Amanda da Costa Marsura

KNOWLEDGE STRUCTURE EVALUATION OF THE ENEM:

DO DIFFERENT GROUPS LEARN DIFFERENTLY?

Support:

CAMPINAS

2023

Amanda da Costa Marsura

KNOWLEDGE STRUCTURE EVALUATION OF THE ENEM: DO DIFFERENT GROUPS LEARN DIFFERENTLY?

Dissertation presented to the *Stricto Sensu* Graduate Program Psychology at San Francisco University, Area of Concentration – Psychological Assessment for obtaining a Master's degree.

Advisor: Prof. PhD. Víthor Rosa Franco

CAMPINAS

2023

157.93 Marsura, Amanda Costa.
M329k Knowledge Structure Evaluation of the ENEM: do different groups learn differently? / Amanda Costa Marsura. – Campinas, 2023.
60 p.

Dissertação (Mestrado) – Programa de Pós-Graduação *Stricto Sensu* em Psicologia da Universidade São Francisco.
Orientação de: Víthor Rosa Franco.

1. Educational Assessment. 2. Psychometrics. 3. Knowledge space theory. 4. Equity. I. Franco, Víthor Rosa. II. Título.

Sistema de Bibliotecas da Universidade São Francisco - USF
Ficha catalográfica elaborada por: Tatiana Santana Matias - CRB-08/8303

UNIVERSIDADE SÃO FRANCISCO - PROGRAMA DE PÓS-GRADUAÇÃO STRICTO SENSU EM PSICOLOGIA

Amanda da Costa Marsura defendeu a dissertação **"*KNOWLEDGE STRUCTURE EVALUATION OF THE ENEM: DIFFERENT GROUPS LEARN DIFFERENTLY*"** **aprovada** pelo Programa de Pós-Graduação Stricto Sensu em Psicologia da Universidade São Francisco em 24 de fevereiro de 2023 pela Banca Examinadora constituída por:

Prof. Dr. Vithor Rosa Franco
Orientador e Presidente

Prof. Dr. Jacob Arie Laros
Examinador

Prof. Dr. Lucas de Francisco Carvalho
Examinador

Abstract

Marsura, A. C. (2023). *Knowledge structure evaluation of the Enem: Do different groups learn differently?* Master's Thesis, Post-Graduate Studies in Psychology, University San Francisco, Campinas, São Paulo.

Psychometrics deals with the development of techniques that aim at measuring socio-psychological constructs. Several psychometric theories are being used in the field of educational assessment. This study addressed Item Response Theory (IRT) and Knowledge Space Theory (KST). While the first assumes learning as a linear process, the second considers learning as a nonlinear process and presumes the existence of "learning paths". In the context of educational assessment in Brazil, various large-scale assessments are undertaken. In this study data has been used of the national high school exam (ENEM), being the main entry point for Brazilians into higher education. In this sense, the objective of the present study is to compare knowledge structures of students from public and private institutions, on basis of the items of the ENEM exam of 2019. This way, evidence was obtained indicating that respondents from different groups learn the same knowledge differently. The implications arising from the use of KST are being discussed.

Keywords: Educational Assessment, Psychometrics, Knowledge space theory, Equity.

Resumo

Marsura, A. C. (2023). *Avaliação das estruturas de conhecimento do Enem: Grupos diferentes aprendem diferentemente?* Dissertação de Mestrado, Programa de Pós-graduação em Psicologia, Universidade São Francisco, Campinas, São Paulo.

A psicometria trata do desenvolvimento de técnicas que visam medir construtos sociopsicológicos. Várias teorias psicométricas estão sendo utilizadas no campo da avaliação educacional. Este estudo abordou a Teoria de Resposta ao Item (TRI) e a Teoria do Espaço de Conhecimento (KST). Enquanto a primeira assume a aprendizagem como um processo linear, a segunda considera a aprendizagem como um processo não linear e pressupõe a existência de "percursos de aprendizagem". No contexto da avaliação educacional no Brasil, diversas avaliações de larga escala são realizadas. Neste estudo foram utilizados dados do Exame Nacional do Ensino Médio (ENEM), sendo a principal porta de entrada dos brasileiros no ensino superior. Nesse sentido, o objetivo do presente estudo é comparar as estruturas de conhecimento de alunos de instituições públicas e privadas, com base nos itens do ENEM de 2019. Dessa forma, foram obtidas evidências que indicam que respondentes de grupos diferentes aprendem o mesmo conhecimento de forma diferente. As implicações decorrentes do uso do KST estão sendo discutidas.

Palavras-chave: Avaliação Educacional, Psicometria, Teoria do espaço do conhecimento,

Equidade.

Summary

List of Figures 7
List of Tables 8
Knowledge Structure Evaluation of the ENEM: 9
Do Different Groups Learn Differently? 9
Large-Scale Educational Assessment 10
The National High School Exam (ENEM) 13
Psychometric Foundations for Large-Scale Assessment 16
Psychometric Theories and their Implications on the Learning Process 20
Method 27
Participants 27
Instruments 27
Templates and Exams 28
“Read me” and Technical Documents 28
Dictionary for Understanding Microdata 29
Inputs 29
Data 29
Procedures 30
Data Analysis 30
Results 33
Discussion 42
Final Remarks 45
References 48

List of Figures

Figure 1		18
Figure 2		22

List of Tables

Table 1 – Fit performance for the knowledge spaces estimated with the KST and the 3-PLM		35
Table 2 - Model comparison of the BLIM fit		36
Table 3 – Model comparison of the equivalency of the scores generated by the KST model and the 3-PLM		37
Table 4 – Fit performance for the knowledge spaces estimated with the KST for private and state schools		39
Table 5 – Model comparison of the BLIM fit		40
Table 6 – Model Comparison of the Equivalency of the Scores Generated by the KST Model with the Private Schools Dataset and with the State Schools Dataset		41

Knowledge Structure Evaluation of the ENEM:

Do Different Groups Learn Differently?

Learning is a process that is influenced by the context in which it takes place (Osher et al., 2020) and it is usually evaluated using educational assessment tools. An educational assessment tool is any type of exam used to assess the knowledge and skills acquired by respondents over a given period of time. In Brazil, the National High School Exam (*Exame Nacional do Ensino Médio,* ENEM; Andrade, 2012) is one of the most important educational assessment tools to compare students coming from different socioeconomic contexts (e.g., Kleinke, 2017; Lima Junior, 2015; Lucena & Santos, 2020; Nascimento et al, 2018). ENEM is a Brazilian large-scale educational assessment tool used for admission to most public universities in Brazil and for obtaining scholarships or funding to attend private universities (INEP, 2019a). The scores of ENEM are calculated using Item Response Theory (IRT; Valle, 2000) methods, which in case the assumption of unidimensionality is met (Condé & Laros, 2007), turns the scores independent of the sample of individuals and items.

From a theoretical perspective, IRT assumes that the learning process is linear. This means that every new concept, skill, or knowledge is acquired one after the other, in an ordered fashion, depending on its difficulties. However, prior studies have shown that the linearity of learning may be questionable (e.g., Doble et al., 2019; Segedinac, 2018). In this scenario, Knowledge Space Theory (KST; Doignon & Falmagne, 2012) can be used as an alternative to IRT for the identification of individuals with more or less mastering of the evaluated skills. Differently from IRT, KST assumes non-linearity of learning, highlighting the possibility that each respondent learns the same knowledge in a different way. Therefore, the objective of this study is to compare the knowledge structures of ENEM's respondents from public and private schools. This study may allow us to understand if the implications of

IRT and KST culminate in different results for the same respondent (e.g., being accepted or not at a university).

Large-Scale Educational Assessment

A large-scale educational assessment is characterized as an educational assessment that affects a large number of people (Emler et. al, 2019). In a global context, there are several types of large-scale educational assessment that act as admission instruments to help universities choose their students. These include not only high school exit exams, but also certification exams or entrance exams and standardized tests of aptitudes. In some cases, students can be admitted without any exam, but still selected through some type of demonstration of requirements obtained throughout their formal basic education (Soares & Soares, 2020).

The Scholastic Aptitude Test (SAT) is a good example of an effective large-scale educational assessment. Administered by the College Entrance Examination Board of the United States of America since 1926, it is a test that focuses on reasoning skill (Liu et al, 2007). According to the College Board, the SAT measures the skills and knowledge that research shows to be the most important for success in college and career. The SAT includes the following sections: Reading, Writing and Language, and Mathematics. In addition, it has optional essay (or essay) sections, which measure reading, analysis and writing skills.

Another example of an effective large-scale assessment is the General Test (GRE), which is administered by the Educational Testing Service (ETS; Klieger et al., 2018) and is a test normally required for admission to US graduate programs or undergraduate courses. The GRE has two versions: printed and computerized. The GRE measures verbal reasoning, quantitative reasoning, critical thinking, and writing skills. It is organized into three batteries: Verbal Reasoning; Quantitative Reasoning; and Analytical Writing. In general, the GRE is used to obtain information that complements undergraduate information, since its

respondents are candidates for postgraduate programs or so-called 2nd cycle graduates (Soares & Soares, 2020).

In general, large-scale educational assessments have four characteristics. First is its uniformity, which is necessary so that the assessment can be applied to different audiences with preserved validity and reliability. Secondly, the high absolute cost, which concerns the expenses to design, develop, administer, score and report the results of the exam. However, it should be noted that the relative cost can be considered cheap, as it is less expensive than, for instance, interviewing all respondents. The last two characteristics are the broad impact and high stakes. Respectively, these characteristics refer to the high number of respondents who perform it, as well as the great influence that these assessments exert on the life of the population (Emler et al, 2019).

The emergence of large-scale educational assessments occurred due to the need for diagnosing and planning directions for policies aimed at decision-making on education (Vianna, 2003). In this context, the concept of “quality” is fundamental. When it comes to education, the concept of quality is not a given; it depends on the individuals involved and their respective objectives (Horta Neto, 2010). This means that the quality of teaching would not be associated with the student's personal assessment of it, but with the achievement of its objective (e.g., entering the desired higher course through the performance obtained in a given assessment). Related to the concept of quality in the context of an assessment, there are also the concepts of efficiency, efficacy and effectiveness (3Es). The first concept, efficiency, is cost-benefit. It means the possibility of producing more with fewer resources. The second is efficacy, which has to do with achievement of goals. Finally, effectiveness refers to how well something works under real conditions (Garcia et al., 2016; Marley, 2000; Matias et al., 2019; Vinha et al., 2016). In that same context, there is a need to establish universal quality criteria (i.e., a reference standard for comparing the results). Despite criticisms about large-

scale assessment forcing a quality standard for the system that is not adequate for any single individual, it is important to consider operational costs (Emler et al., 2019) and the fact that the competencies for attending and performing well at Higher Education are somewhat standardized (Cunha & Muller, 2018; Rodríguez-Hernández et al., 2020).

In Brazil, large-scale educational assessments have been implemented since the 1990's with the purpose of improving the acquisition of skills and competences, as well as to increase students' performance (Vianna, 2003). There are two most important large-scale educational assessments created and implemented in Brazil. First is the Basic Education Assessment System (SAEB; *Sistema de Avaliação da Educação Básica*; Heck, 2018), which aims to define priorities and improve the quality of teaching. SAEB is subdivided into the National Assessment of School Achievement (ANRESC; *Avaliação Nacional do Rendimento Escolar;* Silva & Carvalho, 2020), the National Education Assessment (ANEB; *Avaliação Nacional da Educação;* Villani & Oliveira, 2018), and the National Literacy Assessment (ANA; *Avaliação Nacional de Alfabetização;* Dickel, 2016).

The second most important large-scale educational assessment implemented in Brazil is the ENEM and the National Higher Education Assessment System (SINAES; *Sistema Nacional de Avaliação da Educação Superior*, Teixeira Junior & Rios, 2017), which aim to draw an overview of the quality of courses and higher education institutions in the country. Such assessments have in common the aim of gathering subsidies for the formulation of new educational policies, aiming at deepening the knowledge about different education systems, allowing the government to define intervention priorities, in addition to inducing changes or consolidating previously structured educational reforms (Minhoto, 2016).

In this context, the question remains whether the system is oriented and sufficiently organized for the development of skills and competences. More specifically, questions remain regarding the alignment between the objectives of the large-scale educational

assessments and the school's objectives (Vianna, 2003). Many schools in Brazil do not have the capacity of managing or adequately preparing students for large-scale educational assessments. On the other hand, the results of large-scale assessments culminate in the identification of educational problems, but not what the corrective pedagogical strategies may be (Sousa, 2019). Social actors also question the existence of equity in large-scale assessments, as well as the adequacy of the procedures of grading and their applicability to all education systems in Brazil (Almeida, 2020). Although there are constant attempts to make the Brazilian educational system more uniform, the diversity of educational realities, due to social and economic factors, is evident.

In order to guarantee a standardized quality of such assessments, the use of theories and foundations of psychometrics is pertinent. Both in the Brazilian context and in the international context, large-scale assessments have relied mainly on the methods of two theories, which integrate, respectively, classical and contemporary psychometrics: Classical Test Theory (CTT); and Item Response Theory (IRT; Soares & Soares, 2020). Both comprise theories for the measurement of socio-psychological constructs and their methods are used with the intention of providing fair and standardized results. However, while CTT focuses on sum scores and test's reliability, IRT focuses on items' properties. Currently, both theories are used in the context of large-scale educational assessments (Souza, 2019). However, the present study will give greater emphasis to IRT than to CTT, because information is derived from the ENEM using the IRT.

The National High School Exam (ENEM)

The promulgation of the Brazilian Federal Constitution took place in 1988 (CF/88). With this, education was introduced as the first of social rights, a right for all and a duty of the State. Furthermore, the constitutional text was detailed in terms of the right to education

with principles, the division of responsibility among federal entities, as well as forms of financing. There were also advances in guaranteeing the right to education with the implementation of policies aimed at mainly expanding access to the stages of basic education (Gonçalves & Silveira, 2021).

Among the principles postulated in the letter of the law, which should guide the educational process, are: (a) equal conditions for access and permanence in school; (b) freedom to learn, teach, research and spread thought, art and knowledge; (c) pluralism of ideas and pedagogical conceptions, and coexistence of public and private educational institutions; (d) free public education in official establishments; (e) appreciation of school education professionals, guaranteed, in the form from the law, career plans, with admission exclusively through public examinations and titles, to those of public networks; (f) democratic management of public education, in accordance with the law; (g) quality standard assurance and (h) national professional salary floor for school education professionals under federal law (Constituição do Brasil, 1988).

Although there have been many improvements in the context of public basic education, the private sector is several steps ahead (Sampaio & Guimarães, 2009). Due to its larger budget, the private educational sector is generally able to offer better study conditions to students, both in terms of structural quality and in terms of teaching quality. Although there are exceptions (e.g., Folha, 2018), this disparity can directly influence the enrollment of students in higher education. In this context of Brazilian basic education, there are educational assessments. In this study, attention will be paid to the ENEM.

ENEM is a large-scale educational assessment which is currently used in Brazil as an entrance exam in public and private Higher Education Institutions (HEIs; Andrade, 2012). It was created in 1998 by the Brazilian federal government to be a tool for evaluating the performance of students in completing basic education. The test was intended to be an aid to

the Ministry of Education (*Ministério da Educação, MEC*) in the construction of specific and structural policies for the improvement of Brazilian education through the National Curriculum Parameters (*Parâmetros Curriculares Nacionais, PCNs*) of high schools and elementary schools. For over ten years, performance on this test was used solely to assess the skills and abilities of high school graduates, far from the current purpose of selecting for higher education (Silveira et al, 2015). Then, from 2009 onwards, the new ENEM was implemented and then, changes were consolidated not only in terms of structure, but also in objectives and correction methodology (Andrade, 2012).

Since 2009, the ENEM test has increased from 63 to 180 questions. With this increase, the application time has also augmented: the test started to be administered during two days instead of one, as is the case until today. On the first day, Human Sciences and Their Technologies, Languages, Codes and Their Technologies and Writing are administered. On the second day, it is time for Mathematics, Codes and Their Technologies, and Natural Sciences and Their Technologies. In each of the four areas 45 questions are applied. In addition, the test that until 2008 was corrected using Classical Test Theory (CTT), started to be corrected using Item Response Theory (IRT). In short, this change enabled the production of more refined scores, as well as the opportunity to compare participants from different contexts who perform different tests (Sousa, 2019).

Nowadays, the primary objective of ENEM is to obtain empirical evidence whether respondents demonstrate mastery of the scientific and technological principles that integrate modern production and whether they are aware of contemporary forms of language (INEP, 2019a). In the same sense, it is pointed out that the ENEM results can allow the construction of parameters for the participant to self-evaluate, as well as the constitution of a national reference in order to improve the high school curriculum. Furthermore, the ENEM results can also be used as a complementary, unique or alternative way of accessing high school. There

is also the possibility for the participant to gain access to funding or even a type of support for higher education students. Finally, the results can enable the development of indicators and studies about Brazilian education (INEP, 2019b).

ENEM is structured in four areas of knowledge. The first area is Human Sciences and its Technologies, which has four curricular components: History, Geography, Philosophy and Sociology. The second area of knowledge of the ENEM is the Natural Sciences and its Technologies, which has three curricular components: Chemistry, Physics and Biology. Third is the Language, Codes and their technologies, whose curricular components are: Portuguese Language, Literature, Foreign Language (English or Spanish), Arts, Physical Education and Information and Communication Technologies. Finally, the fourth area of knowledge of ENEM is Mathematics and its technologies, whose curricular component is Mathematics (INEP, 2019a).

In order to take the test, respondents of ENEM have the possibility to request specific assistance (if they have any special educational needs, or even limited mobility, vision and hearing), in addition to being allowed to choose to use their social name, all factors that support social inclusion (INEP, 2019b). Initially, ENEM gave the student the possibility of entering public institutions of higher education. Currently, it also allows admission to private higher education institutions and receiving scholarships. Thus, the target audience of ENEM exists of students who graduate from basic education, both in the public education system and in the private system.

Psychometric Foundations for Large-Scale Assessment

In the context of large-scale educational assessments, studies that address differences in performance between individuals from different socioeconomic backgrounds have been developed. Lucena & Santos (2020) conducted a study investigating the relationship between

performance at ENEM and the socioeconomic profile of the respondents through the database of the 2016 edition of the exam. The results indicate that some specific factors of the respondents' socioeconomic profile, such as educational institution of origin (public/ private), color, family income and whether the respondent works are related to their performance on the exam. It was found that white respondents, who do not work, who attended high school primarily in a private school, and with higher family income tend to obtain better performance on the exam. Corroborating these results, we identified, for example, two other studies that found the influence of socioeconomic issues on the performance of ENEM respondents (Carmo et al., 2019; Marcom & Kleinke, 2017). In this same direction, we identified some studies that specifically focus on the performance of ENEM respondents on the English subtest. A significant part of these studies are of a qualitative character and emphasize only the process of teaching and learning the foreign language in basic education, after its insertion in ENEM (Pinheiro & Quevedo-Camargo, 2017).

However, studies were also identified that assess the social impact of the insertion of English items in ENEM (Blanco, 2013) and even the reasons that lead many respondents to choose the Spanish subtest, despite having studied English during basic education (Mendes & Nunes, 2019). Recently, an analysis carried out by Folha (2021) pointed out English as an obstacle for students from public schools. Although the English items represent only 3% of the total exam, these items represent 46% of the items that have most affected students from public schools in comparison to students from private schools. These empirical results must be evaluated taking into account the limitations of traditional psychometrics.

Psychometrics is the field of study concerned with the development of techniques and methods for measuring sociopsychological constructs (Pasquali, 2017). Conventionally, sociopsychological constructs are assumed to be continuous latent variables (Hutz et al., 2015). This means that, to latent variables, a metric measure is imposed without being

directly observed. Even for special cases where latent variables are considered to be categorical, like in Latent Class Analysis (Porcu & Giambona, 2017), the latent variables are assumed to originate from the same linear underlying space. This means that for latent classes, respondents can be categorized in, e.g., “high” and “low” regarding some psychological construct.

In the context of education, the default procedure in psychometrics is reflected by the scoring system that is used in most parts of the world (Sartes & Souza-Formigoni, 2013): items in a test receive scores which can be summed or averaged to achieve what is called a “test score” or “sum score”. There are more refined ways of estimating these scores, such as by means of methods derived from Item Response Theory (IRT), but these methods have the same basis: someone with a higher score always has more knowledge or magnitude of the measured latent variable.

For Doignon and Falmagne (2015), the Knowledge Space Theory (KST), on the other hand, distances itself from the most common numerical evaluation approaches derived in traditional psychometrics. KST is based on a combinatorial approach regarding the evaluation of knowledge, depending on only two assumptions. The first assumption, known as Learning Smoothness, implies that a student with less knowledge can catch up with a student who has more knowledge by learning the missing concepts, one at a time. The second assumption, known as Learning Consistency, says that any new concept that a student with less knowledge is ready to learn either was already mastered by a student with more knowledge, or the latter is also ready to learn it.

The aforementioned assumptions allow clarifying the main differences between conventional psychometrics, particularly expressed in terms of IRT, and KST. The result of an evaluation that uses IRT as a methodology aims to reveal the measure of a respondent's particular latent trait (Andrade & Valle, 1998). IRT also assumes that knowledge is acquired

in the same way by all test respondents, which culminates, for example, in the attribution of lower scores for those who miss items of lesser difficulty and higher scores for those who correctly answer items of greater difficulty. On the other hand, KST works with the idea that there are a number of paths that a student can take to acquire certain knowledge (Falmagne et al., 2013). In spite of the "difficulty" of items, each student may master the specific knowledge by a different "learning path". Therefore, KST offers an opportunity to consider the respondents' particularities in an evaluation. An assessment, according to KST, should reveal the exact set of dominated items, defined as the knowledge state.

Despite the many advantages of applying IRT in the context of large-scale assessments, some limitations remain. A number of studies on invariance properties (Andrade et al., 2010) and differential item functioning (DIF) of ENEM items (Vieira, 2020) support the idea that, despite the many advantages of using IRT methods in scoring large scale assessments, analysis of item invariance and DIF must be realized. Both are ways of verifying that large-scale assessments actually rate fairly.

From a technical perspective, ENEM is evaluated by applying IRT methods (INEP, 2019a). From a theoretical perspective, the point is that, depending on the socioeconomic background, or even on personal and other contextual factors (Figueirêdo et al., 2014), learning may be different for each one. Although one knowledge after another is mastered, the linearity of learning is discussed and, thus, the possibility of learning to occur in a non-linear way (Doignon & Falmagne, 2015; Falmagne, 2013). Traditional psychometrics in general, and IRT specifically, assumes a linear learning space, scores, even those corrected by the guessing parameter, will not consider that people may learn differently (i.e., non-linearly). Traditional psychometrics in general, and IRT specifically, assumes a linear learning space and in doing so is not considering that people may learn differently.

On the other hand, KST allows one to evaluate knowledge when learning happens in an order that is not necessarily linear. Further still, in KST, the order in which learning takes place may or may not be different for each person. This means that with KST it is possible for a student to take a different learning path to obtain certain knowledge. This is not possible with IRT. Therefore, using KST instead of IRT in the context of ENEM would allow one to take into account learning peculiarities when assessing the respondents, moving towards a more equalitarian and fair assessment (Falmagne et al., 2006).

Applying IRT to generate scores for large-scale educational assessments is assuming that all the respondents learn exactly in the same way, disregarding the diverse contexts students come from. As KST assumes the existence of "learning paths", which can be different for each student in a common knowledge space, KST allows considering the diversity of social contexts and its influence on the learning process. In the context of a large-scale assessment, the differences between IRT's and KST's implications could directly impact the results and the social consequences derived from them.

Psychometric Theories and their Implications on the Learning Process

IRT is a set of methods used to represent the relations between the likelihood that a respondent will correctly answer a given item and ~~what is~~ the magnitude of the respondents' latent traits or, in the context of educational assessments, knowledge (Andrade et al., 2000). In this context, knowledge is assumed to be a continuous variable that increases in a linear fashion. This means that, independent of the items' content, increasing your knowledge in a given domain should increase the probability of answering any item of that domain correctly. This is readily verified by evaluating a nonparametric formula that represents the family of the logistic item response theory models (Valle, 2000):

$$P(X_i = 1) = f(\theta_j - b_i) \quad (1)$$

where $P(X_i = 1)$ is the probability of answering item i correctly, θ_j is the knowledge of respondent j, b_i is the difficulty of item i, and f is usually an increasing logistic function.

IRT emerged as an alternative to Classical Test Theory, which is the most used theory in the context of assessment (Borsboom, 2006). However, both theories depart from the same Latent Variable Theory (McDonald, 2013). Therefore, all approaches that can be called "traditional psychometrics" (e,g, sum scores, facet theory, factor analysis, latent profile/class analysis, multidimensional scaling, and others) follow the same underlying linear and additive structure.

In general, psychometric methods, with IRT specific's case represented in Equation 1, formalize an additive relation between the essential elements (i.e., the respondents' latent trait and the items' parameters) of the situation in which a person responds to an item (Primi, 2004). It is possible to see, then, that the difficulty of an answer is represented by the value of b_i, which posits a linear dependence between items (Le, 2013). This means that as θ_j increases, the probability of getting any item right also increases, and items are mastered (i.e., $P(X_i = 1)$ tends to 1) depending on the order of b_i for all i.

To estimate θ_j and b_i, a number of IRT models can be used. In the context of educational assessment, the three-parameter logistic model (3-PLM; Maris & Bechger, 2009) is commonly used as it includes a "guessing" parameter and can penalize the scores of respondents for guessing. As a consequence, the 3-PLM values respondents who answer coherently (i.e., answers correctly only items that are not more difficult than their values of θ_j). This is a direct consequence of IRT's assumption that knowledge is learned in a linear fashion, as represented in Figure 1. According to this assumption, respondents that correctly answer more difficult items should also correctly answer easier items, as they have "more knowledge" (Andrade et al., 2000).

Figure 1

Acquisition of Knowledge from a Linear Fashion (from Item Response Theory)

Note. The arrows correspond to the only "learning path" possible according to the IRT, which assumes that learning occurs in a linear way for everyone. Each letter represents a new learned concept, "a" being the easiest concept, and "h" the most difficult concept.

This assumption, combined with the 3-PLM, allows respondents with a higher number of correct answers to receive lower scores than respondents with a lower number of correct answers, as long as their response patterns are not "coherent". This means that, for traditional IRT, the mechanism that explains this type of "incoherent" response patterns is the probability of "guessing" of respondents with lower knowledge.

In KST, on the other hand, an assessment should reveal the respondent's knowledge state, defined as the exact set of concepts dominated by the respondent (Falmagne et al., 2013). The concepts have a hierarchical order of mastering, but this order does not need to be linear. This means that, while in IRT a respondent with a lower score should have a low probability of correctly answering a "difficult" item, in KST it is possible for a respondent with more knowledge to not be able to correctly answer an "easy" item. Therefore, in KST, depending on the respondents' knowledge state and the populations' underlying learning space, it is plausible that, without guessing, a respondent that does not master more difficult items in general, may get a particularly more difficult item correctly.

More concretely, in KST, the learning space is constituted by the collection of all states of knowledge that, by convention, will always have at least two states. The first is the empty state (represented by Ø), the state that corresponds to knowing nothing about the domain. The second is the state containing the entire domain (represented by Q), corresponding to the state where one knows everything about the domain. Besides, between

the empty state and the item set domain, there could be an infinite number of states, depending on the number of existing states. Immediately after the empty state, there are the first states composed of the items that are the initial concepts of that domain and from which an infinity of knowledge states can be derived. Several learning paths are possible to advance to the next states. However, some items will be requirements for advancement from one state to another.

It should be clear at this point that a knowledge state is not a quantitative estimate of how much a student has learned, but rather a description of what a student knows and does not know at any given time (Doignon & Falmagne, 2015). This, of course, contrasts with IRT's estimates of θ_j. Due to this characteristic of KST, it can be complemented by the Learning Space Theory (LST; Falmagne et al., 2013). According to LST, a knowledge state shows exactly what the respondent is ready to learn. Each knowledge state (with the exception of Q) has an "external fringe", which is the set of concepts (skills, methods, facts) that the student in that state can start learning. Also, each knowledge state (with the exception of Ø) has an "internal fringe", which is the set of concepts (skills, methods, facts) that the student has just mastered. In the formal analysis of the knowledge space, the letter K is used to represent one of the states that are feasible for that particular knowledge space. In practical terms, a respondent with a knowledge state K can, in principle, correctly answer all items in K and fail to correctly answer any item that is not within K (Cosyn et al., 2021).

For a better understanding, think of a hypothetical and abstract situation regarding items of basic English grammar. In this case, one could expect at least 10 items to compose the knowledge space. The 10 items could involve: (a) the use of the verb to be; (b) verb tenses and the correct use of the pronouns; (c) building sentences in the Present perfect; (d) building sentences in the Simple past; (e) building sentences in the Past perfect; (f) the use of the Modal verbs; (g) third conditional; (h) using direct and indirect speech; (i) passive voice;

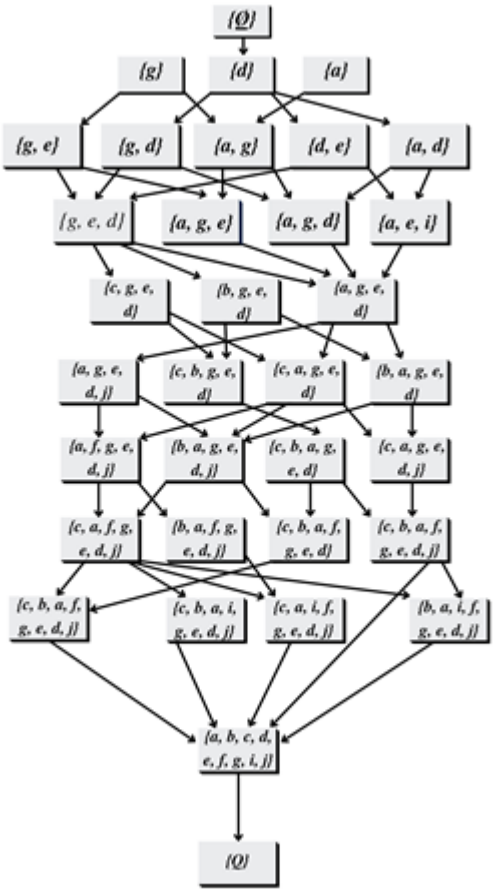
and (j) irregular verbs. It is worth noting that the term “item” in KST contains types of problems or abstract representations of specific concepts. This contrasts with the IRT definition of “item”, which corresponds to particular cases of a specific subject. In KST, particular cases are named as “instances”.

Given these 10 items, a knowledge and a learning space similar to that represented in Figure 2, could be derived. At the top of Figure 2, there is the empty state, Ø, which represents the individuals that know nothing about basic English grammar. At the bottom of Figure 2, there is the full/maximum state, *Q*, which presents the individuals that know everything there is to know about basic English grammar (defined by the 10 items). The arrows correspond to the possible learning paths for the given knowledge state (represented by the square nodes). It is also possible to see that *g*, *d*, and *a* are the easiest items, as they are the only items that can be learned directly from the empty state. It is also clear that *h* is the most difficult item, as it only can be learned after all other items are mastered.

Another issue would be that when we consider that no individual can master the item *e* from the initial state, it becomes clear that item *e* is more difficult than items *a* and *d.* However, after mastering any of items *g* or *d*, the individual is able to learn item *e* (i.e., it is possible to learn it from the same previous state), but item *a* does not provide what is needed to learn item *e*. In the case of IRT, if items *g*, *d* and *a* had the same difficulty, mastering any of these items would already allow the possibility of mastering item *e*.

Figure 2

Learning Space Corresponding to the Example of the 10 Items on Basic English (Grammar)



Note. The arrows correspond to the possible learning paths for the given knowledge state (represented by the square nodes). It is also possible to see that *g*, *d*, and *a* are the easiest items, and that *h* is the most difficult item.

Despite these different theoretical aspects that lead to different applied consequences, KST is seldomly used or tested in real case scenarios. No specific discussion or evidence to justify this scenario was found in the literature. However, at least three justifications can be hypothesized. First, the lack of quantitative e training in psychology courses (Borsboom, 2006; Townsend, 2008). KST has a more advanced mathematical foundation than IRT, and besides this there are not many statistical software programs implementing its methods. This makes KST more difficult to be used by researchers with less knowledge on statistics and statistical programming. The second justification is that even traditional psychometrics is not

well used by researchers in general (Flake & Fried, 2020). Even with a simpler theory, traditional psychometric methods abound with questionable measurement practices (QMP). QMPs are psychometric decisions made by researchers that decrease the robustness of the validity of measures used in a study.

Finally, due to the lack of quantitative training, researchers use statistical or mathematical tools without thoughtful considerations about theory (Sijtsma, 2012). Statistical methods are commonly defined simply as “tools" for researchers from other disciplines. However, the underlying mathematics can also be thought of as a language used to describe the nature of the studied process or phenomena. Therefore, when using a “standard” statistical method (such as IRT models), many implicit assumptions are made and seldomly tested (Michell, 2008). As a consequence, different measurement methods should be used and tested in order to allow stronger tests about the phenomena of interest.

In spite of its limited applications in real-world settings so far, KST has been used in organizing educational goals in the context of curriculum development (e.g., Milan et al, 2011). Furthermore, KST is also implemented in a widespread educational system called ALEKS, which is intended for online tutoring and assessment in subjects including math, introductory statistics, finance and chemistry (Harati et al., 2021). The ALEKS operation is based on Learning Space Theory (Falmagne et al., 2013). In ALEKS, the items potentially used in an assessment are, by construction, a completely comprehensive coverage of a curriculum, which is usually based on consultation of standard textbooks (Doignon & Falmagne, 2015).

Summing up, linearity of learning assumed by IRT and the influence of respondents' socioeconomic context on their performance are relevant factors to be taken into account when scoring or ordering respondents of ENEM. In this context, KST presents an alternative to try to work around these issues. Therefore, using the KST in the context of ENEM could

be a way to obtain fairer scores that actually take into account the fundamental idea of equity, since it is considered that learning can happen differently for different groups. It is important to emphasize that the present study does not intend to suggest that one psychometric theory is objectively better or inherently more coherent than the other. The main intention with this study is to compare the consequences that arise from the application of both theories (KST and IRT) in terms of both theoretical and practical implications.

Method

Participants

In the analysis of all subtests a stratified subsample of the 2019 edition of ENEM was used. Respondents were stratified according to the variables sex, socioeconomic status, and Federative Union. It is intended to use a sample of 5000 respondents. We opted for a sample of 5,000 respondents, due to the lack of knowledge of specific rules for choosing the sample size using KST. As an inclusion criterion, we only selected respondents who are high school graduates in the same year of the exam. This inclusion criterion is used in order to exclude those who did the ENEM only for training and also those who had already completed high school and might have been studying for longer, as is the case of respondents of pre-college courses or those aiming at getting a second degree. Another inclusion criterion that was used consisted of the family income declared in the socioeconomic questionnaire completed during the registration of the exam. Respondents with special educational needs were not included in the research sample.

Instruments

The analyzed data were available for download in the ENEM database (INEP, 2019). This database contained information on the tests dated from 1998 to 2019, gathering data that

includes everything from questions related to the test itself to answers given to the socioeconomic questionnaire. However, due to the extension of the datasets, as well as the change in ENEM's structure from 2009, we used only data from 2019.

On the INEP webpage, it is possible to obtain a very complete range of information regarding the ENEM test. Among the available data, information is distributed in the following folders: Templates and Exams, "Read me" and Technical Documents, Dictionary for understanding the microdata, the Inputs and the Microdata themselves, which are contained in the Data folder, along with the proof items. In the following topics, we will give a brief description of the content of the ENEM database, specifying each of the folders available (in every edition of the race) and the respective information present in each one of them.

Templates and Exams

The folder "Templates and Exams" contains all kinds of exams applied in a specific year, as well as the respective templates. The tests are identified by the colors: blue, pink, orange, grey, yellow, white and green, in addition to the specification about being for the first or second day of the test. The expanded versions (i.e., with a larger text font) of the tests are also found in this folder. In this sense, the complete exam has 180 questions, which are divided into four areas of knowledge (45 questions per area), covering the disciplines taught in high school. These questions are applied over two days, with 90 questions plus writing on the first day and another 90 questions on the second day.

"Read me" and Technical Documents

The folder called "Read me" and Technical Documents contains: the notice of exam, which provides for guidelines, procedures and deadlines of ENEM 2019; the "Read me", that is a file containing general provisions about the exam and the database; the ENEM writing

manual with tips, notes to assist the student in preparing for it; the reference matrix of the test that discusses the competencies and abilities from each area of knowledge of the exam.

Dictionary for Understanding Microdata

The Dictionary for understanding microdata is a folder composed of dictionary of variables, containing the name of the variable, its description and category, presenting how the information inherent to the participants, the school, requests for specialized care (e.g. for students with partial or total blindness, hearing loss and others), requests for specific care (e.g. for pregnant, lactating, elderly students or enrollment in a Hospital Unit), requests for specialized and specific resources to perform the tests (e.g. braille proof request, proof with enlarged font, reading aid, proof in a room with easy access, among others). In addition, the place of application of the tests, the objective test, the writing test and the socioeconomic questionnaire, as well as the items of the test themselves, are available in the microdata.

Inputs

These are used to read the files and were created using the software R, SAS and SPSS. The inputs allow the data to be loaded with the labels, which facilitates their handling, since it makes their use immediate and more intuitive. Nevertheless, it is still essential to consult the Exams and the Dictionary of variables for a good understanding of the information.

Data

Finally, we have the folder that contains the microdata themselves and the exam items. Microdata include all available information about the exam as a whole and the students who took it that year. On the basis of exam items, for each type of test, the information of

skill, color of the test, position of the item in the test, area of knowledge and template were loaded.

Procedures

In this research, the microdata of the 2019 edition of ENEM was managed and analyzed. Data management—download, extraction, importation, and initial treatment was done with the statistical software R (R Core Team, 2023). In the initial treatment, the exclusion of absent students, as well as students who took the test for training purposes and students who are not high school graduates in the year of the test was performed. In addition, the response vectors will be corrected according to the test's templates, according to the type of the exam. Then, stratification will be performed according to what was proposed in the Participants section. Finally, inductive Knowledge Structure recovering methods will be applied to the private and public school's subgroups to evaluate the study's aims.

Data Analysis

Knowledge Space Estimation

To learn the structure of the knowledge space with the ENEM data, we first intended on using the Inductive Item Tree Analysis (IITA; Ünlü & Schrepp, 2020) algorithm implemented in the DAKS R Package (Ünlü & Sargin, 2016). IITA is a set of methods of discrete optimization of the knowledge structure of questionnaires, which aims to obtain logical implications between dichotomous items (i.e., items where only two responses are possible). Contrasting with Latent Class Analysis, IITA does not try to identify latent categories, but rather a series of surmise relations of the type $i \rightarrow j$. These implications can be interpreted as if a subject can solve instances of item i, then such subject is also able to solve instances of item j.

However, pilot tests with the data showed the IITA algorithm to be very inefficient when the number of items in a test increases. This is due to the fact that IITA tries to identify both the size of the knowledge space as well as the possible knowledge states (i.e., possible response patterns). Because there are 2^k possible response patterns for a test with k items, the computational cost of model estimation grows quickly with the number of items. To reduce the computational cost of estimation, but still ensure a reasonable analysis, we developed an estimation method that uses a fixed size for the knowledge space before trying to estimate what are the possible knowledge states.

Because we have no reason to assume any particular size for the knowledge space of ENEM, we forced the algorithm to search for the most appropriate knowledge space that has a size no larger than the size derived from IRT models. The linear knowledge space derived from IRT models will always have a size equal to $k + 1$, assuming all k items have different difficulties. This means that the algorithm we will use cannot describe knowledge spaces more complex than the one derived using IRT models. This puts strong constraints on the potential KST models that can be estimated, but it also works as a lower bound estimate of the quality of these models. This is true because if a model with a constraint in the size of the knowledge space can perform better than an IRT model, then it is reasonable to assume that the true knowledge space, which is probably larger, can perform even better.

Therefore, to use our estimation method, first one has to define the size of the knowledge space that will be searched; in our case, $k + 1$. This knowledge space has two states that are fixed a priori: the empty and the full-knowledge spaces. This means that the estimation method has to estimate only $k - 1$ states. Because response patterns are composed only of 0s and 1s (i.e., incorrect and correct answers, respectively), and because there are no continuous parameters (such as the aptitude and item parameters from IRT), the KST model has to be estimated using discrete optimization (Parker & Rardin, 2014). More specifically,

we used a greedy binary algorithm (Rajeev & Krishnamoorthy, 1992), which is a deterministic optimization procedure that changes every parameter that improves the goodness-of-fit, and keeps fixed every parameter that decreases the goodness-of-fit. For measuring the goodness-of-fit, we used the *diff* coefficient, the same fit measure used by the minimized corrected IITA algorithm (Ünlü & Sargin, 2008), which is a maximum likelihood estimator based on the average sums of the quadratic differences between the observed and expected numbers of counterexamples (i.e., surmise relations that could also fit one intermediary estimate of the knowledge state).

Model Fit Comparison

To compare the goodness-of-fit of the knowledge space estimated with our method and the knowledge space derived from an estimated 3-PLM, we fitted the estimated spaces to the data using the basic local independence model (BLIM; Doignon & Falmagne, 1999). The BLIM is a probabilistic version of the knowledge space, which allows for lucky guesses and careless errors. We used the minimum discrepancy maximum likelihood (MDML; Heller & Wickelmaier; 2013) estimation method, which maximizes the likelihood while constraining the response errors to a minimum. The fit of the models was compared using the Akaike information criterion (AIC), the variation of AIC (ΔAIC), the likelihood ratio (LR), and the model weights (*w*; Lewandowsky & Farrell, 2010). The AIC is an interval measure of the error of the model, with smaller values representing a better comparative fit. The ΔAIC and LR provide the same information, i.e., the difference between the quality of the fit of the best model with the reference model. However, ΔAIC is represented in an interval scale while LR is in a ratio scale. Finally, *w* is the likelihood of the models, assuming that all the tested models represent the sample space of possible models. To fit the BLIM we used the pks package (Wickelmaier et al., 2022).

Model Performance Comparison

We used the Hamming distance (Norouzi et al., 2012) between the characteristic response pattern of the knowledge state and the response pattern of the respondents to estimate the knowledge state to each respondent. The Hamming distance is the number of cases where the knowledge state and the respondent response patterns are different. We use this approach as it is also used, for instance, in the probabilistic Guttman model (Proctor, 1970) and, therefore, make the estimated performances of the KST and the 3-PLM comparable.

To estimate the similarity of the scores estimated using the 3-PLM and the estimated knowledge spaces we will use generalized additive models (GAMs; Pya & Wood, 2015). GAMs are nonparametric regression models that can be used to estimate nonlinear relations. More specifically, for each subtest and subsample, we estimated a restricted and an unrestricted model. The restricted model will be used to estimate monotonically increasing relations. The unrestricted model will be used to estimate relations that can be nonmonotonic.

The model comparison reasoning is: if the monotonic model gives a better fit than the nonmonotonic model, then the scoring procedures of the KST and the 3-PLM are somewhat equivalent (i.e., they provide a similar ordering of the respondents). If the nonmonotonic model provides a better fit, then the KST and 3-PLM provide scores that are not equivalent. The fit of the models will also be compared using AIC, ΔAIC, LR, and *w*. To fit these models we used the mgcv (Wood, 2022) and the scam (Pya, 2022) R packages.

Results

Comparison of IRT and KST Performance Estimates

Table 1 shows that the goodness-of-fit indices are always smaller for the estimated KST models than for the estimated 3-PLMs, which indicate a better fit of the KST models than the 3-PLM models. It should also be noted that the error rates (ER) are, in general,

similar between the models, despite the existence of some inconsistencies. The ER represents the estimated probability of random errors for the given knowledge space. Ideally, ER should be similar between both models to indicate that they are comparable in terms of violations of their implied supposed relations. The two different versions of the *diff* coefficient shown in Table 1, *diff*1 and *diff*2, are based on the corrected and minimized corrected IITA, respectively. Because the estimation was conducted using *diff*1, *diff*2 was added to check if the results are stable, considering a second criterion.

Table 1

Fit performance for the knowledge spaces estimated with the KST and the 3-PLM

Dataset	Model	*diff1*	ER1	*diff2*	ER2
OCH2	3-PLM	73,998.54	0.525	60,596.42	0.469
	KST	35,086.62	0.554	31,362.79	0.489
OCN2	3-PLM	33,930.03	0.654	28,285.04	0.608
	KST	10,032.48	0.703	9,533.31	0.670
OLC2	3-PLM	107,472.32	0.468	9,2791.06	0.417
	KST	58,063.75	0.444	4,7836.68	0.339
OMT2	3-PLM	15411.09	0.703	15000.62	0.689
	KST	5499.73	0.726	5495.32	0.722
OCH4	3-PLM	173073.27	0.395	143670.15	0.341
	KST	76414.29	0.383	61713.23	0.305
OCN4	3-PLM	129734.82	0.538	106335.90	0.473
	KST	49658.87	0.553	40745.30	0.458
OLC4	3-PLM	215578.60	0.363	186312.88	0.315
	KST	74781.12	0.327	64244.77	0.253
OMT4	3-PLM	46168.01	0.579	43988.19	0.559
	KST	29541.62	0.540	29361.17	0.520

Note. OCH2 = Human Sciences and its Technologies for students in state schools; OCN2 = Natural Sciences and its Technologies for students in state schools; OLC2 = Language, Codes and their technologies for students in state schools; OMT2 = Mathematics and its Technologies for students in state schools; OCH4 = Human Sciences and its Technologies for students in private schools; OCN4 = Natural Sciences and its Technologies for students in private schools; OLC4 = Language, Codes and their technologies for students in private schools; OMT4 = Mathematics and its Technologies for students in private schools; ER = error rate; diff = goodness of fit coefficient.

To take the careless errors and lucky guesses into account, we have fitted the BLIMs, the results of which are shown in Table 2. Overall, KST performed better than the 3-PLM. The only exception occurred with the dataset representing the subtest Language, Codes and their technologies for students from private schools (OLC4).

Table 2

Model comparison of the BLIM fit

Dataset	Model	AIC	ΔAIC	LR	*w*
OCH2	3-PLM	199,054.04	2,933.56	0.000	0.000
	KST	196,120.47	0.00	1.000	1.000
OCN2	3-PLM	207,345.12	387.02	0.000	0.000
	KST	206,958.10	0.00	1.000	1.000
OLC2	3-PLM	190,300.79	1,406.02	0.000	0.000
	KST	188894.77	0.00	1.000	1.000
OMT2	3-PLM	207259.76	101.89	0.000	0.000
	KST	207157.88	0.00	1.000	1.000
OCH4	3-PLM	212658.61	4,815.62	0.000	0.000
	KST	207842.98	0.00	1.000	1.000
OCN4	3-PLM	222492.43	716.67	0.000	0.000
	KST	221775.76	0.00	1.000	1.000
OLC4	3-PLM	218614.58	0.00	1.000	1.000
	KST	219717.53	1,102.94	0.000	0.000
OMT4	3-PLM	246813.16	4794.10	0.000	0.000
	KST	242019.06	0.00	1.000	1.000

Note. OCH2: Human Sciences and its Technologies for students in state schools. OCN2: Natural Sciences and its Technologies for students in state schools. OLC2: Language, Codes and their technologies for students in state schools. OMT2: Mathematics and its technologies for students in state schools. OCH4: Human Sciences and its Technologies for students in private schools. OCN4: Natural Sciences and its Technologies for students in private schools. OLC4: Language, Codes and their technologies for students in private schools. OMT4: Mathematics and its technologies for students in private schools. AIC: Akaike information criterion. ΔAIC: difference of AIC. LR: likelihood ratio. *w*: model weight.

The results in Table 3 indicate that, overall, the scores generated by the KST model and the 3-PLM are not equivalent. The only case where the scores are somewhat equivalent occurs in the dataset of the subtest Mathematics, and its technologies for students from private schools (OMT4).

Table 3

Model comparison of the equivalency of the scores generated by the KST model and the 3-PLM.

Dataset	Model	AIC	ΔAIC	LR	*w*
OCH2	Monotonic	7268.43	15.13	0.001	0.001
	Nonmonotonic	7253.29	0.00	1.000	0.999
OCN2	Monotonic	9312.83	86.09	0.000	0.000
	Nonmonotonic	9226.75	0.00	1.000	1.000
OLC2	Monotonic	8715.17	162.88	0.000	0.000
	Nonmonotonic	8552.29	0.00	1.000	1.000
OMT2	Monotonic	10668.04	68.23	0.000	0.000
	Nonmonotonic	10599.81	0.00	1.000	1.000
OCH4	Monotonic	8311.89	30.15	0.000	0.000
	Nonmonotonic	8281.75	0.00	1.000	1.000
OCN4	Monotonic	10583.78	63.38	0.000	0.000
	Nonmonotonic	10520.39	0.00	1.000	1.000
OLC4	Monotonic	11365.64	317.67	0.000	0.000
	Nonmonotonic	11047.98	0.00	1.000	1.000
OMT4	Monotonic	9327.51	0.33	0.846	0.458
	Nonmonotonic	9327.17	0.00	1.000	0.542

Note. OCH2: Human Sciences and its Technologies for students in state schools. OCN2: Natural Sciences and its Technologies for students in state schools. OLC2: Language, Codes and their technologies for students in state schools. OMT2: Mathematics and its technologies for students in state schools. OCH4: Human Sciences and its Technologies for students in private schools. OCN4: Natural Sciences and its Technologies for students in private schools. OLC4: Language, Codes and their technologies for students in private schools. OMT4: Mathematics and its technologies for students in private schools. AIC: Akaike information criterion. ΔAIC: difference of AIC. LR: likelihood ratio. *w*: model weight.

State and Private Schools KST Comparison

Table 4 shows that the goodness-of-fit indices are half congruent with the estimated KST model and half incongruent. This means that for about half the datasets, the fit is better when the model assessed is the same model fitted to the dataset. Then, the model fitted to the dataset of the students of the state schools usually provides a better fit to the dataset of the students of the state schools than for the dataset of the students of the private schools. The same can be said about the model fitted to the dataset of the students of the private schools. But the datasets ~~OCH2~~, OLC2, OCH4, OLC4, and OMT 4 are incongruent with this pattern.

Table 4

Fit performance for the knowledge spaces estimated with the KST for private and state schools.

Dataset	Model	*diff1*	ER1	*diff2*	ER2
OCH2	Private	35280.57	0.522	30285.20	0.454
	State	35086.62	0.554	31362.79	0.489
OCN2	Private	14134.03	0.672	12175.27	0.612
	State	10032.48	0.703	9533.31	0.670
OLC2	Private	43723.14	0.440	39559.29	0.370
	State	58063.75	0.444	47836.68	0.339
OMT2	Private	5875.15	0.681	5861.24	0.673
	State	5499.73	0.726	5495.32	0.722
OCH4	Private	85998.36	0.423	76579.52	0.357
	State	76414.29	0.383	61713.23	0.305
OCN4	Private	37880.12	0.598	36451.02	0.558
	State	49658.87	0.553	40745.30	0.458
OLC4	Private	107,840.35	0.350	90320.47	0.259
	State	74,781.12	0.327	64244.77	0.253
OMT4	Private	30418.02	0.601	30394.22	0.594
	State	29541.62	0.540	29361.17	0.520

Note. OCH2: Human Sciences and its Technologies for students in state schools. OCN2: Natural Sciences and its Technologies for students in state schools. OLC2: Language, Codes and their technologies for students in state schools. OMT2: Mathematics and its technologies for students in state schools. OCH4: Human Sciences and its Technologies for students in private schools. OCN4: Natural Sciences and its Technologies for students in private schools. OLC4: Language, Codes and their technologies for students in private schools. OMT4: Mathematics and its technologies for students in private schools. ER: error rates.

However, when careless errors and lucky guesses are taken into account with BLIMs, we see from Table 5 that the best fitting model is always congruent with the originating dataset, with no exceptions.

Table 5

Model comparison of the BLIM fit.

Dataset	Model	AIC	ΔAIC	LR	*W*
OCH2	Private	200479.41	4358.94	0.000	0.000
	State	196120.47	0.00	1.000	1.000
OCN2	Private	211598.93	4640.83	0.000	0.000
	State	206958.10	0.00	1.000	1.000
OLC2	Private	189701.99	807.22	0.000	0.000
	State	188894.77	0.00	1.000	1.000
OMT2	Private	209253.67	2095.79	0.000	0.000
	State	207157.88	0.00	1.000	1.000
OCH4	Private	203823.26	0.00	1.000	1.000
	State	207842.98	4019.73	0.000	0.000
OCN4	Private	211110.01	0.00	1.000	1.000
	State	221775.76	10665.75	0.000	0.000
OLC4	Private	211174.93	0.00	1.000	1.000
	State	219717.53	8542.59	0.000	0.000
OMT4	Private	240503.12	0.00	1.000	1.000
	State	242019.06	1515.94	0.000	0.000

Note. OCH2: Human Sciences and its Technologies for students in state schools. OCN2: Natural Sciences and its Technologies for students in state schools. OLC2: Language, Codes and their technologies for students in state schools. OMT2: Mathematics and its technologies for students in state schools. OCH4: Human Sciences and its Technologies for students in private schools. OCN4: Natural Sciences and its Technologies for students in private schools. OLC4: Language, Codes and their technologies for students in private schools. OMT4: Mathematics and its technologies for students in private schools. AIC: Akaike information criterion. ΔAIC: difference of AIC. LR: likelihood ratio. *w*: model weight.

The results in Table 6 indicate that, overall, the scores generated by the KST model with the private schools dataset and with the state schools dataset are not equivalent. For OCH2, OLC2 and OCN4 the results are more ambiguous, but for all the other cases, the

relation between the scores are usually nonmonotonic, indicating strong discrepancies between the estimated scores.

Table 6

Model Comparison of the Equivalency of the Scores Generated by the KST Model with the Private Schools Dataset and with the State Schools Dataset

Dataset	Model	AIC	ΔAIC	LR	*w*
OCH2	Monotonic	49255.99	0.00	1.000	0.872
	Nonmonotonic	49259.82	3.83	0.148	0.129
OCN2	Monotonic	52967.31	916.18	0.000	0.000
	Nonmonotonic	52051.13	0.00	1.000	1.000
OLC2	Monotonic	50996.82	0.04	0.982	0.495
	Nonmonotonic	50996.78	0.00	1.000	0.505
OMT2	Monotonic	48088.61	83.37	0.000	0.000
	Nonmonotonic	48005.24	0.00	1.000	1.000
OCH4	Monotonic	59079.75	111.93	0.000	0.000
	Nonmonotonic	58967.82	0.00	1.000	1.000
OCN4	Monotonic	63291.14	0.00	1.000	0.618
	Nonmonotonic	63292.10	0.96	0.619	0.382
OLC4	Monotonic	61444.74	208.66	0.000	0.000
	Nonmonotonic	61236.08	0.00	1.000	1.000
OMT4	Monotonic	60194.12	25.57	0.000	0.000
	Nonmonotonic	60168.54	0.00	1.000	1.000

Note. OCH2: Human Sciences and its Technologies for students in state schools. OCN2: Natural Sciences and its Technologies for students in state schools. OLC2: Language, Codes and their technologies for students in state schools. OMT2: Mathematics and its technologies for students in state schools. OCH4: Human Sciences and its Technologies for students in private schools. OCN4: Natural Sciences and its Technologies for students in private schools. OLC4: Language, Codes and their technologies for students in private schools. OMT4: Mathematics and its technologies for students in private schools. AIC: Akaike information criterion. ΔAIC: difference of AIC. LR: likelihood ratio. *w*: model weight.

Discussion

This study aimed to compare the knowledge structure of ENEM 2019 respondents from two different groups: students form public and private schools. For this, a stratified subsample of 5,000 respondents was used, analyzing all four subtests of the exam. The knowledge space was estimated and, subsequently, the model fit and the model performance were compared. Overall, evidence suggest that in the context of the ENEM, the Knowledge Space Theory (KST) would be better than the Item Response Theory (IRT) in describing the response patterns found in the data. In addition, the use of the KST instead of IRT would imply different results in terms of passing or failing the test. Furthermore, the results indicate that respondents from public and private schools may learn the same knowledge differently.

Should ENEM be scored using IRT or KST?

The results in Table 1 lead us to the conclusion that KST may be more appropriate than the 3-PLM to describe and assess the performance of respondents on ENEM. In a sense, our results support the results and arguments of previous studies that criticize the use of IRT in ENEM (e.g., Gomes et al., 2020), and educational large-scale assessment in general (e.g., Doble et al., 2019; Segedinac, 2018). KST, in principle, takes into account the ways in which different groups of people learn and, therefore, could be a better representation of the skills developed by students and respondents in a given knowledge field. However, our results have also shown that the error rate is quite similar between the KST and the linear learning space, what could indicate that, if random error (i.e., careless errors and lucky guesses) is not taken into account, both models describe the data quite similarly.

In order to consider careless errors and lucky guesses when comparing the goodness-of-fit of the knowledge space estimated by KST and by 3-PLM, we used the basic local independence model (BLIM), fitting the estimated spaces to our datasets. As shown in Table

2, in general, KST still performed better than IRT in all datasets, with the only exception being the data from the Languages, Codes and their technologies subtest of respondents from private schools. This result could indicate one of at least two possibilities. The first one, more naïve, is that only for this specific exam, the 3-PLM describes the data better than the KST in the presence of random error. The second possibility, which is less obvious, is that the knowledge space for Languages, Codes and Technologies subtest is more complex that what was allowed to be estimated by our method (which can only estimate knowledge spaces as complex as the one determined by the 3-PLM). This second hypothesis seems to be more reasonable, given that language is a type of fluid knowledge and it is quite sensible to contextual influences, especially in a country as large as Brazil (Carmo et al., 2019; Lucena & Santos, 2020; Marcom & Kleinke, 2017).

The results shown in Table 3 are related to another important point of this study: whether the scores generated by KST and 3-PL models are equivalent or not. Our results indicate that the scores derived from KST and IRT are not equivalent to each other. The only exception found was in the Mathematics and their technologies subtest of respondents from private schools. In this case, the results were not conclusive if the scores can be considered to be equivalent or not. In contrast to what was discussed in regards to the Languages, Codes and their technologies subtest of respondents from private schools, mathematics is a more standardized field and it would be expected that the learning in this area is more linear than in other areas (e.g., Vieira & Drigo, 2021).

The fact that the scores generated by the KST and 3-PL models were not equivalent for most datasets means that, in practical terms, depending on the correction methodology used in the ENEM, the same respondent could obtain different scores on the test. This implication signifies that using a different scoring method could directly impact the lives of a large number of people who take ENEM in order to enter higher education. Therefore,

despite the fact that IRT is, in general, considered a great improvement over other traditional psychometrics approaches (Sousa, 2020), our results raise questions regarding the necessity of applying a scoring methodology that could be more adequate for ENEM. Of course, as happened in the context of this research, other types of assessments, large or small scale, could also require different types of psychometric assessments to guarantee that the scoring, or classification methodology, is valid (Heck, 2018; Liu et al, 2007; Soares & Soares, 2020).

Differences Between the Knowledge Spaces of Private and Public Schools

The results in Table 4 indicate that half of the goodness-of-fit indices are congruent with the estimated KST model, while the other half is not. This means that when the model fitted to the state school dataset is assessed in regards to the datasets of state schools, it will typically have a better fit. The same goes for the datasets and models of private schools. However, we found that some datasets were incompatible with this pattern: Human Sciences and its Technologies for students in state schools; Language, Codes and their technologies for students in state schools; Human Sciences and its Technologies for students in private schools; Language Codes and their technologies for students in private schools; Mathematics and its technologies for students in private schools. However, when considering careless errors and lucky guesses with BLIMs in Table 5, the best fitting-model is always congruent with the source data, with no exceptions.

In terms of practical implications, the results presented in Tables 4 and 5 indicate that students from private and public schools learn differently. These results are supported by many previous theoretical (Figueiredo et al., 2014; Marcom & Kleinke, 2017; Osher et al., 2020) and empirical (e.g., Kleinke, 2017; Lima Junior, 2015; Lucena & Santos, 2020; Nascimento et al, 2018) studies. It should be noted that, again, these results are dependent on the validity of our procedure for estimating the knowledge space. Because it is a greedy optimization procedure, it may result in overfitting (e.g., Norouzi et al., 2015). Nevertheless,

even previous studies from IRT (e.g., Lucena & Santos, 2020) and CTT (e.g., Sousa & Braga, 2020) perspectives have found evidence of the performance on the ENEM, in terms of possible response patterns, is sensible to the socioeconomic group a person comes from. The main innovation of this study in this regard is that a KST model also tries to explain how exactly each group learns.

The results in Table 6 indicate that, overall, the scores generated by the KST model with state and private schools are not equivalent to each other. More ambiguous results were found in Human Sciences and their Technologies at a state school, Languages, Codes and their technologies at a state school, and Natural Sciences at a private school. In this sense, this evidence reinforces the previous conclusion that there are distinctions between the spaces of knowledge of respondents from private and state schools. Considering all our results, this study indicates that KST could add improvements in the context of large-scale educational assessments. The main contribution comes from the fact that KST allows for the identification of different learning paths for each individual, which seems to be a process that occurs in a non-linear way (Doignon & Falmagne, 2015). This is a strong paradigm shift from the linearity characteristic of the IRT and the idea that everyone learns in the same way (Valle, 2000).

Final Remarks

In this study, we discuss large-scale educational assessments in the context of ENEM. More specifically, we compared two psychometric theories: Item Response Theory, the current theory applied in ENEM; and Knowledge Space Theory, a theory proposed as an alternative for linearly based psychometric theories. In addition to verifying and comparing the theoretical, practical, and methodological implications of the theories, we also compared the knowledge structures of respondents from different groups, from state and private

schools, in order to test if different groups learn the same knowledge differently. With this study, we found evidence suggesting that different groups learn the same knowledge differently and, therefore, that KST would be better than IRT in analyzing the results of ENEM. Finally, the use of each theory would directly affect the result of the exam, allowing the same respondent to have different scores depending on the correction methodology used.

To correctly reflect the implications of our evidences, some limitations about this study should be taken into account. First, the study was carried out using only a subsample of 5,000 respondents from one edition of the ENEM. Considering the promising evidence that was found, we point out that further studies should be carried out using the complete sample, or even other editions of the ENEM. Another limitation is that there was no in-depth testing of the method we used to fit the KST. In future studies, it makes sense that the method we developed is further evaluated, or even that a more efficient version of the IITA algorithm is developed. A final limitation worth mentioning would be that, at least for now, we are not aware of a widespread and evidently efficient way of comparing the performances of KST and IRT models. In this sense, we encourage future studies to test further if the approaches we followed in the current study are generalizable and capable of finding true distinctions.

The main theoretical implication of this study is that individuals from state schools and private schools learn differently. This can sound as a trivial conclusion, but it is not uncommon for statistical fitting and testing of models applied to large-scaled educational assessments to not take into account this issue. Also, because we derived this implication from a KST model, rather than an IRT model, we have also found evidence of the non-linearity of learning. Of course, these implications are quite dependent on the methodological implication of our study, which indicates that our analytical method to fit a KST model can be useful in other research contexts. Even though it is necessary to test the method more

thoroughly in future studies, we consider that the development of this method can help in popularizing KST applications and research.

The most important implication of this study is that, despite the popularity and efficiency of IRT in the context of large-scale educational assessments, our evidence suggests that this theory is possibly not the most suitable for ENEM. This means that there is a chance that the test respondents are not being selected for access to public and private higher education institutions in the best possible way. This practical implication reinforces the urgency and necessity for future studies comparing the KST with the IRT. As a final account of the innovations of this study, it worth mentioning two of its main characteristics. First, this is one of the first studies comparing KST directly with IRT, even taken into account an international research context. For the best of our knowledge, this is study is also the first study that proposed to apply the KST to the data of ENEM. We hope that our analysis and discussion will provide some guidance for future studies and in doing so to keep improving the quality of the results of large-scale educational assessments.

References

Almeida, F. M. D. (2020). Avaliação da equidade dos cadernos de matemática do Enem 2017 [Equity assessment of ENEM 2017 math notebooks]. [Master's thesis, Federal University of Ceará]. http://repositorio.ufc.br/handle/riufc/51155

Andrade, G. G. (2012). A metodologia do Enem: uma reflexão [The ENEM methodology: A reflection]. *Revista Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB*, *33,* 68-76. https://serie-estudos.ucdb.br/serie-estudos/article/view/71

Andrade, J. M., Laros, J. A., & Gouveia, V. V. (2010). O uso da Teoria de Resposta ao Item em avaliações educacionais: diretrizes para pesquisadores [The use of item-response theory in educational assessments: Guidelines for researchers]. *Avaliação Psicológica*, *9*(3), 421-435. https://www.redalyc.org/pdf/3350/335027284009.pdf

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria de Resposta ao Item: conceitos e aplicações* [*Item Response Theory: Concepts and applications*]. SINAPE.

Andrade, D. F., & Valle, R. C. (1998). Introdução a Teoria de Resposta ao Item [Introduction to Item Response Theory]. *Estudos em Avaliação Educacional*, (18), 13-32. https://doi.org/10.18222/eae01819982250

Blanco, J. (2013). A avaliação de Língua Inglesa no ENEM: efeitos de seu impacto social no contexto escolar [English language assessment at ENEM: Effects of its social impact on the school context]. [Master 's thesis, Federal University of São Carlos]. https://repositorio.ufscar.br/handle/ufscar/5768

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425. https://doi.org/10.1007/s11336-006-1447-6

Carmo, R. V., Heckler, W. F., & de Carvalho, J. V. (2020). Uma análise do desempenho dos estudantes do Rio Grande do Sul no ENEM de 2019 [An analysis of the performance

of Rio Grande do Sul students at ENEM 2019]. *RENOTE*, *18*(2), 378-387. DOI: 10.22456/1679-1916.110257

Constituição do Brasil (1988). (Constituição da República Federativa do Brasil, 1988/2001).

Cosyn, E., Uzun, H., Doble, C., & Matayoshi, J. (2021). A practical perspective on knowledge space theory: ALEKS and its data. *Journal of Mathematical Psychology*, *101*, 102512. https://doi.org/10.1016/j.jmp.2021.102512

Cunha, E. C. S., & Müller, E. R. (2018). Avaliações em larga escala: uma tentativa de controle, regulação, captura e padronização do cotidiano escolar [Large-scale evaluations: An attempt to control, regulate, capture and standardize school life]. *Cadernos da FUCAMP*, *17*(29). http://www.fucamp.edu.br/editora/index.php/cadernos/article/view/1317/943

Dickel, A. (2016). The national literacy assessment in the context of the basic education assessment system and the National Pact for Literacy in the right age: Accountability and control. *Cadernos Cedes*, *36*, 193-206. https://doi.org/10.1590/CC0101-32622016162940

Doble, C., Matayoshi, J., Cosyn, E., Uzun, H., & Karami, A. (2019). A data-based simulation study of reliability for an adaptive assessment based on knowledge space theory. *International Journal of Artificial Intelligence in Education*, *29*(2), 258-282. https://doi.org/10.1007/s40593-019-00176-0

Doignon, J. P. & Falmagne, J. C. (1999). *Knowledge spaces*. Springer.

Doignon, J. P., & Falmagne, J. C. (2015). Knowledge spaces and learning spaces. *arXiv preprint arXiv:1511.06757*, 1-51.

Emler, T. E., Zhao, Y., Deng, J., Yin, D., & Wang, Y. (2019). Side effects of large-scale assessments in education. *ECNU Review of Education*, *2*(3), 279-296. https://doi.org/10.1177/2096531119878964

Falmagne, J. C., Cosyn, E., Doignon, J. P., & Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. In *Formal concept analysis* (pp. 61-79). Springer. ~~Berlin, Heidelberg.~~

Falmagne, J-C., Albert, D., Doble, C., Eppstein, D., & Hu, X. (2013). *Knowledge Space: Application in Education*. (1 ed.). Springer.

Figueirêdo, E., Nogueira, L., & Santana, F. L. (2014). Igualdade de oportunidades: analisando o papel das circunstâncias do desempenho do ENEM [Equal opportunities: Analyzing the role of circumstances in the performance of ENEM.] *Revista Brasileira de Economia*, *68*(3), 373-392. https://doi.org/10.1590/s0034-71402014000300005

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 456–465. https://doi.org/10.1177/2515245920952393

Folha (2021). Inglês no ENEM é obstáculo entre aluno de escola pública e a faculdade [English in ENEM is an obstacle to enter college for public school students]. Retrieved 22 May 2021, from https://www.folhape.com.br/noticias/ingles-no-enem-e-obstaculo-entre-aluno-de-escola-publica-e-a-faculdade/174392/.

Garcia, P. S., Prearo, L. C., do Carmo Romero, M., Secco, A., & Bassi, M. S. (2016). School performance: An IDEB analysis of the seven municipalities in the ABC region. *Revista Eletrônica de Educação*, *10*(2), 115-134. https://doi.org/10.14244/198271991784

Gomes, C. M. A., Golino, H. F., & de Souza Peres, A. J. (2020). Fidedignidade dos escores do Exame Nacional do Ensino Médio (ENEM) [Score reliability of the national high school examination (ENEM)]. *Psico*, *51*(2), e31145-e31145. https://doi.org/10.15448/1980-8623.2020.2.31145

Gonçalves, A.D.B.V., & Silveira, A.A.D. (2021). A exigibilidade do direito à educação básica no Brasil: estado da arte das teses e dissertações de 1988 a 2018 [The enforceability of the right to basic education in Brazil: State of the art of theses and dissertations from 1988 to 2018] *Revista Educação e Políticas em Debate*, *10*(2), 922-940. https://doi.org/10.14393/REPOD-v10n2a2021-58553

Harati, H., Sujo-Montes, L., Tu, C. H., Armfield, S. J., & Yen, C. J. (2021). Assessment and Learning in Knowledge Spaces (ALEKS): Adaptive system impact on students' perception and self-regulated learning skills. *Education Sciences*, *11*(10), 603. https://doi.org/10.3390/educsci11100603

Heck, M. F. (2018). Reflexões acerca do Sistema Nacional de Avaliação da Educação Básica [Reflections on the national basic education assessment system (SAEB)]. *REAMEC-Rede Amazônica de Educação em Ciências e Matemática*, *6*(1), 124-141. http://doi.org/10.26571/REAMEC.a2018.v6.n1.p124-141.i6183

Heller, J., & Wickelmaier, F. (2013). Minimum discrepancy estimation in probabilistic knowledge structures. *Electronic Notes in Discrete Mathematics, 42*, 49-56. https://doi.org/10.1016/j.endm.2013.05.145

Horta Neto, J. (2010). Avaliação externa de escolas e sistemas: questões presentes no debate sobre o tema [External evaluation of schools and systems: Current issues in the debate on the topic]. *Revista Brasileira de Estudos Pedagógicos*, *91*(227). https://doi.org/10.24109/2176-6681.rbep.91i227.604

Hutz, C. S., Bandeira, D. R., & Trentini, C. M. (2015). *Psicometria* [*Psychometrics*]. Artmed Editora.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2019). *Microdados do Enem 2019. Leia-me* [*Microdata of Enem 2019. Read me*]. http://portal.inep.gov.br/web/guest/microdados.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2019). *Microdados do Enem 2019. Edital* [*Microdata of Enem 2019. Edital*]. http://portal.inep.gov.br/web/guest/microdados.

Kleinke, K. (2017). Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics*, *42*(4), 371-404. https://doi.org/10.3102/1076998616687084

Klieger, D. M., Bridgeman, B., Tannenbaum, R. J., Cline, F. A., & Olivera-Aguilar, M. (2018). The validity of GRE® general test scores for predicting academic performance at US law schools. *ETS Research Report Series, 2018*(1), 1-28. https://doi.org/10.1002/ets2.12213

Le, D. T. (2013). Applying item response theory modeling in educational research [Doctoral Thesis, Iowa State University]. https://lib.dr.iastate.edu/etd/13410/

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE Publications.

Lima Junior, P. (2015). Crítica sociológica ao Exame Nacional do Ensino Médio: uma análise bourdiana [Sociological critique of the National High School Exam: A Bourdian analysis]. *Encontro Nacional De Pesquisa Em Educação Em Ciências*, *10*, 1-8. http://www.abrapecnet.org.br/enpec/x-enpec/anais2015/resumos/R1971-1.PDF

Liu, J., Harris, D. J., & Schmidt, A. (2007). 33 Statistical procedures used in college admissions testing. *Handbook of statistics*, *26*, 1057-1091.

Lucena, J., & dos Santos, H. (2020). A relação entre o desempenho no Exame Nacional do Ensino Médio e o perfil socioeconômico: um estudo com os microdados de 2016. [The relationship between performance on the National High School Exam and

socioeconomic profile: A study with microdata of 2016]. *Revista de Gestão e Secretariado, 11*(2), 1-23. https://doi.org/10.7769/gesec.v11i2.994

Marcom, G. S., & Kleinke, M. U. (2017). Gênero e status socioeconômico: reflexões sobre o desempenho dos candidatos na prova de ciências da natureza do ENEM 2014. [Gender and socioeconomic status: Reflections on the performance of candidates on the ENEM science test of 2014]. *Perspectiva Sociológica: A Revista de Professores de Sociologia*, (19), 44-52. http://dx.doi.org/10.33025/rps.v0i19.1174

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement*, *7*(2), 75-88. https://doi.org/10.1080/15366360903070385

Marley, J. E. (2000). Efficacy, effectiveness, efficiency. https://digital.library.adelaide.edu.au/dspace/bitstream/2440/32757/1/hdl_32757.pdf

Matias, A. B., Quaglio, G. M., Oliveira, B. G., Lima, J. P. R., & Bertolin, R. V. (2019). Expenditures' level and efficiency in public education: A study of São Paulo municipalities using data envelopment analysis. *Brazilian Journal of Management*, *11*(4), 1051-1067. https://doi.org/10.5902/1983465916448

McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press.

Mendes, M., & da Costa Nunes, M. A. (2019). English or Spanish? What factors do students privileged in choosing a language for ENEM? *Vivências*, *15*(28), 124-134. https://doi.org/10.31512/vivencias.v15i28.20

Michell, J. (2008). Is psychometrics pathological science? *Measurement*, *6*(1-2), 7-24. https://doi.org/10.1080/15366360802035489

Minhoto, M. A. P. (2016). Política de avaliação da educação brasileira: limites e perspectivas [Brazilian education assessment policy: Limits and perspectives]. *Jornal de Políticas Educacionais*, *10*(19).

Nascimento, M. M., Cavalcanti, C., & Ostermann, F. (2018). Uma busca por questões de Física do ENEM potencialmente não reprodutoras das desigualdades socioeconômicas [A search for Physics questions of ENEM that are potentially non-reproductive of socioeconomic inequalities]. *Revista Brasileira de Ensino de Física*, *40*. https://doi.org/10.1590/1806-9126-RBEF-2017-0237

Norouzi, M., Collins, M., Johnson, M. A., Fleet, D. J., & Kohli, P. (2015). Efficient non-greedy optimization of decision trees. *Advances in Neural Information Processing Systems, 28*. https://proceedings.neurips.cc/paper/2015/hash/1579779b98ce9edb98dd85606f2c119d-Abstract.html

Norouzi, M., Fleet, D. J., & Salakhutdinov, R. (2012). Hamming distance metric learning. In F. Pereira, C.J. Burges, L. Bottou & K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems (NIPS) 25* (p. 1070-1078). NeurIPS Proceedings

Osher, D., Cantor, P., Berg, J., Steyer, L., & Rose, T. (2020). Drivers of human development: How relationships and context shape learning and development. *Applied Developmental Science*, *24*(1), 6-36. https://doi.org/10.1080/10888691.2017.1398650

Pasquali, L. (2017). *Psicometria: teoria dos testes em psicologia e educação.* [*Psychometrics: Theory of tests in psychology and education*]. Editora Vozes Limitada.

Parker, R. G., & Rardin, R. L. (2014). *Discrete optimization*. Elsevier.

Pinheiro, L. L. S., & Quevedo-Camargo, G. (2017). Efeito retroativo e multimodalidade no ENEM: análise das questões de inglês e espanhol [Retroactive effect and multimodality at ENEM: Analysis of English and Spanish issues]. *Signum: Estudos da Linguagem*, *20*(1), 136-166. http://dx.doi.org/10.5433/2237-4876.2017v20n1p136

Primi, R. (2004). Advances in the interpretation of scales with the application of Item Response Theory. *Avaliação Psicológica, 3*(1), 53-58. https://bv.fapesp.br/en/publicacao/7226/advances-in-scale-interpretation-with-the-application-of-ite/

Porcu, M., & Giambona, F. (2017). Introduction to latent class analysis with applications. *The Journal of Early Adolescence*, *37*(1), 129-158. https://doi.org/10.1177/0272431616648452

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika, 35*(1), 73-78. https://doi.org/10.1007/BF02290594

Pya, N. (2022). *scam: Shape Constrained Additive Models*. https://cran.r-project.org/package=scam

Pya, N., & Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing, 25*(3), 543-559. https://doi.org/10.1007/s11222-013-9448-7

Quevedo-Silva, F., & Sauer, L. (2012). Competição justa? A relação entre desempenho no vestibular e perfil socioeconômico [Fair competition? The relationship between vestibular performance and socioeconomic profile]. *Pensamento & Realidade*, *27*(1). https://revistas.pucsp.br/pensamentorealidade/article/view/11588

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rajeev, S., & Krishnamoorthy, C. S. (1992). Discrete optimization of structures using genetic algorithms. *Journal of Structural Engineering, 118*(5), 1233-1250. https://doi.org/10.1061/(ASCE)0733-9445(1992)118:5(1233)

Rodriguez-Hernandez, C. F., Cascallar, E., & Kyndt, E. (2020). Socioeconomic status and academic performance in higher education: A systematic review. *Educational Research Review*, *29*, 100305. https://doi.org/10.1016/j.edurev.2019.100305

Sampaio, B., & Guimarães, J. (2009). Diferenças entre a eficiência da educação pública e privada no Brasil [Differences in efficiency between public and private education in Brazil]. *Economia Aplicada*, *13*, 45-68.

Sartes, L. M. A., & Souza-Formigoni, M. L. O. D. (2013). Avanços em psicometria: da teoria clássica dos testes à teoria de resposta ao item [Advances in psychometrics: From classical theory of tests to item response theory]. *Psicologia: Reflexão e Crítica*, *26*(2), 241-250. https://doi.org/10.1590/S0102-79722013000200004

Segedinac, M., Segedinac, M., Konjović, Z., & Savić, G. (2011). A formal approach to organization of educational objectives. *Psihologija*, *44*(4), 307-323. http://doi.org/10.2298/PSI1104307S

Segedinac, M. T., Horvat, S., Rodić, D. D., Rončević, T. N., & Savić, G. (2018). Using knowledge space theory to compare expected and real knowledge spaces in learning stoichiometry. *Chemistry Education Research and Practice*, *19*(3), 670-680. DOI: 10.1039/C8RP00052B

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*. https://doi.org/10.1177/0959354312454353

Silva, M. S., & de Carvalho, M. C. A. (2020). Concepção dos professores sobre a Avaliação Nacional do Desempenho Escolar - Anresc [Teachers' conception of the National Assessment of School Performance – Anresc]. *Revista de Instrumentos, Modelos e Políticas em Avaliação Educacional*, *1*(2), e020011-e020011. https://doi.org/10.51281/impa.e020011

Silveira, F. L. D., Barbosa, M. C. B., & Silva, R. D. (2015). Exame Nacional do Ensino Médio: uma análise crítica [National High School Exam (ENEM): A critical analysis]. https://doi.org/10.1590/S1806-11173710001

Soares, T. M., & Soares, M. C. N. (2020). Mensuração em sistemas admissionais no ensino superior e avaliações de impacto [Measurement in admission systems in higher education and impact assessments]. *Pesquisa e Debate em Educação*, *10*(1), 1190-1223. https://doi.org/10.34019/2237-9444.2020.v10.32029

Sousa, L. A. d. (2019). Comparative analysis of the National High School Exam via classical test theory and item response theory [Doctoral Thesis, Universidade Federal do Ceará]. http://www.repositorio.ufc.br/handle/riufc/48275

Sousa, L. A., & Braga, A. E. (2020). Teoria clássica dos testes e teoria de resposta ao item em avaliação educacional [Classical Test Theory and Item Response Theory in educational assessment]. *Revista de Instrumentos, Modelos e Políticas em Avaliação Educacional*, *1*(1), e020002-e020002.

Teixeira, P. R., & Rios, M. P. G. (2017). Dez anos do SINAES: um mapeamento das teses e dissertações defendidas no período de 2004-2014 [Ten years of SINAES: A mapping of theses and dissertations defended in the period 2004-2014]. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, *22*, 793-816. https://doi.org/10.1590/S1414-40772017000300012

Townsend, J. T. (2008). Mathematical psychology: Prospects for the 21st century: A guest editorial. *Journal of Mathematical Psychology*, *52*(5), 269-280. https://doi.org/10.1016/j.jmp.2008.05.001

Ünlü, A., & Sargin, A. (2008). Maximum likelihood methodology for diff fit measures for quasi orders. *Unpublished manuscript*. https://opus.bibliothek.uni-augsburg.de/opus4/files/1194/mpreprint_08_030.pdf

Ünlü, A. & Sargin, A. (2016). *daks: Data Analysis and Knowledge*. https://CRAN.R-project.org/package=DAKS

Ünlü, A., & Schrepp, M. (2020, September 28). Generalized inductive item tree analysis. https://doi.org/10.31234/osf.io/75cxn

Valle, R. d. C. (2000). Teoria de Resposta ao Item [Item Response Theory]. *Estudos em Avaliação Educacional*, *21*, 7-92. https://doi.org/10.18222/eae02120002225

Vianna, H. M. (2003). Avaliações educacionais em larga escala: análises e propostas. [Large-scale national assessments: analysis and proposals.] *Estudos em Avaliação Educacional*, (27), 41-76. https://doi.org/10.18222/eae02720032177

Vieira, C. E. D. A. (2020). Metodologia unificada para detecção de DIF: Aplicação de itens do ENEM 2017 para candidatos com déficit de atenção [Unified methodology for detection of DIF: An application for items of ENEM 2017 for candidates with déficit of attention]. [Master's thesis, Federal University of Pará]. https://www.ppgme.propesp.ufpa.br/ARQUIVOS/dissertacoes/2020/CHARLES%20EDUARDO%20DE%20ALBUQUERQUE%20VIEIRA.pdf

Vieira, D. D. O. L., & Drigo, M. O. (2021). Dificuldades de ensino e aprendizagem em matemática no ensino superior na perspectiva de docentes e discentes [Teaching and learning difficulties in mathematics in higher education from the perspectives of professors and students]. *Série-Estudos*, *26*(58), 323-340. https://doi.org/10.20435/serie-estudos.v26i58.1569

Villani, M., & Oliveira, D. A. (2018). Avaliação nacional e internacional no Brasil: os vínculos entre o PISA e o IDEB [National and International Assessment in Brazil: The links between PISA and IDEB]. *Educação & Realidade*, 43, 1343-1362. https://doi.org/10.1590/2175-623684893

Vinha, L. G. D. A., Karino, C. A., & Laros, J. A. (2016). Factors associated with Mathematics performance in Brazilian basic education. *Psico-USF*, *21*, 87-100. https://doi.org/10.1590/1413-82712016210108

Wickelmaier, F., Heller, J., Mollenhauer, J., & Anselmi, P. (2022). *pks: Probabilistic Knowledge Structures*. https://cran.r-project.org/package=pks

Wood, S. (2022). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. https://cran.r-project.org/package=mgcv