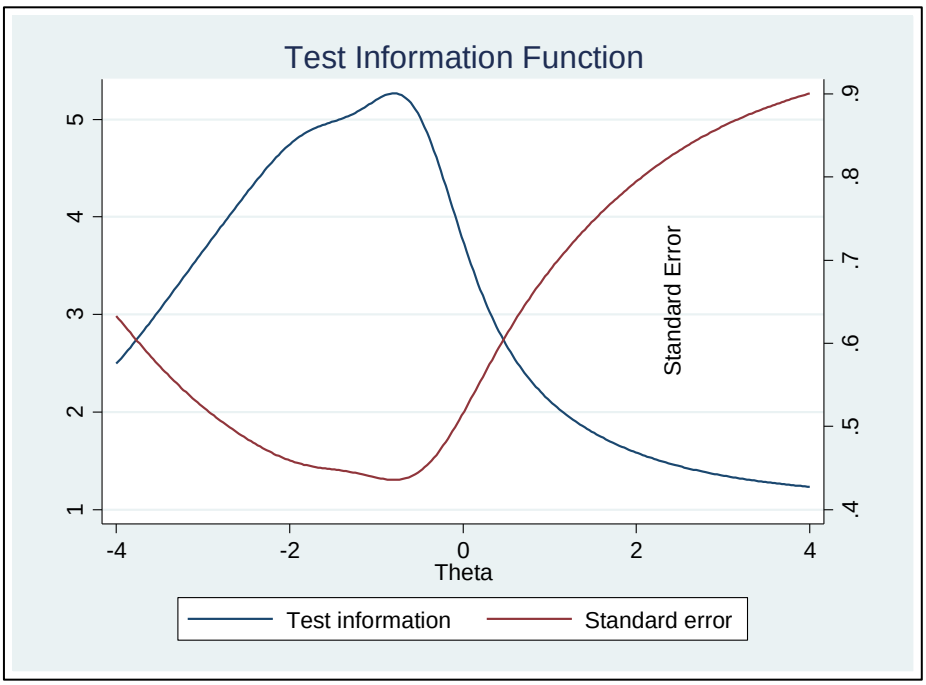


Figure 28: TIF for students

The ICCs and IIFs for items with discrimination values close to the cut-off value are presented in Figure 29 and Figure 30. In the same way as before, items with lower values than the cut-off value and/or values statistically close to zero are excluded from the new instrument.

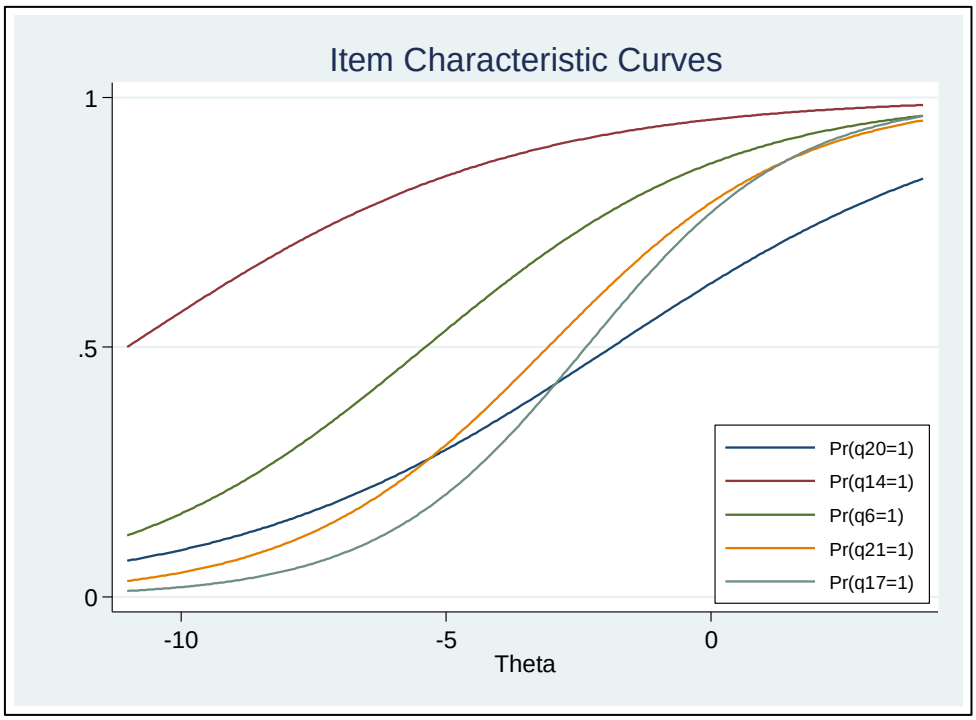


Figure 29: ICCs for items with discrimination values close to the cut-off value for students

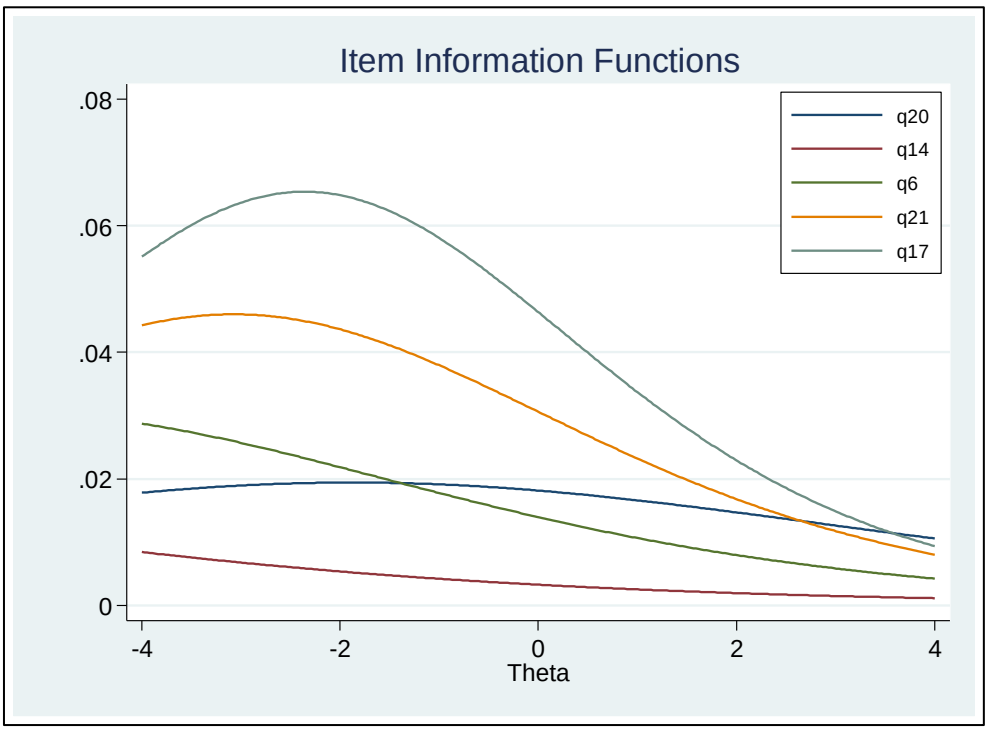


Figure 30: IIFs for items with discrimination values close to the cut-off value for students

The 17 items that were finally included were analyzed with the 2PL model. The results (Table 12) show again an overall improvement.

Two-parameter logistic model — Number of obs = 219
Log likelihood = -1703.3011

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q21	.4938435	.2194812	2.25	0.024	.0636683	.9240187
q25	.5188841	.2263314	2.29	0.022	.0752827	.9624854
q17	.5309863	.2220401	2.39	0.017	.0957957	.9661768
q26	.5438239	.2026795	2.68	0.007	.1465794	.9410685
q16	.6493568	.2101497	3.09	0.002	.237471	1.061243
q8	.7209043	.2449699	2.94	0.003	.2407721	1.201037
q12	.8285573	.2450463	3.38	0.001	.3482753	1.308839
q3	.8313867	.3811847	2.18	0.029	.0842785	1.578495
q7	.8462269	.3006766	2.81	0.005	.2569115	1.435542
q18	.942693	.2853496	3.30	0.001	.3834181	1.501968
q24	1.02369	.2440022	4.20	0.000	.5454549	1.501926
q1	1.056096	.289856	3.64	0.000	.4879884	1.624203
q28	1.114381	.2651887	4.20	0.000	.5946211	1.634142
q9	1.278329	.4217485	3.03	0.002	.4517167	2.10494
q11	1.560849	.6941144	2.25	0.025	.2004096	2.921288
q4	1.949642	.5862004	3.33	0.001	.8007105	3.098574
q2	2.765739	.8369931	3.30	0.001	1.125262	4.406215
Diff						
q21	-2.706545	1.142238	-2.37	0.018	-4.94529	-.4678003
q25	2.753103	1.130357	2.44	0.015	.5376447	4.968561
q17	-2.28173	.9052162	-2.52	0.012	-4.055921	-.5075392
q26	-1.442598	.5507967	-2.62	0.009	-2.522139	-.3630559
q16	-1.344666	.4395425	-3.06	0.002	-2.206153	-.4831786
q8	-1.833057	.5648472	-3.25	0.001	-2.940137	-.7259771
q12	-1.640941	.4349651	-3.77	0.000	-2.493457	-.7884252
q3	-3.570074	1.387274	-2.57	0.010	-6.289082	-.8510662
q7	-2.458523	.7339577	-3.35	0.001	-3.897053	-1.019992
q18	-2.050212	.5132956	-3.99	0.000	-3.056253	-1.044171
q24	-.5460312	.1860211	-2.94	0.003	-.9106258	-.1814365
q1	-1.492081	.3380435	-4.41	0.000	-2.154634	-.8295277
q28	-.0976318	.1526229	-0.64	0.522	-.3967673	.2015036
q9	-2.303711	.5399262	-4.27	0.000	-3.361947	-1.245475
q11	-2.938654	.8471737	-3.47	0.001	-4.599084	-1.278224
q4	-1.794765	.2953821	-6.08	0.000	-2.373703	-1.215826
q2	-.5847867	.1144536	-5.11	0.000	-.8091117	-.3604618

Table 12: 2PL model results for students before distribution

The TCC (Figure 31) is now steeper and covers a greater range of the expected score but the TIF (Figure 32) does not differentiate a lot.

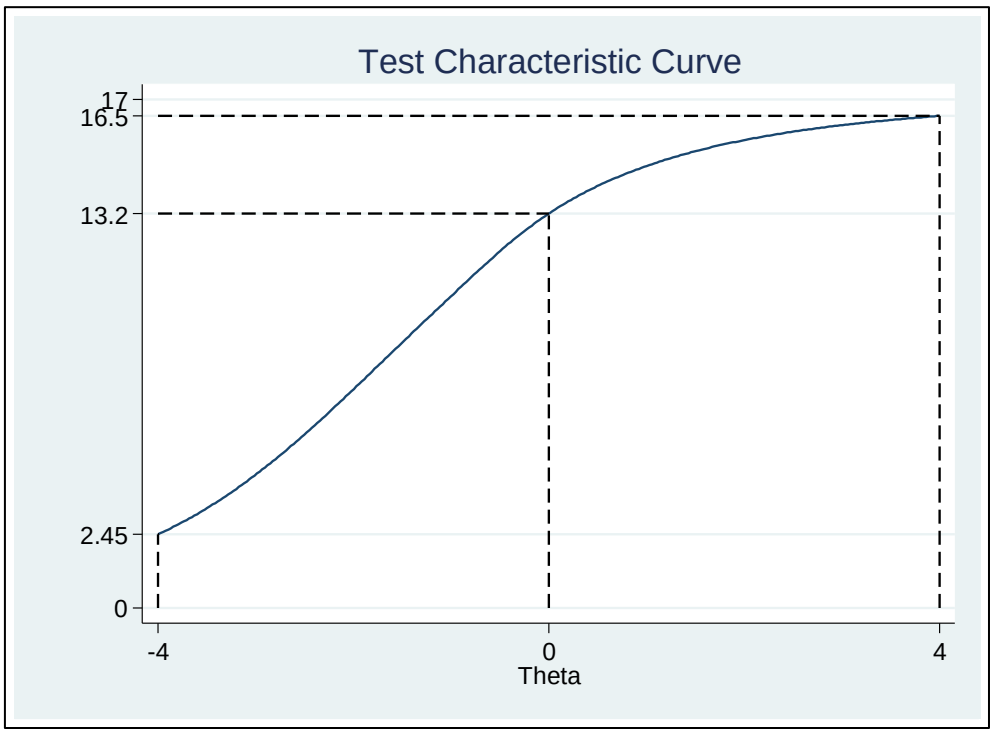


Figure 31: TCC for students before distribution

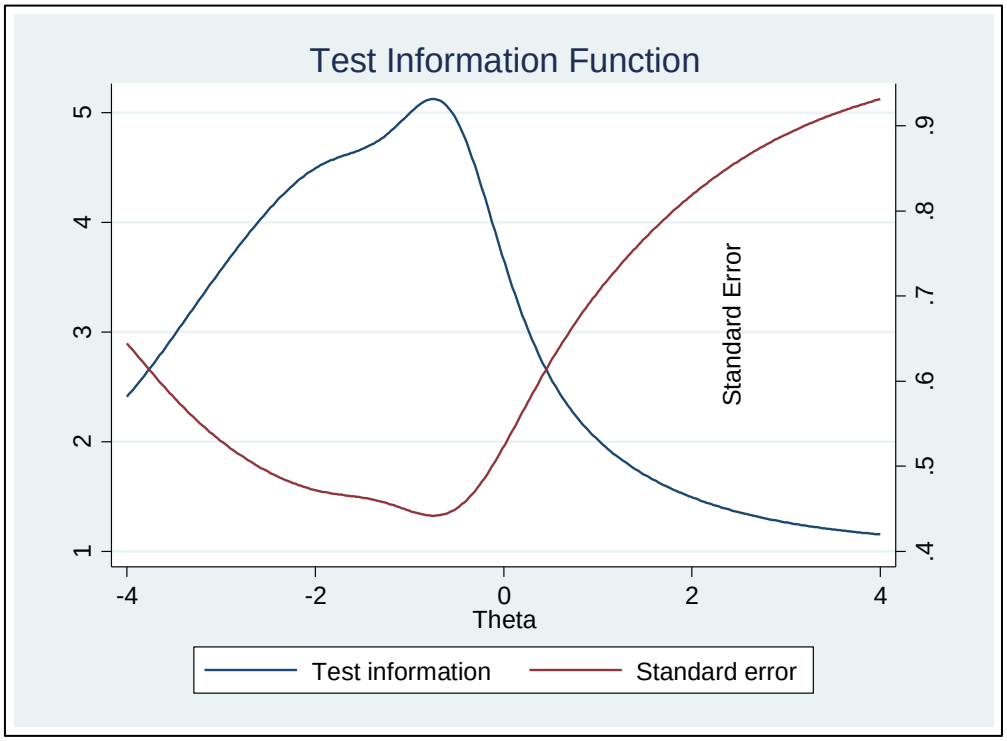


Figure 32: TIF for students before distribution

The descriptive statistics for the data collected with the new instrument are found in the Appendix. These data were fitted to the 2PL model and the results are shown in Table 13. The new instrument appears to have a great improvement regarding the discrimination values, as it was expected from the analysis prior to its distribution. There are no negative values and those who are statistically close to zero (q1, q2, q18 and q21), this is, however, at a relatively low significance level.

Two-parameter logistic model — Number of obs = 155
Log likelihood = -1373.3759

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q1	.2934187	.2143406	1.37	0.171	-.1266812	.7135186
q2	.413108	.2201092	1.88	0.061	-.0182981	.844514
q18	.463073	.2806985	1.65	0.099	-.0870859	1.013232
q21	.4953284	.2816624	1.76	0.079	-.0567197	1.047376
q28	.5033587	.2251217	2.24	0.025	.0621282	.9445891
q17	.5653916	.2472904	2.29	0.022	.0807113	1.050072
q25	.5747853	.2770397	2.07	0.038	.0317975	1.117773
q26	.6219042	.240293	2.59	0.010	.1509385	1.09287
q11	.7499028	.3076169	2.44	0.015	.1469848	1.352821
q12	.9271402	.3007174	3.08	0.002	.3377449	1.516536
q3	.9661062	.3978734	2.43	0.015	.1862887	1.745924
q24	1.069711	.3105854	3.44	0.001	.4609751	1.678447
q8	1.372725	.4081833	3.36	0.001	.5727003	2.172749
q9	1.379626	.4390327	3.14	0.002	.5191381	2.240115
q4	1.392175	.4350559	3.20	0.001	.5394812	2.244869
q16	1.631954	.4558509	3.58	0.000	.7385022	2.525405
q7	2.152124	.7376468	2.92	0.004	.7063629	3.597885
Diff						
q1	-2.0798	1.570885	-1.32	0.186	-5.158679	.9990784
q2	-1.158417	.7048498	-1.64	0.100	-2.539897	.2230631
q18	-3.506302	2.011828	-1.74	0.081	-7.449412	.4368072
q21	-3.204527	1.713774	-1.87	0.062	-6.563463	.1544087
q28	.4614537	.38835	1.19	0.235	-.2996983	1.222606
q17	-1.520113	.6668657	-2.28	0.023	-2.827145	-.2130799
q25	2.575533	1.153745	2.23	0.026	.3142352	4.836832
q26	-.4799408	.3231452	-1.49	0.137	-1.113294	.1534122
q11	-2.236668	.811027	-2.76	0.006	-3.826251	-.6470838
q12	-1.213931	.367142	-3.31	0.001	-1.933517	-.4943462
q3	-2.584256	.8619787	-3.00	0.003	-4.273704	-.8948092
q24	-.2607388	.1920571	-1.36	0.175	-.6371638	.1156862
q8	-1.19654	.2781735	-4.30	0.000	-1.74175	-.6513304
q9	-1.616432	.3715609	-4.35	0.000	-2.344678	-.8881858
q4	-1.522605	.3463005	-4.40	0.000	-2.201341	-.8438681
q16	-.0009745	.1448286	-0.01	0.995	-.2848333	.2828842
q7	-1.569285	.2889871	-5.43	0.000	-2.13569	-1.002881

Table 13: 2PL model results of the new instrument for students

The same applies for the TCC (Figure 33), which is almost identical to the one from the prior analysis. This fact verifies the improvement of the instrument, as expected, providing a steeper curve. The TIF (Figure 34) is close to the expectations as well, with the only difference that it provides the greatest amount of information for HSP with a bit lower ability (approximately from -2 to -0.5).

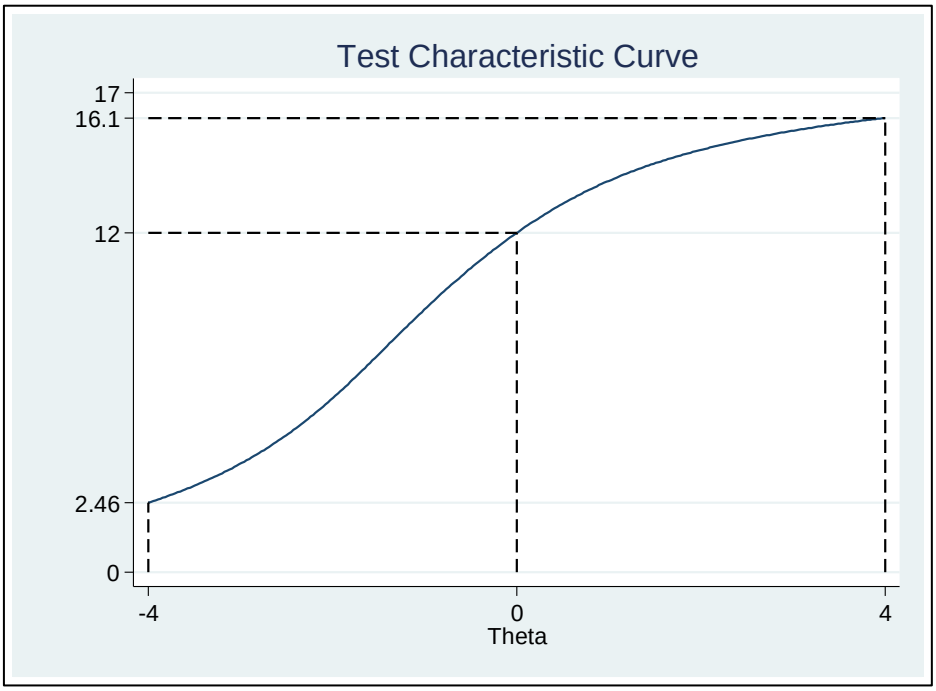


Figure 33: TCC of the new instrument for students

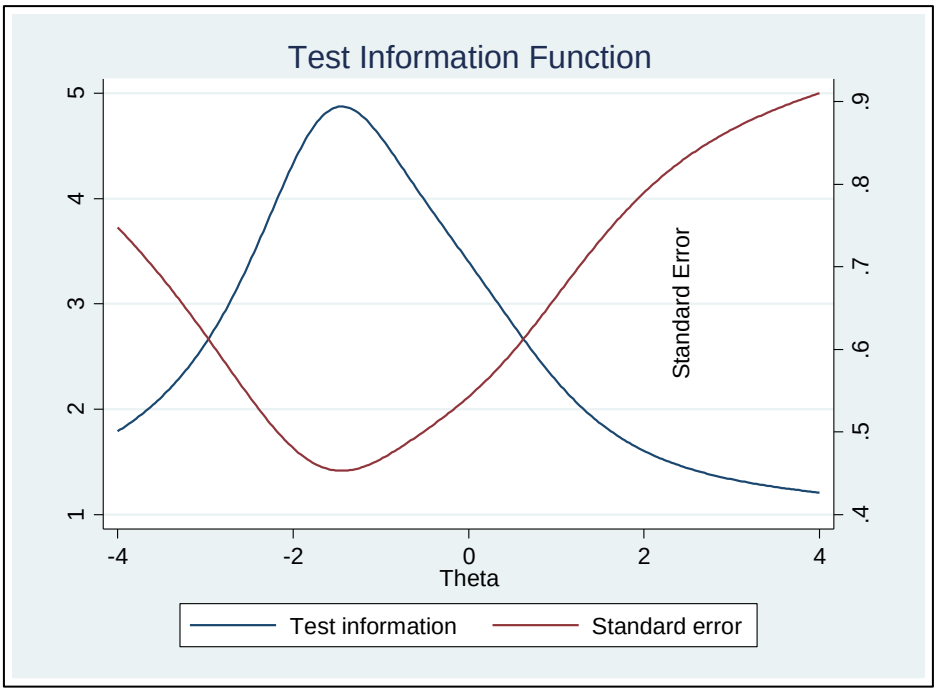


Figure 34: TIF of the new instrument for students

In order to observe the variations of the discrimination values through the different phases of the analysis, the following table was created (Table 14). For phases 1 and 2, the discrimination values are more or less the same. For phase 3, these values are again different. Nevertheless, as discussed before, the overall image of the test is indeed improved as expected.

	Coef.		
Discrim	Phase 1	Phase 2	Phase 3
q1	1.062129	1.056096	.2934187
q2	2.722854	2.765739	.413108
q3	.8520338	.8313867	.9661062
q4	2.094593	1.949642	1.392175
q7	.922574	.8462269	2.152124
q8	.750907	.7209043	1.372725
q9	1.284475	1.278329	1.379626
q11	1.538411	1.560849	.7499028
q12	.7990266	.8285573	.9271402
q16	.7287725	.6493568	1.631954
q17	.5113757	.5309863	.5653916
q18	.9360947	.942693	.463073
q21	.4290073	.4938435	.4953284
q24	1.054561	1.02369	1.069711
q25	.5250382	.5188841	.5747853
q26	.5283342	.5438239	.6219042
q28	1.087678	1.114381	.5033587

Table 14: Comparing table for students

5.5. Teachers

As expected, the instrument seemed to be easier for the teachers. The results of the 2PL model are presented in Table 15.

Two-parameter logistic model — Number of obs = 209
Log likelihood = -2513.2166

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q21	.0553491	.2049121	0.27	0.787	-.3462712	.4569694
q3	.0723901	.3024348	0.24	0.811	-.5203713	.6651515
q6	.201099	.3862499	0.52	0.603	-.5559369	.9581349
q13	.2316442	.2128653	1.09	0.276	-.1855641	.6488525
q14	.3297775	.5699532	0.58	0.563	-.7873102	1.446865
q22	.3467995	.1829844	1.90	0.058	-.0118434	.7054423
q19	.3759424	.1720504	2.19	0.029	.0387297	.713155
q1	.3785361	.1874825	2.02	0.043	.0110772	.7459951
q27	.3900327	.2225414	1.75	0.080	-.0461404	.8262057
q10	.4324339	.5851802	0.74	0.460	-.7144981	1.579366
q26	.5633964	.194266	2.90	0.004	.1826419	.9441508
q17	.6114528	.2263446	2.70	0.007	.1678255	1.05508
q20	.624015	.2239987	2.79	0.005	.1849856	1.063044
q2	.6338079	.1878134	3.37	0.001	.2657005	1.001915
q15	.6584363	.1936379	3.40	0.001	.2789129	1.03796
q23	.8663684	.338017	2.56	0.010	.2038672	1.52887
q5	.9026331	.2158883	4.18	0.000	.4794999	1.325766
q28	.9993582	.2238009	4.47	0.000	.5607165	1.438
q9	1.0926	.2693005	4.06	0.000	.564781	1.62042
q12	1.209469	.3098041	3.90	0.000	.6022644	1.816674
q4	1.261927	.3928632	3.21	0.001	.4919293	2.031925
q24	1.302416	.2662681	4.89	0.000	.7805397	1.824291
q8	1.520982	.3249516	4.68	0.000	.8840885	2.157876
q7	1.682217	.3743688	4.49	0.000	.9484681	2.415967
q16	1.840165	.3694282	4.98	0.000	1.116099	2.564231
q11	1.844377	.6751705	2.73	0.006	.5210671	3.167687
q18	2.17847	.6245381	3.49	0.000	.9543978	3.402542
q25	2.801442	.6843258	4.09	0.000	1.460188	4.142696
Diff						
q21	-26.05269	96.37588	-0.27	0.787	-214.9459	162.8406
q3	-34.4295	143.6284	-0.24	0.811	-315.9361	247.0771
q6	-15.5128	29.49262	-0.53	0.599	-73.31728	42.29167
q13	-6.852512	6.210681	-1.10	0.270	-19.02522	5.320199
q14	-12.09562	20.41984	-0.59	0.554	-52.11778	27.92653
q22	-2.627003	1.387951	-1.89	0.058	-5.347336	.0933311
q19	-.3432563	.4085882	-0.84	0.401	-1.144074	.4575619
q1	-2.67567	1.313789	-2.04	0.042	-5.250649	-.1006914
q27	-4.330363	2.377807	-1.82	0.069	-8.990778	.3300528
q10	-9.310139	12.10264	-0.77	0.442	-33.03088	14.4106
q26	-1.773431	.6090498	-2.91	0.004	-2.967146	-.579715
q17	-2.591318	.8861871	-2.92	0.003	-4.328213	-.8544228
q20	-2.546289	.8436041	-3.02	0.003	-4.199723	-.8928556
q2	-.0167738	.2384752	-0.07	0.944	-.4841766	.4506291
q15	-1.002844	.3443577	-2.91	0.004	-1.677772	-.3279148
q23	-3.222564	1.052546	-3.06	0.002	-5.285516	-1.159612
q5	-.5438425	.2076776	-2.62	0.009	-.9508831	-.136802
q28	-.5539993	.1923542	-2.88	0.004	-.9310067	-.176992
q9	-1.616614	.3316813	-4.87	0.000	-2.266697	-.9665304
q12	-1.934635	.3783469	-5.11	0.000	-2.676181	-1.193088
q4	-2.455772	.5513881	-4.45	0.000	-3.536473	-1.375072
q24	-.5917033	.1636089	-3.62	0.000	-.9123709	-.2710358
q8	-1.147382	.1978935	-5.80	0.000	-1.535247	-.7595183
q7	-1.398957	.2156041	-6.49	0.000	-1.821534	-.9763811
q16	-1.004765	.1618481	-6.21	0.000	-1.321981	-.6875487
q11	-2.477011	.5115713	-4.84	0.000	-3.479673	-1.47435
q18	-1.783048	.2549444	-6.99	0.000	-2.28273	-1.283366
q25	-.091956	.1025001	-0.90	0.370	-.2928525	.1089405

Table 15: 2PL model results for teachers

All difficulty values are negative or statistically equal to zero, which implies that the items where too easy for this group, and the TCC (Figure 35) provides high values of expected score. Nevertheless, the test appears to be satisfyingly discriminative, taking into account the fact that there is no negative discrimination value and that the TCC is relatively steep.

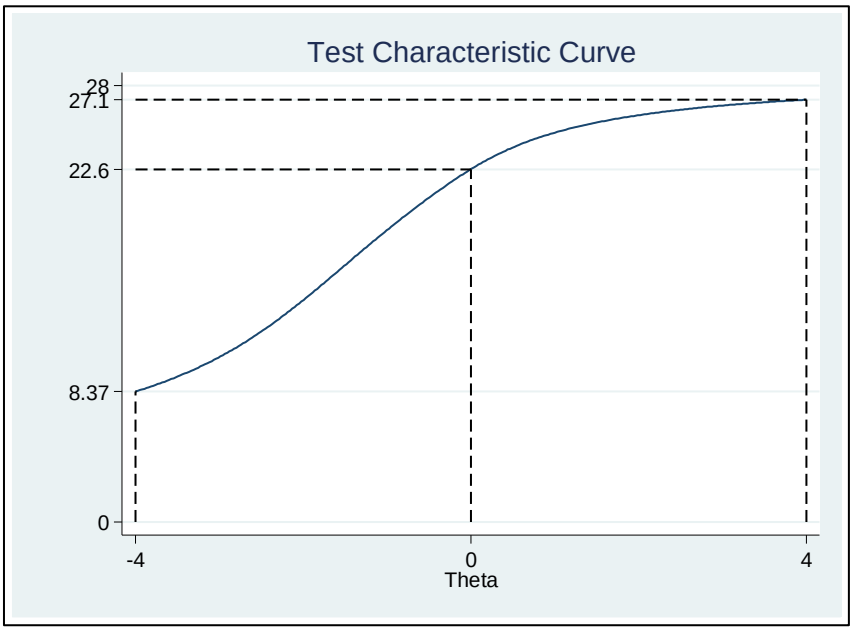


Figure 35: TCC for teachers

A greater amount of information about the latent construct is provided for this group in combination with a lower standard error (TIF in Figure 36), especially for individuals belonging to HSPs between -2 and 0.

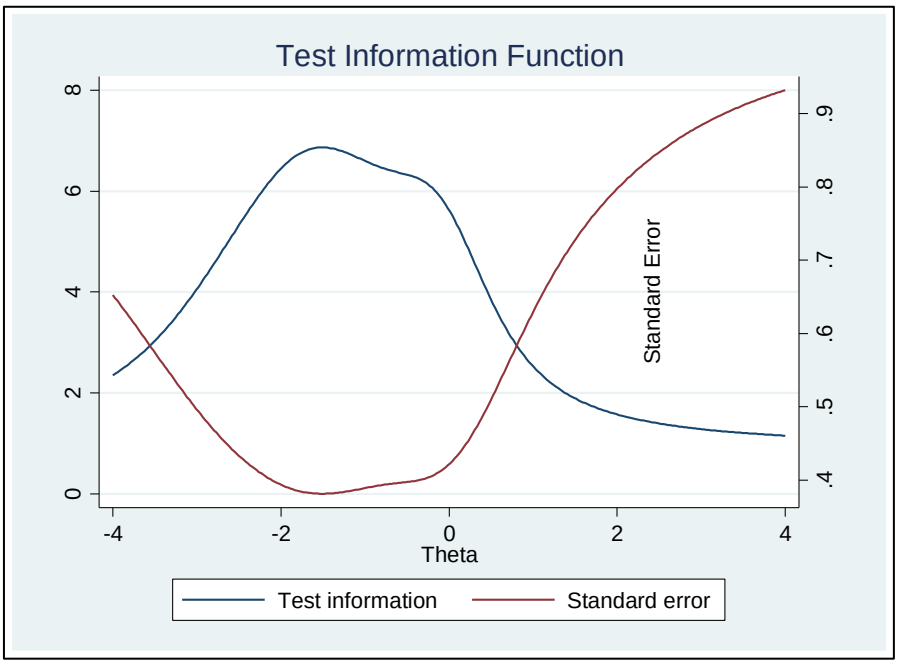


Figure 36: TIF for teachers

For the first six items that have discrimination values greater than the cut-off value, the ICCs and IIFs are shown in Figure 37 and Figure 38. Following the same methodology, items with low discrimination values and/or with values statistically close to zero are the ones to be excluded.

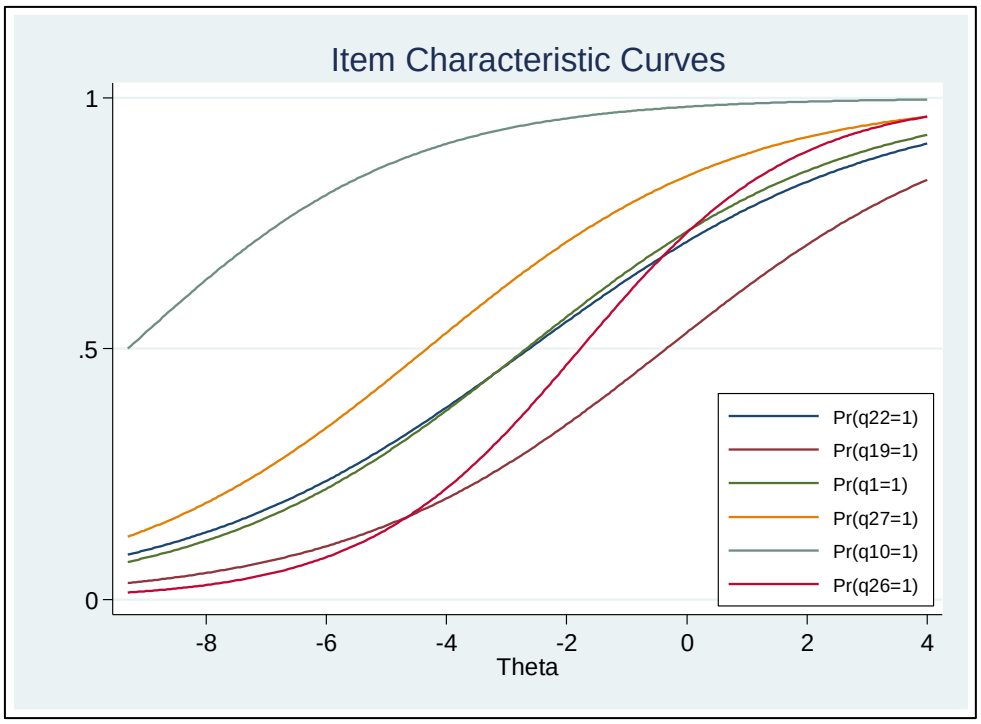


Figure 37: ICCs for items with discrimination values close to the cut-off value for teachers

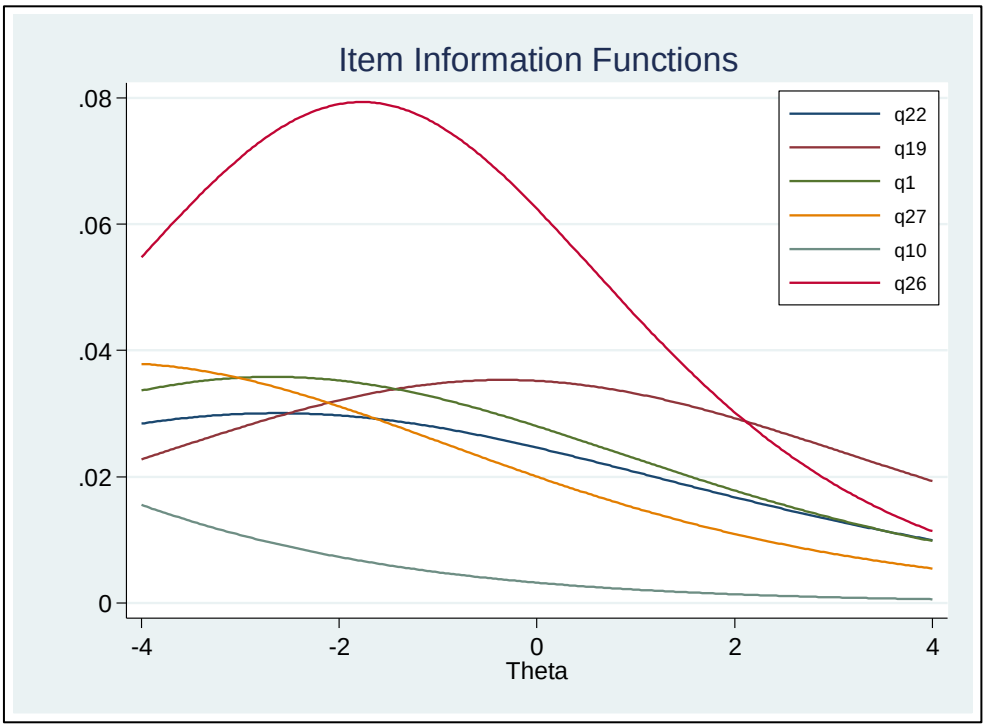


Figure 38: IIFs for items with discrimination values close to the cut-off value for teachers

That leads to a new instrument with 20 items for this group. The 2PL model was applied for this subset of items. The results (Table 16) are now showing an even more discriminating instrument.

Two-parameter logistic model — Number of obs = 209
Log likelihood = -1967.6321

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q19	.3404867	.1698669	2.00	0.045	.0075537	.6734196
q1	.3665226	.1867755	1.96	0.050	.0004493	.7325959
q26	.5639282	.1939828	2.91	0.004	.1837288	.9441276
q20	.6026623	.2216657	2.72	0.007	.1682055	1.037119
q17	.6456546	.2282679	2.83	0.005	.1982577	1.093052
q2	.6555368	.1898699	3.45	0.001	.2833988	1.027675
q15	.6659267	.1939696	3.43	0.001	.2857532	1.0461
q23	.8198506	.3310652	2.48	0.013	.1709746	1.468727
q5	.9269887	.2181475	4.25	0.000	.4994276	1.35455
q28	.9534765	.2190757	4.35	0.000	.5240961	1.382857
q9	1.072594	.2654053	4.04	0.000	.5524093	1.592779
q12	1.201229	.3074094	3.91	0.000	.5987178	1.803741
q4	1.244782	.3860589	3.22	0.001	.4881204	2.001444
q24	1.285628	.2636705	4.88	0.000	.7688437	1.802413
q8	1.499899	.3209672	4.67	0.000	.8708144	2.128983
q7	1.676435	.3712733	4.52	0.000	.9487528	2.404118
q11	1.828796	.6607273	2.77	0.006	.5337945	3.123798
q16	1.914253	.3836305	4.99	0.000	1.162351	2.666155
q18	2.350412	.674094	3.49	0.000	1.029212	3.671612
q25	2.707805	.6571196	4.12	0.000	1.419874	3.995735
Diff						
q19	-.3766146	.4547625	-0.83	0.408	-1.267933	.5147034
q1	-2.757965	1.394212	-1.98	0.048	-5.490571	-.0253584
q26	-1.771717	.6075193	-2.92	0.004	-2.962433	-.5810012
q20	-2.623982	.895378	-2.93	0.003	-4.37889	-.8690732
q17	-2.473052	.8020684	-3.08	0.002	-4.045077	-.9010266
q2	-.0158258	.2318245	-0.07	0.946	-.4701935	.4385418
q15	-.9929781	.3390399	-2.93	0.003	-1.657484	-.3284721
q23	-3.369486	1.157495	-2.91	0.004	-5.638134	-1.100837
q5	-.5324444	.2025718	-2.63	0.009	-.9294779	-.1354109
q28	-.5718243	.201134	-2.84	0.004	-.9660398	-.1776088
q9	-1.636417	.3395934	-4.82	0.000	-2.302007	-.9708259
q12	-1.943292	.3812341	-5.10	0.000	-2.690497	-1.196087
q4	-2.479098	.5585514	-4.44	0.000	-3.573839	-1.384357
q24	-.5947653	.1654407	-3.60	0.000	-.9190233	-.2705074
q8	-1.15424	.2009137	-5.74	0.000	-1.548024	-.7604567
q7	-1.400339	.216285	-6.47	0.000	-1.824249	-.9764277
q11	-2.492299	.5106482	-4.88	0.000	-3.493151	-1.491447
q16	-.9860568	.1575838	-6.26	0.000	-1.294915	-.6771982
q18	-1.739673	.2348403	-7.41	0.000	-2.199951	-1.279394
q25	-.0913803	.1034044	-0.88	0.377	-.2940493	.1112886

Table 16: 2PL model results for teachers before distribution

This is also endorsed by the TCC (Figure 39), which is now steeper but still providing relatively high values of expected score. The TIF (Figure 40) shows minimal differentiation.

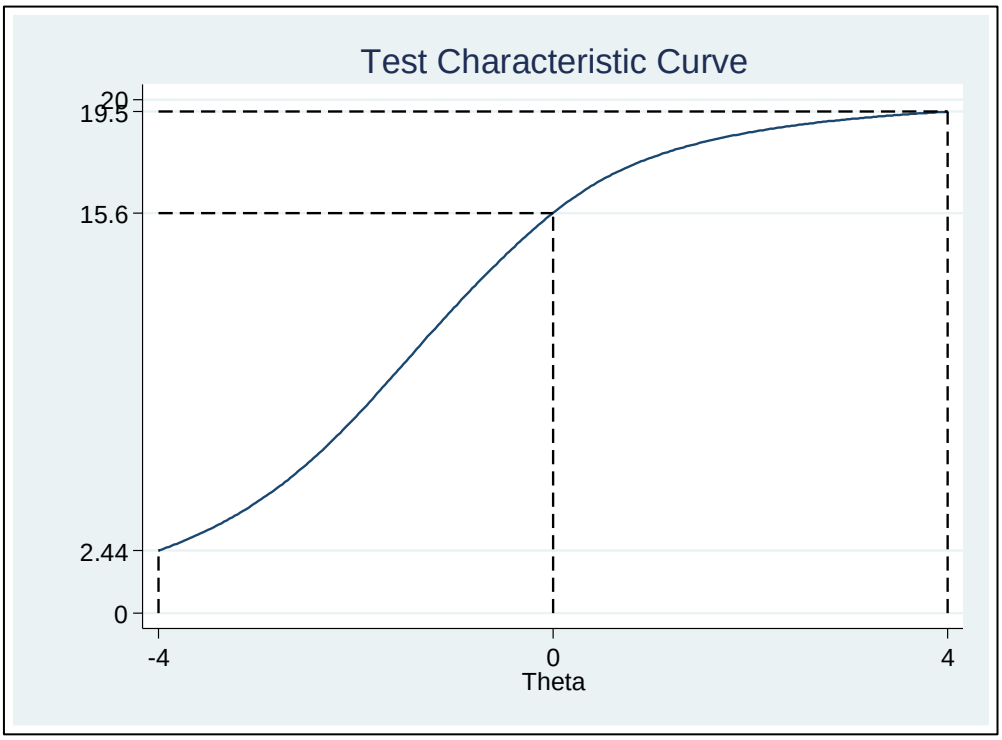


Figure 39: TCC for teachers before distribution

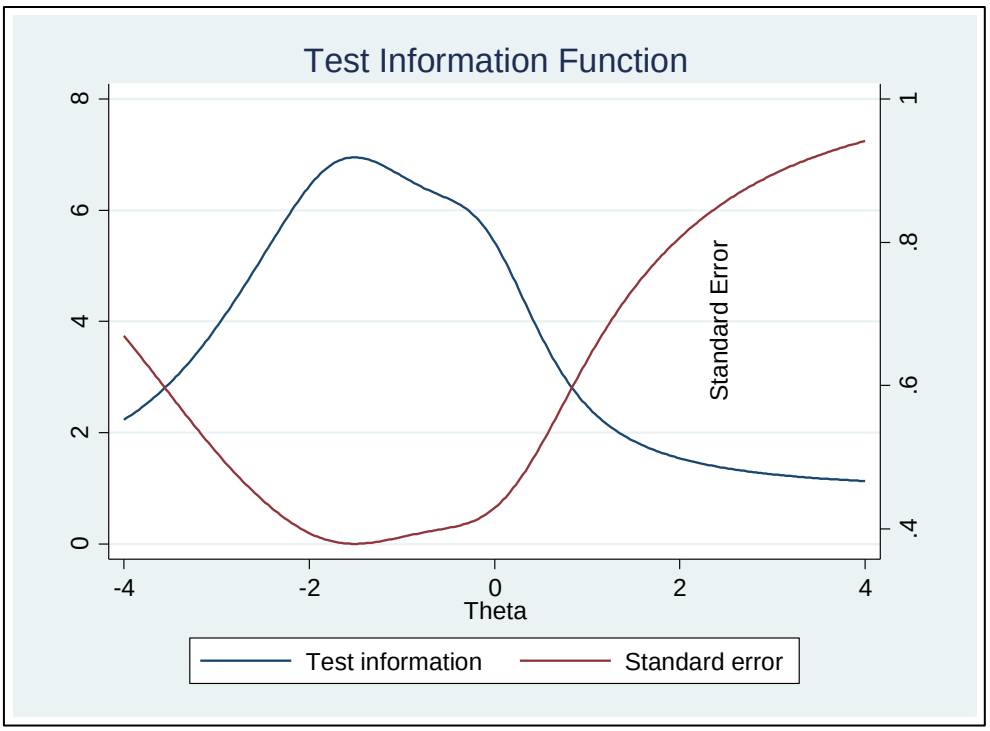


Figure 40: TIF for teachers before distribution

Once again, the new data were fitted to the 2PL model and the results are presented in Table 17. The discrimination values are here in general greater, as well, but there are still a few items that have very low and/or statistically zero values.

Two-parameter logistic model — Number of obs = 96
Log likelihood = -904.77674

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
q5	-.1006378	.2704742	-0.37	0.710	-.6307575	.4294819
q20	.2757723	.2467052	1.12	0.264	-.2077609	.7593055
q23	.4746954	.2968815	1.60	0.110	-.1071816	1.056572
q2	.7904695	.3216985	2.46	0.014	.159952	1.420987
q17	.8654515	.439981	1.97	0.049	.0031046	1.727798
q15	.8927203	.4683598	1.91	0.057	-.0252481	1.810689
q26	.93213	.3384773	2.75	0.006	.2687266	1.595533
q28	1.124721	.3600547	3.12	0.002	.4190267	1.830415
q11	1.173217	.4417713	2.66	0.008	.3073612	2.039073
q18	1.29077	.5362124	2.41	0.016	.2398125	2.341727
q19	1.314802	.3937642	3.34	0.001	.5430379	2.086565
q25	1.45698	.3987661	3.65	0.000	.6754133	2.238548
q12	1.470423	.784103	1.88	0.061	-.0663906	3.007237
q24	1.574341	.5081708	3.10	0.002	.5783446	2.570338
q16	1.745462	.493471	3.54	0.000	.778277	2.712648
q7	1.787959	.8004806	2.23	0.026	.2190462	3.356872
q8	1.913809	.5417984	3.53	0.000	.8519033	2.975714
q9	2.121217	.663866	3.20	0.001	.8200638	3.422371
q1	2.514555	.8202718	3.07	0.002	.9068522	4.122259
q4	3.307915	1.76716	1.87	0.061	-.1556549	6.771484
Diff						
q5	10.9416	29.36629	0.37	0.709	-46.61527	68.49847
q20	-1.563404	1.549203	-1.01	0.313	-4.599786	1.472977
q23	-2.430865	1.472976	-1.65	0.099	-5.317845	.4561143
q2	-1.352848	.5328821	-2.54	0.011	-2.397278	-.3084183
q17	-2.320241	.9918945	-2.34	0.019	-4.264318	-.3761634
q15	-2.598792	1.137985	-2.28	0.022	-4.829201	-.3683824
q26	-1.013341	.3768084	-2.69	0.007	-1.751872	-.2748103
q28	-.6927678	.2698554	-2.57	0.010	-1.221675	-.1638609
q11	-1.501956	.4480926	-3.35	0.001	-2.380201	-.6237105
q18	-1.753926	.5294071	-3.31	0.001	-2.791545	-.7163077
q19	.7866679	.2655597	2.96	0.003	.2661804	1.307155
q25	.2577673	.2049627	1.26	0.209	-.1439522	.6594867
q12	-2.233517	.7997176	-2.79	0.005	-3.800934	-.666099
q24	-1.120381	.2745864	-4.08	0.000	-1.658561	-.5822018
q16	-.5668424	.1928946	-2.94	0.003	-.9449089	-.1887758
q7	-1.81962	.4862011	-3.74	0.000	-2.772557	-.8666836
q8	-.5135401	.1812358	-2.83	0.005	-.8687557	-.1583244
q9	-.8747206	.1989123	-4.40	0.000	-1.264582	-.4848596
q1	-.9510551	.1908075	-4.98	0.000	-1.325031	-.5770794
q4	-1.459711	.2749002	-5.31	0.000	-1.998506	-.9209168

Table 17: 2PL model results of the new instrument for teachers

The TCC (Figure 41) is almost exactly as expected, but the TIF (Figure 42) implies that the test provides more information regarding the latent construct than expected.

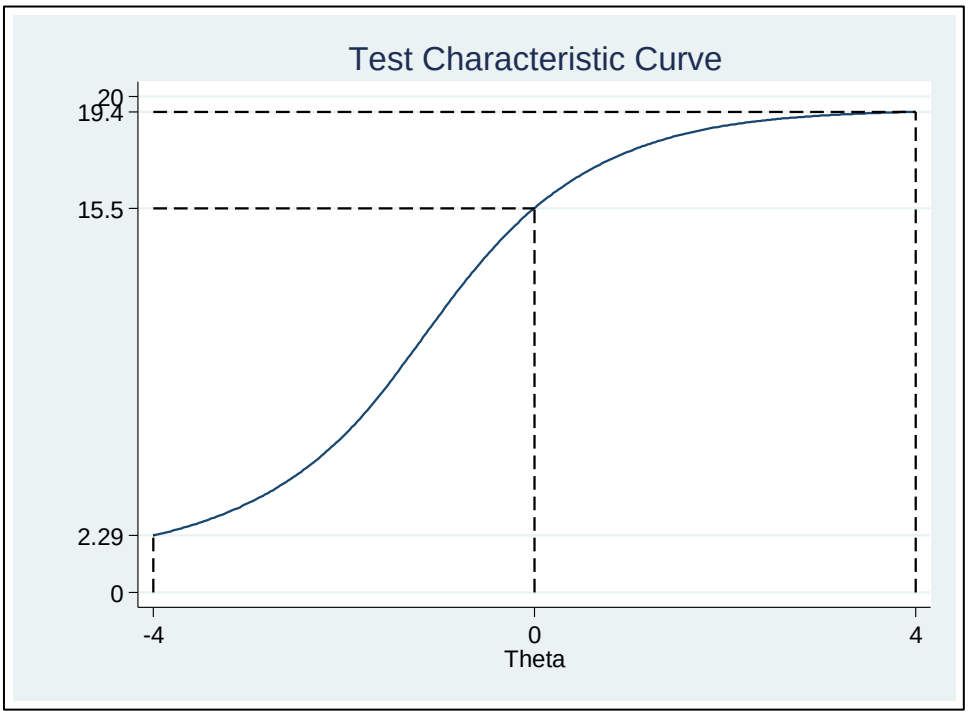


Figure 41: TCC of the new instrument for teachers

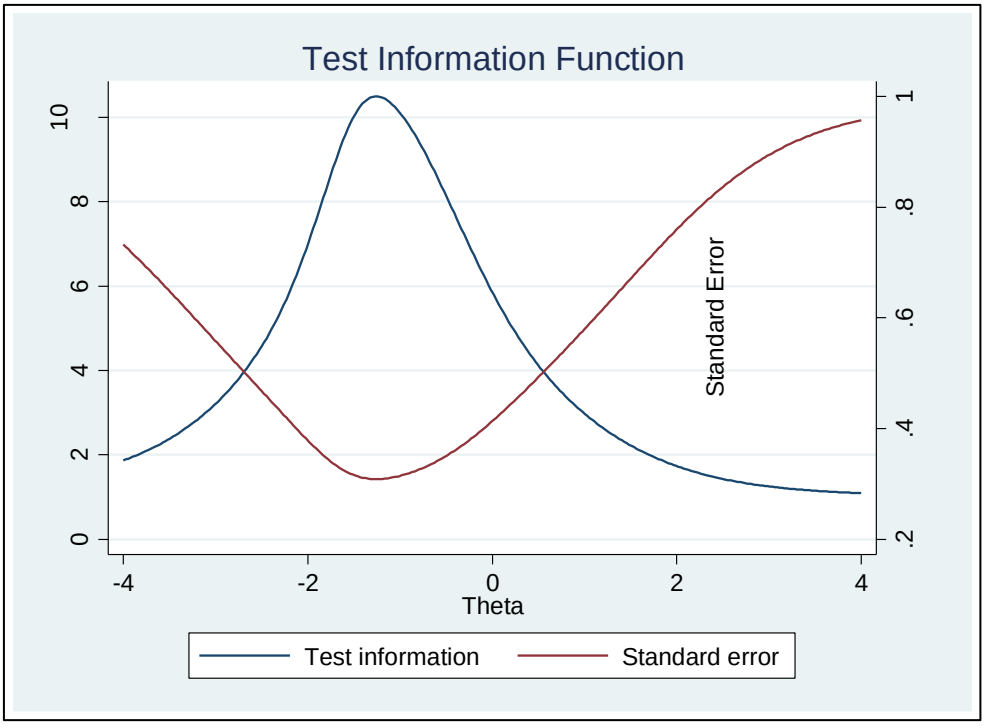


Figure 42: TIF of the new instrument for teachers

Taking a look at the progress of the discrimination values through the different phases of the analysis (Table 18), it is noticeable that most of the items have now improved values. Although, there are still these items that have statistically zero values (q4, q5, q12, q15, q20 and q23), or simply a bit lower values than before (q8, q11, q16, q18 and q25).

	Coef.		
Discrim	Phase 1	Phase 2	Phase 3
q1	.3785361	.3665226	2.514555
q2	.6338079	.6555368	.7904695
q4	1.261927	1.244782	3.307915
q5	.9026331	.9269887	-.1006378
q7	1.682217	1.676435	1.787959
q8	1.520982	1.499899	1.913809
q9	1.0926	1.072594	2.121217
q11	1.844377	1.828796	1.173217
q12	1.209469	1.201229	1.470423
q15	.6584363	.6659267	.8927203
q16	1.840165	1.914253	1.745462
q17	.6114528	.6456546	.8654515
q18	2.17847	2.350412	1.29077
q19	.3759424	.3404867	1.314802
q20	.624015	.6026623	.2757723
q23	.8663684	.8198506	.4746954
q24	1.302416	1.285628	1.574341
q25	2.801442	2.707805	1.45698
q26	.5633964	.5639282	.93213
q28	.9993582	.9534765	1.124721

Table 18: Comparing table for teachers

6. Discussion and Conclusions

The purpose of this study, as it is already mentioned, is to calibrate, improve and validate already existing instruments in the area of mechanics.

Analyzing the original instrument for each group, moving gradually from the lowest to the highest age group, we noticed every time a rising improvement, regarding the average of correct answers, the discrimination values and the appropriateness of the instrument. This finding, converges with the conclusion of the study of Κώτσης (2011), where the results are getting better in relation to the age group.

The instrument created for the primary school students appeared to be much more effective than expected. The analysis for the initial instrument for this group revealed that more than half of the items were not good enough at measuring the academic ability in mechanics. In addition, the overall image of the instrument was not that positive, as the TCC showed a low discrimination power (not steep enough slope) and in connection to that, the TIF did not provide a great amount of information about the latent construct. We excluded a number of 17 items (from a total of 28) from the new instrument, as they provided negative, statistically zero or too low discrimination values. The discrimination values of the rest 11 items in the new instrument, were significantly greater than before (>1,28), with only exception the items q11 and q22 that did not seem to fulfill the expectations. As a result, the TCC is much steeper than expected, showing that the new test is possessing a greater discrimination power. Of course, greater discrimination power, means more information about the late construct, fact which is also supported by the TIF which now provides almost three times more information about it. Important is also the fact that, according to the TIF, the new instrument provides the greatest amount of information for almost the same HSPs as before. Therefore, the new instrument is more appropriate for gathering information about the same ability range of θ comparing to the initial.

Regarding the instrument for the junior high school students, the results were not that positive, as for the previous group. Although the results of the initial instrument showed that the instrument was better at measuring the academic ability in mechanics for this group as for the previous one, it was still at a not satisfying degree. One item, (q10) was not varying enough, in order to perform the 2PL model analysis, therefore, we excluded it directly from the analysis. From the rest 27 items, 11 had discrimination values statistically equal

to zero and the 16 items left, had at least sufficient values and we distributed them to a new sample as a new instrument. The results from the new sample were not as good as expected, but they were still showing an improvement of the instrument. There were 7 items with values statistically equal to zero. Most of the items that had initially low discrimination values did not seem to improve (q20, q23, q26 and q3), but one did (q28). Some other items, with greater discrimination values, did not work at all (q6, q12 and q14) or at least not as well as before (q7, q8 and q11). The rest did not differentiate a lot through the phases of the analysis. The TCC was steeper for the new instrument, but not as much as expected and the TIF provided a bit less information about the latent construct. The range of θ that the most information is provided is approximately from 0 to 0.5, where the corresponding one for the initial instrument is approximately from -2 to -0.5. That means, that the new instrument is more appropriate for gathering information about individuals with an average ability.

As mentioned in Section 5.3., neither did we calibrate nor did we validate any instrument for the senior high school students, due to serious statistical violations during the analysis of the initial instrument.

About the instrument for the primary school education students, the results revealed a great improvement, almost as expected. The initial instrument seemed to be relatively easy for this group. At the same time, a total of 11 items had discrimination values close to zero, so that the new instrument would consist of the 17 remaining. After the distribution, the instrument appeared to be more effective at measuring the latent construct by providing greater discrimination values. In addition, even the 4 items that had discrimination values statistically equal to zero (q1, q2, q18 and q21), that was at a relatively low significance level. The majority of the remaining items, obtained a greater or maintained their discrimination power in the new instrument, with only exceptions, items q4, q11 and q28. The last ones, could not maintain their power, but they were still sufficient enough. The TCC was almost as expected, only with a bit lower expected score for the zero value of θ. The TIF provided almost the same amount of information (this time coming from fewer items), but the range where the most information is provided was including students with a bit lower ability (before for a range of approximately -1 to 0, now for a range of approximately -2 to -1).

The fourth instrument, the one distributed to teachers, showed also improvement, especially regarding the amount of information provided. The original one for this group, seemed to be very easy, as the items possessed

highly negative difficulty values or at least statistically equal to zero. This fact is also supported by the relatively high expected score (22,6/28) for the ability level of θ=0. Nevertheless, the instrument is still sufficiently discriminative as all the discrimination values of the items were over the cut-off value or at least statistically close to zero. After we excluded 8 items and distributed the new instrument to a new sample, we noticed that the new TCC was almost exactly as we expected, but the TIF was providing this time a much greater amount of information (and also a lower standard error). It is also important to note, that the new instrument is providing the greatest amount of information for a more specific group than the original one (before from approximately -2 to 0, now from approximately -1,5 to -1). As for the 20 items included to the new instrument, most of them had improved discrimination values. However, some discrimination values were statistically equal to zero (q4, q5, q12, q15, q20 and q23) and some others were a bit lower than expected, but still satisfyingly discriminative (q8, q11, q16, q18 and q25).

Worth mentioning, is the fact that the discrimination value of the item q11 was the only one that worsen at every one of the new instruments. Although it was discriminative at a great degree, only in the case of primary school students its value dropped dramatically (from ≈1,18 to statistically zero).

As a general conclusion of this study, we would say that every instrument showed improvement, some of them at a greater degree and some at a lower. Through this conclusion, it is once again highlighted how IRT and its statistical models can be of great importance at the development of instruments in the area of Physics' research (and furthermore of educational research). The implementation of this theory in the area of educational research can lead to more accurate measurement instruments and consequently to more accurate measurement and conclusions.

As suggestions for further research, we propose the removal of the items that did not seemed to be discriminative or their replacement with new ones and subsequently, the validation of these modified instruments. Especially, about the instrument for the senior high school students, we propose initially the redistribution of the initial one, followed by its calibration and the validation of the one coming from the analysis.

References

Alagumalai, S., & Curtis, D. D. (2005). CLASSICAL TEST THEORY. *Applied Rasch Measurement: A Book of Exemplars*, 1-14. Netherlands: Springer.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association. In Goodwin, L. D., & Leech, N. L. (2003). The Meaning of Validity in the New Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development*, 36, 181-191.

Baker, F. B. (2001). *The Basics of Item Response Theory*. U.S.A.: ERIC Clearinghouse on Assessment and Evaluation, College Park, MD. doi:10.1111/j.1365-2702.2011.03893.x

Boardley, D., Fox, C. M., & Robinson, K. L. (1999). Public policy involvement of nutrition professionals. *Journal of Nutrition Education,* 31, 248–254. In StataCorp. (2015). *Stata: Release 14*. Statistical Software. College Station, TX: StataCorp LP.

Callingham, R., & Bond, T. (2006). Research in mathematics education and Rasch measurement. *Mathematics Education Research Journal*, 18(2), 1–10. In Long, C., Wendt, H., & Dunne, T. (2011). Applying Rasch measurement in mathematics education research: steps towards a triangulated investigation into proficiency in the multiplicative conceptual field, *Educational Research and Evaluation*, 17(5), 387-407, DOI: 10.1080/13803611.2011.632661

Carson, R., & Rowlands, S. (2005). Mechanics as the logical point of entry for the enculturation into scientific thinking. *Science & Education*, 14 (3-5), 473-493. In Κώτσης, Κ. Θ. (2011). *Ερευνητική προσέγγιση του διαχρονικού χαρακτήρα των εναλλακτικών ιδεών στη διδακτική της φυσικής*. Ιωάννινα: Εκδόσεις Πανεπιστημίου Ιωαννίνων.

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care*, 44 (11 Suppl 3), 50-59. doi:10.1097/01.mlr.0000245426.10853.30

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics - Physics Education Research*, 5(2), 1–17. doi:10.1103/PhysRevSTPER.5.020103. In Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1), 10116. doi:10.3847/AER2010024

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(SUPPL. 1), 5–18. doi:10.1007/s11136-007-9198-0

Embretson S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum

Frisbie, D. A. (1988). Reliability of Scores From Teacher-Made Tests. *Educational Measurement: Issues and Practice*, 7, 25–35. doi:10.1111/j.1745-3992.1988.tb00422.x

Goodwin, L. D., & Leech, N. L. (2003). The Meaning of Validity in the New Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development*, 36, 181-191.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 39–47. doi:10.1097/01.mlr.0000245426.10853.30

Harvey, J. R., & Hammer, L. A. (1999). Item Response Theory. *THE COUNSELING PSYCHOLOGIST*, 27 (3), 353-383.

Hatzilygeroudis, I., Koutsojannis, C., Papavlasopoulos, C., & Prentzas, J. (2006). Knowledge-Based Adaptive Assessment in a Web-Based Intelligent Educational System. *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)*, 651–655. doi:10.1109/ICALT.2006.1652526

Hills, J. P. (1991). Apathy concerning grading and testing. *Phi Delta Kappa*, 72(7), 540-545. In Simsek, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education,* 4 (4), 477-489.

Κώτσης, Κ. Θ. (2004). Διαφορές αντιλήψεων σε έννοιες της Μηχανικής, Φοιτητών Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης, οι οποίοι εισήχθησαν στο Πανεπιστήμιο με τα δύο τελευταία εισαγωγικά συστήματα εξετάσεων. *Πρακτικά 4ου Πανελλήνιου Συνεδρίου για την «Διδακτική των Φυσικών Επιστημών και των Νέων Τεχνολογιών στην Εκπαίδευση»*, Αθήνα, Τόμος Α', 422-428. In Κώτσης, Κ. Θ. (2011). *Ερευνητική προσέγγιση του διαχρονικού χαρακτήρα των εναλλακτικών ιδεών στη διδακτική της φυσικής*. Ιωάννινα: Εκδόσεις Πανεπιστημίου Ιωαννίνων.

Κώτσης, Κ. Θ. (2011). *Ερευνητική προσέγγιση του διαχρονικού χαρακτήρα των εναλλακτικών ιδεών στη διδακτική της φυσικής*. Ιωάννινα: Εκδόσεις Πανεπιστημίου Ιωαννίνων.

Κώτσης, Κ. Θ., & Βέμης, Κ. (2002). Οι εναλλακτικές αντιλήψεις των παιδιών, η εννοιολογική αλλαγή και η διάρκεια γνώσης από την διδασκαλία στο Δημοτικό για φαινόμενα που στηρίζονται στον τρίτο νόμο του Νεύτωνα. *Πρακτικά 3ου Πανελλήνιου Συνεδρίου για την «Διδακτική των Φυσικών Επιστημών και των Νέων Τεχνολογιών στην Εκπαίδευση»*, Ρέθυμνο, 257-262. In Κώτσης, Κ. Θ. (2011). *Ερευνητική προσέγγιση του διαχρονικού χαρακτήρα των εναλλακτικών ιδεών στη διδακτική της φυσικής*. Ιωάννινα: Εκδόσεις Πανεπιστημίου Ιωαννίνων

Κώτσης, Κ. Θ., & Κολοβός, Χ. (2002). Οι εναλλακτικές αντιλήψεις των παιδιών, η εννοιολογική αλλαγή και η διάρκεια γνώσης από την διδασκαλία στο Δημοτικό στην έννοια της δύναμης. *Πρακτικά 3ου Πανελλήνιου Συνεδρίου για την «Διδακτική των Φυσικών Επιστημών και των Νέων Τεχνολογιών στην Εκπαίδευση»*, Ρέθυμνο, 250-256. In Κώτσης, Κ. Θ. (2011). *Ερευνητική προσέγγιση του διαχρονικού χαρακτήρα των εναλλακτικών ιδεών στη διδακτική της φυσικής*. Ιωάννινα: Εκδόσεις Πανεπιστημίου Ιωαννίνων

King, J., & Bond, T. G. (1996). A Rasch analysis of a measure of computer anxiety. *Journal of Educational Computing Research*, 14, 49–65. In StataCorp. (2015). *Stata: Release 14*. Statistical Software. College Station, TX: StataCorp LP.

Lee, Y.-J., Palazzo, D., Warnakulasooriya, R., & Pritchard, D. (2008). Measuring student learning with item response theory. *Physical Review Special Topics - Physics Education Research*, 4(1), 1–6. doi:10.1103/PhysRevSTPER.4.010102. In Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1), 10116. doi:10.3847/AER2010024

Long, C., Wendt, H., & Dunne, T. (2011). Applying Rasch measurement in mathematics education research: steps towards a triangulated investigation into proficiency in the multiplicative conceptual field, *Educational Research and Evaluation*, 17(5), 387-407, DOI: 10.1080/13803611.2011.632661

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). In Harvey, J. R. & Hammer, L. A. (1999). Item Response Theory. *THE COUNSELING PSYCHOLOGIST*, 27 (3), 353-383.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley. In Embretson S.E. & Reise S.P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum

Marshall, J. A., Hagedorn, E. A., & O'Connor, J. (2009). Anatomy of a physics test: Validation of the physics items on the Texas Assessment of Knowledge and Skills. *Physical Review Special Topics - Physics Education Research*, 5(1), 1–11. doi:10.1103/PhysRevSTPER.5.010104. In Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1), 10116. doi:10.3847/AER2010024

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment*, 14(1), 50–59. In Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(SUPPL. 1), 5–18. doi:10.1007/s11136-007-9198-0

O'Sullivan, R. G., & Chalnick, M. K. (1991). Measurement-related course work requirements for teacher certification and recertification. *Educational Measurement: Issues and Practices*, 10(1), 17-19. In Simsek, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education*, 4 (4), 477-489.

Pek, P.-K., & Poh, K.-L. (2000). Framework of a decision -theoretic tutoring system for learning of mechanics. *Journal of Science Education and Technology*, 9(4), 343–356. In Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1), 10116. doi:10.3847/AER2010024

Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch model based analysis of the Force Concept Inventory. *Physical Review Special Topics - Physics Education Research*, 6(1), 1–11. doi:10.1103/PhysRevSTPER.6.010103. In Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1), 10116. doi:10.3847/AER2010024

Simsek, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education*, 4 (4), 477-489.

StataCorp. (2015). *Stata: Release 14*. Statistical Software. College Station, TX: StataCorp LP.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond groupmean differences: The concept of item bias. *Psychological Bulletin*, 99(1), 118–128. In Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(SUPPL. 1), 5–18. doi:10.1007/s11136-007-9198-0

Traub, R. E. (1997). Classical Test Theory in Historical Perspective. Educational Measurement: Issues and Practice, 16 (4), 8-14. doi:10.1111/j.1745-3992.1997.tb00603.x

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11, 1-13. In Harvey, J. R. & Hammer, L. A. (1999). Item Response Theory. *THE COUNSELING PSYCHOLOGIST*, 27 (3), 353-383.

Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1), 10116. doi:10.3847/AER2010024

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, 78(10), 1064-1070. doi:10.1119/1.3443565. In Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1), 10116. doi:10.3847/AER2010024

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116. In Harvey, J. R. & Hammer, L. A. (1999). Item Response Theory. *THE COUNSELING PSYCHOLOGIST*, 27 (3), 353-383.

Wu, A. W., Hays, R. D., Kelly S., Malitz, F., & Bozzette, S. A. (1997). Applications of the Medical Outcomes Study health-related quality of life measures in HIV/AIDS. *Quality of Life Research*, 6, 531–554. In StataCorp. (2015). *Stata: Release 14*. Statistical Software. College Station, TX: StataCorp LP.

Φωτάρης, Π. (2011). *Αξιολόγηση πολυμεσικού μαθησιακού υλικού με εφαρμογή της θεωρίας απόκρισης ερωτήματος (IRT) σε συστήματα ηλεκτρονικής μάθησης* (Διδακτορική διατριβή, Πανεπιστήμιο Μακεδονίας). Διαθέσιμο από την ΨΗΦΙΔΑ, την Ψηφιακή Βιβλιοθήκη και Ιδρυματικό Καταθετήριο του Πανεπιστημίου Μακεδονίας (http://dspace.lib.uom.gr/handle/2159/14324).

Zhang, Z. & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, 16(4), 323-342. In Simsek, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education*, 4 (4), 477-489.

Appendix

Questionnaire for Primary School

Original Greek version

1. Πότε ασκούμε μια δύναμη;

Α) Όταν σπρώχνουμε ένα ποδήλατο. Β) Όταν σπρώχνουμε έναν τοίχο.
Γ) Και στις δυο περιπτώσεις.

2. Ένα παιδί πετά μια πέτρα, πότε το παιδί ασκεί δύναμη στην πέτρα;

Α) Όταν φεύγει από το χέρι του. Β) Όταν είναι στον αέρα.

3. Όταν περπατάμε, σπρώχνουμε το έδαφος:

Α) Προς τα εμπρός. Β) Προς τα πίσω.

4. Ένα αυτοκίνητο πότε κινείται με μεγαλύτερη ασφάλεια σε παγωμένο δρόμο; Όταν είναι:

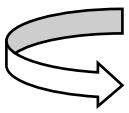
Α) άδειο. Β) φορτωμένο.

5. Ένα ελαφρύ επιβατικό αυτοκίνητο κι ένα βαρύ φορτηγό βρίσκονται σε κόκκινο φανάρι. Όταν ανάψει πράσινο «γκαζώνουν» το ίδιο, ποιο θα ξεκινήσει πιο γρήγορα;

Α) Το επιβατικό. Β) Το φορτηγό.

6. Η βαρύτητα στη Σελήνη είναι μικρότερη από τη βαρύτητα της Γης. Το βάρος μιας σοκολάτας είναι:

Α) μικρότερο στη Γη απ' ότι στη Σελήνη. Β) μεγαλύτερο στη Γη απ' ότι στη Σελήνη.
Γ) το ίδιο και στη Γη και στη Σελήνη.



7. Ένα μήλο έχει:

Α) το ίδιο βάρος στη Γη και στη Σελήνη.
Β) την ίδια μάζα στη Γη και στη Σελήνη.
Γ) ίδιο βάρος και ίδια μάζα στη Γη και στη Σελήνη.

8. Σε μια μηλιά, ένα μήλο στέκεται στο κλαδί του κι ένα άλλο πέφτει προς το έδαφος. Ποιο από τα δυο μήλα παράγει έργο;

Α) Αυτό που πέφτει. Β) Αυτό που στέκεται στο κλαδί.
Γ) Και τα δυο μήλα. Δ) Κανένα από τα δυο

9. Ανεβαίνεις στο δεύτερο όροφο του σπιτιού σου, τη μια φορά άδειος και την άλλη φορτωμένος με πράγματα. Πότε ξοδεύεις μεγαλύτερο έργο;

Α) Όταν είσαι άδειος.
Β) Όταν είσαι φορτωμένος.
Γ) Το ίδιο.

10. Δυο αθλητές με το ίδιο βάρος και το ίδιο ύψος τρέχουν σε απόσταση 100 μέτρων. Ποιος καταναλώνει περισσότερη ενέργεια;

Α) Αυτός που τερματίζει πρώτος.
Β) Αυτός που τερματίζει δεύτερος.
Γ) Καταναλώνουν την ίδια.

11. Ένα φορτηγό πότε έχει μεγαλύτερη ενέργεια;

Α) Όταν κινείται.
Β) Όταν είναι στάσιμο.
Γ) Έχει πάντοτε την ίδια.

Translated English version

1. In which case do we apply a force?

A) When we push a bicycle. B) When we push a wall.
C) In both cases.

2. A child throws a rock, when does the child apply a force on the rock?

A) When it leaves its hand. B) When it is in the air.

3. **When we walk, we push the ground:**

A) to the front. B) to the back.

4. When does a car move with greater safety on an icy road? When it is:

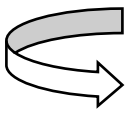
A) empty. B) loaded.

5. A light private car and a heavy truck are stopped at a red traffic light. When it goes green they both hit the gas the same, which one is going to start quicker?

A) The private car. B) The truck.

6. The gravity on the moon is weaker than the one on the Earth. The weight of a chocolate is:

A) smaller on the Earth than on the moon. B) greater on the Earth than on the moon.
C) the same on the Earth and on the moon.



7. An apple has:

A) the same weight on the Earth and on the moon.
B) the same mass on the Earth and on the moon.
C) the same weight and the same mass on the Earth and on the moon.

8. On an apple tree, an apple is hanging on its branch and another one falls towards the ground. Which one of them produces work?

A) The one falling. B) The one hanging on its branch.
C) Both. D) None.

9. **You go up to the second floor of your house, the first time empty-handed and the second one loaded with stuff. When do you produce greater work?**

A) When you are empty-handed.
B) When you are loaded.
C) The same.

10. Two athletes with the same weight and height run a distance of 100 meters. Which one requires more energy?

A) The one finishing first.
B) The one finishing second.
C) They both require the same.

11. When does a truck possess greater amount of energy?

A) When it moves.
B) When it is stopped.
C) It has always the same amount.

Questionnaire for Junior High School

Original Greek Version

1. Δύναμη είναι η αιτία που ένα σώμα:

Α) παραμορφώνεται. Β) αλλάζει την κινητική του κατάσταση.
Γ) που κάνει και τα δυο.

2. Πότε ασκούμε μια δύναμη;

Α) Όταν σπρώχνουμε ένα ποδήλατο. Β) Όταν σπρώχνουμε έναν τοίχο.
Γ) Και στις δυο περιπτώσεις.

3. Πότε ενεργεί μια δύναμη σ' ένα σώμα;

Α) Όταν αρχίζουμε να κινούμε ένα σώμα. Β) Όταν σταματάμε ένα σώμα που κινείται.
Γ) Και στις δυο περιπτώσεις.

4. Ένα παιδί πετά μια πέτρα, πότε το παιδί ασκεί δύναμη στην πέτρα;

Α) Όταν φεύγει από το χέρι του. Β) Όταν είναι στον αέρα.

5. Σκοντάφτω σε μια πέτρα, την οποία και μετακινώ. Η πέτρα:

Α) Ασκεί μια δύναμη και σε μένα. Β) Δεν ασκεί δύναμη σε μένα.

6. Χτυπώ το χέρι μου σε ένα τραπέζι και με πονάει το χέρι, γιατί;

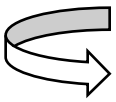
Α) Άσκησα δύναμη στο τραπέζι. Β) Το τραπέζι άσκησε δύναμη σε μένα.

7. Όταν περπατάμε, σπρώχνουμε το έδαφος:

Α) Προς τα εμπρός. Β) Προς τα πίσω.

8. Πότε αναπτύσσεται μεγαλύτερη τριβή μεταξύ ενός αυτοκινήτου και του δρόμου;

Α) Όταν είναι στεγνός. Β) Όταν είναι βρεγμένος.



9. Ένα ελαφρύ επιβατικό αυτοκίνητο κι ένα βαρύ φορτηγό βρίσκονται σε κόκκινο φανάρι. Όταν ανάψει πράσινο «γκαζώνουν» το ίδιο, ποιο θα ξεκινήσει πιο γρήγορα;

Α) Το επιβατικό. Β) Το φορτηγό.

10. Όταν ξεκινάει το αυτοκίνητο, οι επιβάτες κινούνται προς τα πίσω:

Α) εξαιτίας της ταχύτητας. Β) εξαιτίας της αδράνειας. Γ) Δεν ξέρω.

11. Το βάρος ενός σώματος είναι:

Α) δύναμη. Β) ιδιότητα του σώματος. Γ) η μάζα του σώματος.

12. Σε μια μηλιά, ένα μήλο στέκεται στο κλαδί του κι ένα άλλο πέφτει προς το έδαφος. Ποιο από τα δυο μήλα παράγει έργο;

Α) Αυτό που πέφτει. Β) Αυτό που στέκεται στο κλαδί.
Γ) Και τα δυο μήλα. Δ) Κανένα από τα δυο

13. Δυο αθλητές με το ίδιο βάρος και το ίδιο ύψος τρέχουν μια απόσταση 100 μέτρων. Ποιος έχει μεγαλύτερη ισχύ;

Α) Αυτός που τερματίζει πρώτος. Β) Αυτός που τερματίζει δεύτερος.
Γ) Και οι δύο έχουν την ίδια ισχύ.

14. Σε μια πορτοκαλιά ένα πορτοκάλι είναι πάνω στο δέντρο κι ένα άλλο πέφτει. Ποιο από τα δυο πορτοκάλια έχει ενέργεια;

Α) Αυτό που πέφτει. Β) Αυτό που είναι στο δέντρο. Γ) Και τα δυο.

15. Γιατί οι αθλητές που πηδούν άλμα σε μήκος, παίρνουν φόρα και τρέχουν;

Α) Για να ξεπεράσουν την αντίσταση από τον αέρα.
Β) Για να αποκτήσουν μεγαλύτερη ενέργεια.
Γ) Δεν ξέρω.

16. Δυο αρσιβαρίστες σηκώνουν το ίδιο βάρος. Ποιος ξοδεύει μεγαλύτερη ενέργεια;

Α) Αυτός που είναι πιο ψηλός. Β) Αυτός που είναι πιο κοντός.
Γ) Το ίδιο και οι δυο.

Translated English version

1. Force is the cause that a body:

A) is deformed. B) changes its kinetic state.
C) both of the above.

2. In which case do we apply a force?

A) When we push a bicycle. B) When we push a wall.
C) In both cases.

3. When does a force is being applied on a body?

A) When we start moving it. B) When we stop a moving body.
C) In both cases.

4. A child throws a rock, when does the child apply a force on the rock?

A) When it leaves its hand. B) When it is in the air.

5. I fall on a rock, which I move. The rock:

A) Applies a force on me as well. B) Does not apply any force on me.

6. I slam my hand on the table and it hurts, why?

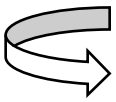
A) I applied a force on the table. B) The table applied a force on me.

7. When we walk, we push the ground:

A) to the front. B) to the back.

8. In which case more friction between a car and the road is produced?

A) When the road is dry. B) When the road is wet.



9. A light private car and a heavy truck are stopped at a red traffic light. When it goes green they both hit the gas the same, which one is going to start quicker?

A) The private car. B) The truck

10. When a car starts, the passengers are moved to the back:

A) because of the velocity. B) because of the inertia. C) I do not know.

11. The weight of a body is:

A) a force. B) a characteristic of the body. C) the mass of the body.

12. On an apple tree, an apple is hanging on its branch and another one falls towards the ground. Which one of them produces work?

A) The one falling. B) The one hanging on its branch.
C) Both. D) None.

13. Two athletes with the same weight and height run a distance of 100 meters. Which has more power?

A) The one finishing first.
B) The one finishing second.
C) Both have the same.

14. On an orange tree one orange is on the tree and the other one is falling down. Which one of them has energy?

A) The one falling. B) The one on the tree. C) Both.

15. Why do athletes of the long jump, speed up and run?

A) To overcome the resistance of the air.
B) To gain greater amount of energy.
C) I do not know.

16. Two weightlifters lift the same weight. Which one consumes more energy?

A) The taller one. B) The shorter one.
C) Both the same.

Questionnaire for Primary School Education Students

Original Greek version

Ερωτηματολόγιο Φοιτητών ΠΤΔΕ

Το ερωτηματολόγιο αυτό είναι ανώνυμο και δεν αποθηκεύει προσωπικές πληροφορίες.

* Απαιτείται

1. **Σε ποιο έτος φοίτησης βρίσκεσαι;** *
   Να επισημαίνεται μόνο μία έλλειψη.
   - ◯ 1ο
   - ◯ 2ο
   - ◯ 3ο
   - ◯ 4ο
   - ◯ 5ο και άνω

2. **Δύναμη είναι η αιτία που ένα σώμα:** *
   Να επισημαίνεται μόνο μία έλλειψη.
   - ◯ παραμορφώνεται.
   - ◯ αλλάζει την κινητική του κατάσταση.
   - ◯ που κάνει και τα δυο.

3. **Με έναν φίλο σου κάνεις «κόλλα-πέντε». Τι κατεύθυνση έχουν οι δυνάμεις που ασκεί ο ένας στον άλλο στα χέρια σας;** *
   Να επισημαίνεται μόνο μία έλλειψη.
   - ◯ Ίδια διεύθυνση και φορά.
   - ◯ Ίδια διεύθυνση και αντίθετη φορά.
   - ◯ Διαφορετική διεύθυνση και φορά.

4. **Πότε ασκούμε μια δύναμη;** *

Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Όταν σπρώχνουμε ένα ποδήλατο.
- ◯ Όταν σπρώχνουμε έναν τοίχο.
- ◯ Και στις δυο περιπτώσεις.

5. **Πότε ενεργεί μια δύναμη σ' ένα σώμα;** *

Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Όταν αρχίζουμε να κινούμε ένα σώμα.
- ◯ Όταν σταματάμε ένα σώμα που κινείται.
- ◯ Και στις δυο περιπτώσεις.

6. **Σκοντάφτω σε μια πέτρα, την οποία και μετακινώ. Η πέτρα:** *

Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Ασκεί μια δύναμη και σε μένα.
- ◯ Δεν ασκεί δύναμη σε μένα.

7. **Χτυπώ το χέρι μου σε ένα τραπέζι και με πονάει το χέρι, γιατί:** *

Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Άσκησα δύναμη στο τραπέζι.
- ◯ Το τραπέζι άσκησε δύναμη σε μένα.

8. **Είσαι ακίνητος στην πρωινή προσευχή. Επέλεξε ποιες δυνάμεις υπάρχουν (μόνο μια απάντηση).** *

Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Η δύναμη που ασκώ στο έδαφος.
- ◯ Η δύναμη που ασκεί το έδαφος σε μένα.
- ◯ Και οι δυο προηγούμενες.
- ◯ Καμιά δύναμη.

9. **Όταν περπατάμε, σπρώχνουμε το έδαφος:** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Προς τα εμπρός.
- ◯ Προς τα πίσω.

10. **Πότε αναπτύσσεται μεγαλύτερη τριβή μεταξύ ενός αυτοκινήτου και του δρόμου;** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Όταν είναι στεγνός.
- ◯ Όταν είναι βρεγμένος.

11. **Το βάρος ενός σώματος είναι:** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ δύναμη.
- ◯ ιδιότητα του σώματος.
- ◯ η μάζα του σώματος.

12. **Η βαρύτητα στη Σελήνη είναι μικρότερη από τη βαρύτητα της Γης. Το βάρος μιας σοκολάτας είναι:** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ μικρότερο στη Γη απ' ότι στη Σελήνη.
- ◯ μεγαλύτερο στη Γη απ' ότι στη Σελήνη.
- ◯ το ίδιο και στη Γη και στη Σελήνη.

13. **Ένα μήλο έχει:** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ το ίδιο βάρος στη Γη και στη Σελήνη.
- ◯ την ίδια μάζα στη Γη και στη Σελήνη.
- ◯ ίδιο βάρος και ίδια μάζα στη Γη και στη Σελήνη.

14. **Ανεβαίνεις στο δεύτερο όροφο του σπιτιού σου, τη μια φορά άδειος και την άλλη φορτωμένος με πράγματα. Πότε ξοδεύεις μεγαλύτερο έργο;** *

Να επισημαίνεται μόνο μία έλλειψη.

- Όταν είσαι άδειος.
- Όταν είσαι φορτωμένος.
- Το ίδιο.

15. **Σε μια πορτοκαλιά ένα πορτοκάλι είναι πάνω στο δέντρο κι ένα άλλο πέφτει. Ποιο από τα δυο πορτοκάλια έχει ενέργεια;** *

Να επισημαίνεται μόνο μία έλλειψη.

- Αυτό που πέφτει.
- Αυτό που είναι στο δέντρο.
- Και τα δυο.

16. **Είστε στο μπαλκόνι του σπιτιού σας και κοιτάζετε το ηλιοβασίλεμα. Που έχετε μεγαλύτερη ενέργεια ως προς το έδαφος της Γης;** *

Να επισημαίνεται μόνο μία έλλειψη.

- Όταν είστε στον πρώτο όροφο.
- Όταν είστε στο δεύτερο όροφο.
- Έχετε την ίδια ενέργεια παντού.
- Δεν έχετε πουθενά ενέργεια.

17. **Γιατί οι αθλητές που πηδούν άλμα σε μήκος, παίρνουν φόρα και τρέχουν;** *

Να επισημαίνεται μόνο μία έλλειψη.

- Για να ξεπεράσουν την αντίσταση από τον αέρα.
- Για να αποκτήσουν μεγαλύτερη ενέργεια.
- Δεν ξέρω.

18. **Δυο αρσιβαρίστες σηκώνουν το ίδιο βάρος. Ποιος ξοδεύει μεγαλύτερη ενέργεια;** *

Να επισημαίνεται μόνο μία έλλειψη.

- Αυτός που είναι πιο ψηλός.
- Αυτός που είναι πιο κοντός.
- Το ίδιο και οι δυο.

Translated English version

1. **On which year of studies are you:**
   - 1st
   - 2nd
   - 3rd
   - 4th
   - 5th or greater.

2. **Force is the cause that a body:**
   - is deformed.
   - changes its kinetic state.
   - both of the above.

3. **With a friend you do "hi-5". What are the directions of the forces applied to your hands by each other?**
   - Same direction and orientation.
   - Same direction and opposite orientation.
   - Different direction and orientation.

4. **In which case do we apply a force?**
   - When we push a bicycle.
   - When we push a wall.
   - In both cases.

5. **When does a force is being applied on a body?**
   - When we start moving it.
   - When we stop a moving body.
   - In both cases.

6. **I fall on a rock, which I move. The rock:**
   - Applies a force on me as well.
   - Does not apply any force on me.

7. **I slam my hand on the table and it hurts, why?**
   - I applied a force on the table.
   - The table applied a force on me.

8. **You are standing in the morning prayer. Choose which forces are applied (only one correct answer):**
   - The force I apply to the ground
   - The force that the ground applies on me
   - Both of them
   - None of them

9. **When we walk, we push the ground:**
   - to the front.
   - to the back.

10. In which case more friction between a car and the road is produced?
- When the road is dry.
- When the road is wet.

11. The weight of a body is:
- a force.
- a characteristic of the body.
- the mass of the body.

12. The gravity on the moon is weaker than the one on the Earth. The weight of a chocolate is:
- smaller on the Earth than on the moon.
- greater on the Earth than on the moon.
- the same on the Earth and on the moon.

13. An apple has:
- the same weight on the Earth and on the moon.
- the same mass on the Earth and on the moon.
- the same weight and the same mass on the Earth and on the moon.

14. You go up to the second floor of your house, the first time empty-handed and the second one loaded with stuff. When do you produce greater work?
- When you are empty-handed.
- When you are loaded.
- The same.

15. On an orange tree one orange is on the tree and the other one is falling down. Which one of them has energy?
- The one falling.
- The one on the tree.
- Both.

16. You are on the balcony of your house and look at the sunset. In which case do you have the greatest amount of energy, in relation to the ground of the Earth?
- When you are on the first floor.
- When you are on the second floor.
- You have the same amount of energy in both cases.
- You have nowhere energy.

17. Why do athletes of the long jump, speed up and run?
- To overcome the resistance of the air.
- To gain greater amount of energy.
- I do not know.

18. Two weightlifters lift the same weight. Which one consumes more energy?
- The taller one.
- The shorter one.
- Both the same.

Online Questionnaire for Teachers

Ερωτηματολόγιο Εκπαιδευτικών

Το ερωτηματολόγιο αυτό είναι ανώνυμο και δεν αποθηκεύει προσωπικές πληροφορίες.

* Απαιτείται

1. **Είμαι εκπαιδευτικός:** *
   Να επισημαίνεται μόνο μία έλλειψη.
   - ◯ πρωτοβάθμιας εκπαίδευσης
   - ◯ δευτεροβάθμιας εκπαίδευσης

2. **Δύναμη είναι η αιτία που ένα σώμα:** *
   Να επισημαίνεται μόνο μία έλλειψη.
   - ◯ παραμορφώνεται.
   - ◯ αλλάζει την κινητική του κατάσταση.
   - ◯ που κάνει και τα δυο.

3. **Με έναν φίλο σου κάνεις «κόλλα-πέντε». Τι κατεύθυνση έχουν οι δυνάμεις που ασκεί ο ένας στον άλλο στα χέρια σας;** *
   Να επισημαίνεται μόνο μία έλλειψη.
   - ◯ Ίδια διεύθυνση και φορά.
   - ◯ Ίδια διεύθυνση και αντίθετη φορά.
   - ◯ Διαφορετική διεύθυνση και φορά.

4. **Πότε ενεργεί μια δύναμη σ' ένα σώμα;** *
   Να επισημαίνεται μόνο μία έλλειψη.
   - ◯ Όταν αρχίζουμε να κινούμε ένα σώμα.
   - ◯ Όταν σταματάμε ένα σώμα που κινείται.
   - ◯ Και στις δυο περιπτώσεις.

5. **Πότε ένας ποδοσφαιριστής ασκεί δύναμη σε μια μπάλα;** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Όταν τη σουτάρει.
- ◯ Όταν κινείται προς τα δίχτυα.
- ◯ Και στις δυο περιπτώσεις.
- ◯ Σε καμιά περίπτωση.

6. **Σκοντάφτω σε μια πέτρα, την οποία και μετακινώ. Η πέτρα:** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Ασκεί μια δύναμη και σε μένα.
- ◯ Δεν ασκεί δύναμη σε μένα.

7. **Χτυπώ το χέρι μου σε ένα τραπέζι και με πονάει το χέρι, γιατί;** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Άσκησα δύναμη στο τραπέζι.
- ◯ Το τραπέζι άσκησε δύναμη σε μένα.

8. **Είσαι ακίνητος στην πρωινή προσευχή. Σημείωσε ποιες δυνάμεις υπάρχουν (μόνο μια απάντηση).** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Η δύναμη που ασκώ στο έδαφος.
- ◯ Η δύναμη που ασκεί το έδαφος σε μένα.
- ◯ Και οι δυο προηγούμενες.
- ◯ Καμιά δύναμη.

9. **Όταν περπατάμε, σπρώχνουμε το έδαφος:** *
Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Προς τα εμπρός.
- ◯ Προς τα πίσω.

10. **Πότε αναπτύσσεται μεγαλύτερη τριβή μεταξύ ενός αυτοκινήτου και του δρόμου;** *

Να επισημαίνεται μόνο μία έλλειψη.

- Όταν είναι στεγνός.
- Όταν είναι βρεγμένος.

11. **Όταν ξεκινάει το αυτοκίνητο, οι επιβάτες κινούνται προς τα πίσω:** *

Να επισημαίνεται μόνο μία έλλειψη.

- εξαιτίας της ταχύτητας.
- εξαιτίας της αδράνειας.
- Δεν ξέρω.

12. **Το βάρος ενός σώματος είναι:** *

Να επισημαίνεται μόνο μία έλλειψη.

- δύναμη.
- ιδιότητα του σώματος.
- η μάζα του σώματος.

13. **Η βαρύτητα στη Σελήνη είναι μικρότερη από τη βαρύτητα της Γης. Το βάρος μιας σοκολάτας είναι:** *

Να επισημαίνεται μόνο μία έλλειψη.

- μικρότερο στη Γη απ' ότι στη Σελήνη.
- μεγαλύτερο στη Γη απ' ότι στη Σελήνη.
- το ίδιο και στη Γη και στη Σελήνη.

14. **Ένα μήλο έχει:** *

Να επισημαίνεται μόνο μία έλλειψη.

- το ίδιο βάρος στη Γη και στη Σελήνη.
- την ίδια μάζα στη Γη και στη Σελήνη.
- ίδιο βάρος και ίδια μάζα στη Γη και στη Σελήνη.

15. **Όταν είσαι στη θάλασσα και σηκώνεις μια πέτρα μέσα από το νερό, το βάρος της πέτρας είναι:** *

Να επισημαίνεται μόνο μία έλλειψη.

- μεγαλύτερο στο νερό.
- μικρότερο στο νερό.
- το ίδιο.

16. **Σε μια μηλιά, ένα μήλο στέκεται στο κλαδί του κι ένα άλλο πέφτει προς το έδαφος. Ποιο από τα δυο μήλα παράγει έργο;** *

Να επισημαίνεται μόνο μία έλλειψη.

- Αυτό που πέφτει.
- Αυτό που στέκεται στο κλαδί.
- Και τα δυο μήλα.
- Κανένα από τα δυο

17. **Δυο αθλητές με το ίδιο βάρος και το ίδιο ύψος τρέχουν μια απόσταση 100 μέτρων. Ποιος έχει μεγαλύτερη ισχύ;** *

Να επισημαίνεται μόνο μία έλλειψη.

- Αυτός που τερματίζει πρώτος.
- Αυτός που τερματίζει δεύτερος.
- Και οι δύο έχουν την ίδια ισχύ.

18. **Σε μια πορτοκαλιά ένα πορτοκάλι είναι πάνω στο δέντρο κι ένα άλλο πέφτει. Ποιο από τα δυο πορτοκάλια έχει ενέργεια;** *

Να επισημαίνεται μόνο μία έλλειψη.

- Αυτό που πέφτει.
- Αυτό που είναι στο δέντρο.
- Και τα δυο.

19. **Είστε στο μπαλκόνι του σπιτιού σας και κοιτάζετε το ηλιοβασίλεμα. Που έχετε μεγαλύτερη ενέργεια ως προς το έδαφος της Γης;** *

Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Όταν είστε στον πρώτο όροφο.
- ◯ Όταν είστε στο δεύτερο όροφο.
- ◯ Έχετε την ίδια ενέργεια παντού.
- ◯ Δεν έχετε πουθενά ενέργεια.

20. **Γιατί οι αθλητές που πηδούν άλμα σε μήκος, παίρνουν φόρα και τρέχουν;** *

Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Για να ξεπεράσουν την αντίσταση από τον αέρα.
- ◯ Για να αποκτήσουν μεγαλύτερη ενέργεια.
- ◯ Δεν ξέρω.

21. **Δυο αρσιβαρίστες σηκώνουν το ίδιο βάρος. Ποιος ξοδεύει μεγαλύτερη ενέργεια;** *

Να επισημαίνεται μόνο μία έλλειψη.

- ◯ Αυτός που είναι πιο ψηλός.
- ◯ Αυτός που είναι πιο κοντός.
- ◯ Το ίδιο και οι δυο.

Printed Questionnaire for Teachers

Original Greek version

1. Δύναμη είναι η αιτία που ένα σώμα:

Α) παραμορφώνεται. Β) αλλάζει την κινητική του κατάσταση.

Γ) που κάνει και τα δυο.

2. Με έναν φίλο σου κάνεις «κόλλα-πέντε». Τι κατεύθυνση έχουν οι δυνάμεις που ασκεί ο ένας στον άλλο στα χέρια σας;

Α) Ίδια διεύθυνση και φορά. Β) Ίδια διεύθυνση και αντίθετη φορά.

Γ) Διαφορετική διεύθυνση και φορά.

3. Πότε ενεργεί μια δύναμη σ' ένα σώμα;

Α) Όταν αρχίζουμε να κινούμε ένα σώμα. Β) Όταν σταματάμε ένα σώμα που κινείται.
Γ) Και στις δυο περιπτώσεις.

4. Πότε ένας ποδοσφαιριστής ασκεί δύναμη σε μια μπάλα;

Α) Όταν τη σουτάρει. Β) Όταν κινείται προς τα δίχτυα.
Γ) Και στις δυο περιπτώσεις. Δ) Σε καμιά περίπτωση.

5. Σκοντάφτω σε μια πέτρα, την οποία και μετακινώ. Η πέτρα:

Α) Ασκεί μια δύναμη και σε μένα. Β) Δεν ασκεί δύναμη σε μένα.

6. Χτυπώ το χέρι μου σε ένα τραπέζι και με πονάει το χέρι, γιατί;

Α) Άσκησα δύναμη στο τραπέζι. Β) Το τραπέζι άσκησε δύναμη σε μένα.

7. Είσαι ακίνητος στην πρωινή προσευχή. Σημείωσε ποιες δυνάμεις υπάρχουν (μόνο μια απάντηση).

Α) Η δύναμη που ασκώ στο έδαφος. Β) Η δύναμη που ασκεί το έδαφος σε μένα.
Γ) Και οι δυο προηγούμενες. Δ) Καμιά δύναμη.

8. Όταν περπατάμε, σπρώχνουμε το έδαφος:

Α) Προς τα εμπρός. Β) Προς τα πίσω.

9. Πότε αναπτύσσεται μεγαλύτερη τριβή μεταξύ ενός αυτοκινήτου και του δρόμου;

Α) Όταν είναι στεγνός. Β) Όταν είναι βρεγμένος.

10. Όταν ξεκινάει το αυτοκίνητο, οι επιβάτες κινούνται προς τα πίσω:

Α) εξαιτίας της ταχύτητας. Β) εξαιτίας της αδράνειας. Γ) Δεν ξέρω.

11. Το βάρος ενός σώματος είναι:

Α) δύναμη. Β) ιδιότητα του σώματος. Γ) η μάζα του σώματος.

12. Η βαρύτητα στη Σελήνη είναι μικρότερη από τη βαρύτητα της Γης. Το βάρος μιας σοκολάτας είναι:

Α) μικρότερο στη Γη απ' ότι στη Σελήνη. Β) μεγαλύτερο στη Γη απ' ότι στη Σελήνη.
Γ) το ίδιο και στη Γη και στη Σελήνη.

13. Ένα μήλο έχει:

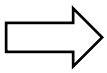
Α) το ίδιο βάρος στη Γη και στη Σελήνη. Β) την ίδια μάζα στη Γη και στη Σελήνη.
Γ) ίδιο βάρος και ίδια μάζα στη Γη και στη Σελήνη.

14. Όταν είσαι στη θάλασσα και σηκώνεις μια πέτρα μέσα από το νερό, το βάρος της πέτρας είναι:

Α) μεγαλύτερο στο νερό. Β) μικρότερο στο νερό. Γ) το ίδιο.

15. Σε μια μηλιά, ένα μήλο στέκεται στο κλαδί του κι ένα άλλο πέφτει προς το έδαφος. Ποιο από τα δυο μήλα παράγει έργο;

Α) Αυτό που πέφτει. Β) Αυτό που στέκεται στο κλαδί.
Γ) Και τα δυο μήλα. Δ) Κανένα από τα δυο



16. Δυο αθλητές με το ίδιο βάρος και το ίδιο ύψος τρέχουν μια απόσταση 100 μέτρων. Ποιος έχει μεγαλύτερη ισχύ;

Α) Αυτός που τερματίζει πρώτος. Β) Αυτός που τερματίζει δεύτερος.
Γ) Και οι δύο έχουν την ίδια ισχύ.

17. Σε μια πορτοκαλιά ένα πορτοκάλι είναι πάνω στο δέντρο κι ένα άλλο πέφτει. Ποιο από τα δυο πορτοκάλια έχει ενέργεια;

Α) Αυτό που πέφτει. Β) Αυτό που είναι στο δέντρο. Γ) Και τα δυο.

18. Είστε στο μπαλκόνι του σπιτιού σας και κοιτάζετε το ηλιοβασίλεμα. Που έχετε μεγαλύτερη ενέργεια ως προς το έδαφος της Γης;

Α) Όταν είστε στον πρώτο όροφο. Β) Όταν είστε στο δεύτερο όροφο.
Γ) Έχετε την ίδια ενέργεια παντού. Δ) Δεν έχετε πουθενά ενέργεια.

19. Γιατί οι αθλητές που πηδούν άλμα σε μήκος, παίρνουν φόρα και τρέχουν;

Α) Για να ξεπεράσουν την αντίσταση από τον αέρα. Β) Για να αποκτήσουν μεγαλύτερη ενέργεια.
Γ) Δεν ξέρω.

20. Δυο αρσιβαρίστες σηκώνουν το ίδιο βάρος. Ποιος ξοδεύει μεγαλύτερη ενέργεια;

Α) Αυτός που είναι πιο ψηλός. Β) Αυτός που είναι πιο κοντός. Γ) Το ίδιο και οι δυο.

Translated English Version

(The first extra question of the online questionnaire is about being a primary or a secondary school teacher.)

1. Force is the cause that a body:

A) is deformed. B) changes its kinetic state.
C) both of the above.

2. With a friend you do "hi-5". What are the directions of the forces applied to your hands by each other?

A) Same direction and orientation.
B) Same direction and opposite orientation.
C) Different direction and orientation.

3. When does a force is being applied on a body?

A) When we start moving it. B) When we stop a moving body.
C) In both cases.

4. When does a football player apply force on a ball?

A) When he kicks it. B) When he moves to the nets.
C) In both cases. D) In neither one of these cases.

5. I fall on a rock, which I move. The rock:

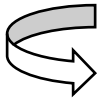
A) Applies a force on me as well. B) Does not apply any force on me.

6. I slam my hand on the table and it hurts, why?

A) I applied a force on the table. B) The table applied a force on me.

7. You are standing in the morning prayer. Choose which forces are applied (only one correct answer):

A) The force I apply to the ground
B) The force that the ground applies on me
C) Both of them
D) None of them



8. When we walk, we push the ground:

A) to the front. B) to the back.

9. In which case more friction between a car and the road is produced?

A) When the road is dry. B) When the road is wet.

10. When a car starts, the passengers are moved to the back:

A) because of the velocity. B) because of the inertia. C) I do not know.

11. The weight of a body is:

A) a force. B) a characteristic of the body. C) the mass of the body.

12. The gravity on the moon is weaker than the one on the Earth. The weight of a chocolate is:

A) smaller on the Earth than on the moon. B) greater on the Earth than on the moon.
C) the same on the Earth and on the moon.

13. An apple has:

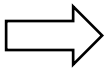
A) the same weight on the Earth and on the moon.
B) the same mass on the Earth and on the moon.
C) the same weight and the same mass on the Earth and on the moon.

14. When you are in the sea and you lift a rock in the water, the weight of the rock is:

A) greater in the water. B) lower in the water. Γ) the same.

15. On an apple tree, an apple is hanging on its branch and another one falls towards the ground. Which one of them produces work?

A) The one falling. B) The one hanging on its branch.
C) Both. D) None.



16. Two athletes with the same weight and height run a distance of 100 meters. Which has more power?

A) The one finishing first.
B) The one finishing second.
C) Both have the same.

17. On an orange tree one orange is on the tree and the other one is falling down. Which one of them has energy?

A) The one falling. B) The one on the tree. C) Both.

18. You are on the balcony of your house and look at the sunset. In which case do you have the greatest amount of energy, in relation to the ground of the Earth?

A) When you are on the first floor.
B) When you are on the second floor.
C) You have the same amount of energy in both cases.

19. Why do athletes of the long jump, speed up and run?

A) To overcome the resistance of the air.
B) To gain greater amount of energy.
C) I do not know.

20. Two weightlifters lift the same weight. Which one consumes more energy?

A) The taller one. B) The shorter one.
C) Both the same.

Descriptive statistics for Primary School Students

Grade

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5th	60	49,2	49,2	49,2
	6th	62	50,8	50,8	100,0
	Total	122	100,0	100,0	

q3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	50	41,0	41,0	41,0
	Correct	72	59,0	59,0	100,0
	Total	122	100,0	100,0	

q6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	34	27,9	27,9	27,9
	Correct	88	72,1	72,1	100,0
	Total	122	100,0	100,0	

q11

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	31	25,4	25,4	25,4
	Correct	91	74,6	74,6	100,0
	Total	122	100,0	100,0	

q13

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	37	30,3	30,3	30,3
	Correct	85	69,7	69,7	100,0
	Total	122	100,0	100,0	

q14

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	30	24,6	24,6	24,6
	Correct	92	75,4	75,4	100,0
	Total	122	100,0	100,0	

q17

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	24	19,7	19,7	19,7
	Correct	98	80,3	80,3	100,0
	Total	122	100,0	100,0	

q18

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	27	22,1	22,1	22,1
	Correct	95	77,9	77,9	100,0
	Total	122	100,0	100,0	

q20

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	49	40,2	40,2	40,2
	Correct	73	59,8	59,8	100,0
	Total	122	100,0	100,0	

q21

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	18	14,8	14,8	14,8
	Correct	104	85,2	85,2	100,0
	Total	122	100,0	100,0	

q22

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	90	73,8	73,8	73,8
	Correct	32	26,2	26,2	100,0
	Total	122	100,0	100,0	

q27

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	27	22,1	22,1	22,1
	Correct	95	77,9	77,9	100,0
	Total	122	100,0	100,0	

Descriptive statistics for Junior High School Students

Grade

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2nd	58	50,0	50,0	50,0
	3rd	58	50,0	50,0	100,0
	Total	116	100,0	100,0	

q1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	58	50,0	50,0	50,0
	Correct	58	50,0	50,0	100,0
	Total	116	100,0	100,0	

q3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	49	42,2	42,2	42,2
	Correct	67	57,8	57,8	100,0
	Total	116	100,0	100,0	

q4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	78	67,2	67,2	67,2
	Correct	38	32,8	32,8	100,0
	Total	116	100,0	100,0	

q6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	13	11,2	11,2	11,2
	Correct	103	88,8	88,8	100,0
	Total	116	100,0	100,0	

q7

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	41	35,3	35,3	35,3
	Correct	75	64,7	64,7	100,0
	Total	116	100,0	100,0	

q8

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	57	49,1	49,1	49,1
	Correct	59	50,9	50,9	100,0
	Total	116	100,0	100,0	

q11

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	25	21,6	21,6	21,6
	Correct	91	78,4	78,4	100,0
	Total	116	100,0	100,0	

q12

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	38	32,8	32,8	32,8
	Correct	78	67,2	67,2	100,0
	Total	116	100,0	100,0	

q14

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	20	17,2	17,2	17,2
	Correct	96	82,8	82,8	100,0
	Total	116	100,0	100,0	

q15

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	60	51,7	51,7	51,7
	Correct	56	48,3	48,3	100,0
	Total	116	100,0	100,0	

q16

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	86	74,1	74,1	74,1
	Correct	30	25,9	25,9	100,0
	Total	116	100,0	100,0	

q20

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	74	63,8	63,8	63,8
	Correct	42	36,2	36,2	100,0
	Total	116	100,0	100,0	

q23

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	62	53,4	53,4	53,4
	Correct	54	46,6	46,6	100,0
	Total	116	100,0	100,0	

q24

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	90	77,6	77,6	77,6
	Correct	26	22,4	22,4	100,0
	Total	116	100,0	100,0	

q26

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	40	34,5	34,5	34,5
	Correct	76	65,5	65,5	100,0
	Total	116	100,0	100,0	

q28

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	70	60,3	60,3	60,3
	Correct	46	39,7	39,7	100,0
	Total	116	100,0	100,0	

Descriptive statistics for Primary School Education Students

Year of studies

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1st	14	9,0	9,0	9,0
	2nd	41	26,5	26,5	35,5
	3rd	59	38,1	38,1	73,5
	4th	32	20,6	20,6	94,2
	5th +	9	5,8	5,8	100,0
	Total	155	100,0	100,0	

q1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	55	35,5	35,5	35,5
	Correct	100	64,5	64,5	100,0
	Total	155	100,0	100,0	

q2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	60	38,7	38,7	38,7
	Correct	95	61,3	61,3	100,0
	Total	155	100,0	100,0	

q3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	16	10,3	10,3	10,3
	Correct	139	89,7	89,7	100,0
	Total	155	100,0	100,0	

q4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	26	16,8	16,8	16,8
	Correct	129	83,2	83,2	100,0
	Total	155	100,0	100,0	

q7

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	17	11,0	11,0	11,0
	Correct	138	89,0	89,0	100,0
	Total	155	100,0	100,0	

q8

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	35	22,6	22,6	22,6
	Correct	120	77,4	77,4	100,0
	Total	155	100,0	100,0	

q9

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	24	15,5	15,5	15,5
	Correct	131	84,5	84,5	100,0
	Total	155	100,0	100,0	

q11

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	28	18,1	18,1	18,1
	Correct	127	81,9	81,9	100,0
	Total	155	100,0	100,0	

q12

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	43	27,7	27,7	27,7
	Correct	112	72,3	72,3	100,0
	Total	155	100,0	100,0	

q16

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	78	50,3	50,3	50,3
	Correct	77	49,7	49,7	100,0
	Total	155	100,0	100,0	

q17

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	48	31,0	31,0	31,0
	Correct	107	69,0	69,0	100,0
	Total	155	100,0	100,0	

q18

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	27	17,4	17,4	17,4
	Correct	128	82,6	82,6	100,0
	Total	155	100,0	100,0	

q21

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	28	18,1	18,1	18,1
	Correct	127	81,9	81,9	100,0
	Total	155	100,0	100,0	

q24

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	69	44,5	44,5	44,5
	Correct	86	55,5	55,5	100,0
	Total	155	100,0	100,0	

q25

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	124	80,0	80,0	80,0
	Correct	31	20,0	20,0	100,0
	Total	155	100,0	100,0	

q26

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	67	43,2	43,2	43,2
	Correct	88	56,8	56,8	100,0
	Total	155	100,0	100,0	

q28

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	86	55,5	55,5	55,5
	Correct	69	44,5	44,5	100,0
	Total	155	100,0	100,0	

Descriptive statistics for Teachers

School Education Level

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Primary	73	76,0	76,0	76,0
	Secondary	23	24,0	24,0	100,0
	Total	96	100,0	100,0	

q1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	22	22,9	22,9	22,9
	Correct	74	77,1	77,1	100,0
	Total	96	100,0	100,0	

q2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	27	28,1	28,1	28,1
	Correct	69	71,9	71,9	100,0
	Total	96	100,0	100,0	

q4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	9	9,4	9,4	9,4
	Correct	87	90,6	90,6	100,0
	Total	96	100,0	100,0	

q5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	24	25,0	25,0	25,0
	Correct	72	75,0	75,0	100,0
	Total	96	100,0	100,0	

q7

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	9	9,4	9,4	9,4
	Correct	87	90,6	90,6	100,0
	Total	96	100,0	100,0	

q8

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	35	36,5	36,5	36,5
	Correct	61	63,5	63,5	100,0
	Total	96	100,0	100,0	

q9

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	25	26,0	26,0	26,0
	Correct	71	74,0	74,0	100,0
	Total	96	100,0	100,0	

q11

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	19	19,8	19,8	19,8
	Correct	77	80,2	80,2	100,0
	Total	96	100,0	100,0	

q12

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	7	7,3	7,3	7,3
	Correct	89	92,7	92,7	100,0
	Total	96	100,0	100,0	

q15

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	11	11,5	11,5	11,5
	Correct	85	88,5	88,5	100,0
	Total	96	100,0	100,0	

q16

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	34	35,4	35,4	35,4
	Correct	62	64,6	64,6	100,0
	Total	96	100,0	100,0	

q17

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	14	14,6	14,6	14,6
	Correct	82	85,4	85,4	100,0
	Total	96	100,0	100,0	

q18

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	14	14,6	14,6	14,6
	Correct	82	85,4	85,4	100,0
	Total	96	100,0	100,0	

q19

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	66	68,8	68,8	68,8
	Correct	30	31,3	31,3	100,0
	Total	96	100,0	100,0	

q20

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	38	39,6	39,6	39,6
	Correct	58	60,4	60,4	100,0
	Total	96	100,0	100,0	

q23

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	24	25,0	25,0	25,0
	Correct	72	75,0	75,0	100,0
	Total	96	100,0	100,0	

q24

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	22	22,9	22,9	22,9
	Correct	74	77,1	77,1	100,0
	Total	96	100,0	100,0	

q25

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	55	57,3	57,3	57,3
	Correct	41	42,7	42,7	100,0
	Total	96	100,0	100,0	

q26

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	30	31,3	31,3	31,3
	Correct	66	68,8	68,8	100,0
	Total	96	100,0	100,0	

q28

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Wrong	34	35,4	35,4	35,4
	Correct	62	64,6	64,6	100,0
	Total	96	100,0	100,0	