UNIVERSIDADE DE BRASÍLIA
Centro de Estudos Avançados Multidisciplinares – CEAM
Programa de Pós-Graduação em Desenvolvimento, Sociedade e Cooperação Internacional – PPGDSCI

O MODELO MULTIFACETAS DE RASCH NA CONSTRUÇÃO DE SUBESCALAS

ANTONIA REGINA DE OLIVEIRA

BRASÍLIA
MARÇO – 2019

ANTONIA REGINA DE OLIVEIRA

O MODELO MULTIFACETAS DE RASCH NA CONSTRUÇÃO DE SUBESCALAS

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Desenvolvimento, Sociedade e Cooperação Internacional do Centro de Estudos Avançados Multidisciplinares da Universidade de Brasília, vinculada à Linha de Pesquisa Desenvolvimento, e Políticas Públicas como parte dos requisitos para a obtenção do título de Mestre sob orientação do Professor Dr. Joaquim José Soares Neto e coorientação da Drª. Camila Akemi Karino.

BRASÍLIA
MARÇO – 2019

ANTONIA REGINA DE OLIVEIRA

O MODELO MULTIFACETAS DE RASCH NA CONSTRUÇÃO DE SUBESCALAS

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Desenvolvimento, Sociedade e Cooperação Internacional do Centro de Estudos Avançados Multidisciplinares da Universidade de Brasília, vinculada à Linha de Pesquisa Desenvolvimento, e Políticas Públicas como parte dos requisitos para a obtenção do título de Mestre sob orientação do Professor Dr. Joaquim José Soares Neto e coorientação da Drª. Camila Akemi Karino.

Banca Examinadora

_____

Prof. Dr. Joaquim José Soares Neto
PPGDSCI - Universidade de Brasília (UnB)

_____

Profª Drª. Marília Miranda Forte Gomes
PPGDSCI - Universidade de Brasília (UnB)

_____

Profª Drª. Girlene Ribeiro de Jesus
Faculdade de Educação – Universidade de Brasília (UnB)

_____

Profª Drª. Maria de Fátima Rodrigues Makiuchi
PPGDSCI - Universidade de Brasília (UnB) - Suplente

À minha família, aos meus amigos e a Deus, que tornam minha caminhada mais leve e confiante; sem eles nada disso faria sentido ou seria possível.

AGRADECIMENTO

Gostaria agradecer, primeiramente, o meu orientador, professor Dr. Joaquim José Soares Neto, por todo conhecimento passado, pelo incentivo, pela paciência, pela presença constante, dedicação e todos os conselhos. Agradeço por ter possibilitado a minha primeira experiência profissional na área de avaliação. E também à minha coorientadora a Dra. Camila Akemi Karino pela amizade, pelo apoio, conhecimentos e paciência em todas as horas.

Às professoras, Dra. Marília Miranda Forte Gomes, Dra. Girlene Ribeiro de Jesus e Dra. Maria de Fátima Rodrigues Makiuchi, que aceitaram gentilmente o convite e a grande responsabilidade de aferir minhas habilidades e conhecimentos.

Ao Dr. Wellington Silva por seu tempo e disponibilidade em ensinar sobre a metodologia e ao Centro de Políticas Públicas e Avaliação da Educação - Caed/UFJF que viabilizou os materiais utilizados no estudo.

A todos os professores do Programa de PGDSCI, pelos seus ensinamentos, pelo apoio, pelo aprendizado que me proporcionaram durante o período de realização do curso.

Aos colegas do Programa de Pós-graduação em Desenvolvimento Sociedade e Cooperação internacional – PPGDSCI, os colegas de sala por todas as experiências e trocas de informações, em especial ao grupo de pesquisa em avaliação educacional: Claudete, Andrea, Alice, Fernanda, Layla, Cristian, Mary e todos que passaram pelas nossas apresentações, pelo suporte e troca de experiências com discussões enriquecedoras para que isso fosse possível. A vocês um imenso obrigado.

Aos funcionários da secretaria de PPGDSCI, em especial ao André, pessoas que trabalham nos bastidores, mas imprescindíveis para que as coisas transcorram da melhor forma possível.

Aos meus amigos que estiveram ao meu lado durante todos os bons momentos e não tão bons assim, que me ajudaram, que conhecem todos os bastidores dessa caminhada e que me demonstraram o valor de se ter amizade. A minha jornada foi mais leve e feliz com vocês ao meu lado. A vocês a minha grande e sincera gratidão.

Aos colegas dos Correios, por terem me apoiado a continuar estudando e principalmente por terem acreditado no meu trabalho nesta instituição. Aproximando e conectando pessoas!

A todos da minha família – em especial à minha mãe Graça, pelo amor incondicional –, que torceram por mim, que acreditaram no meu potencial, que me deram força para persistir, que compreenderam minhas ausências e que estiveram presentes de diversas maneiras ao longo dessa jornada de dois anos. Obrigada pelo carinho e apoio, meu amor por vocês é infinito!

Finalmente, e senão o mais importante, agradeço a Deus pelas bênçãos recebidas, pela saúde, pela proteção nos momentos difíceis e pela serenidade alcançada. Senhor, obrigada, por estar comigo, me dar forças para prosseguir, por ser luz no meu caminho.

RESUMO

A avaliação em larga escala é um importante instrumento que possibilita mensurar características dos diversos atores envolvidos no processo educacional. E seus resultados podem ajudar na criação de políticas públicas educacionais para a melhoria da qualidade educacional. Além disso, são muitos os fatores que podem causar variabilidade nas pontuações atribuídas aos examinandos, podendo comprometer a justiça da avaliação. O Modelo Multifacetas de Rasch (MFRM) se destaca por permitir a inclusão no modelo de variáveis que podem ser geradoras de viés na avaliação, as chamadas *facetas*, permitindo que a análise dos efeitos causados por cada elemento da avaliação seja feita de forma individual. Na área educacional, as facetas normalmente consistem em variações de examinandos, tarefas/itens (dificuldade), critérios, escalas de pontuação (entendimento e utilização) e avaliadores (severidade, tendências a julgamentos sistemáticos e características pessoais). E nesse cenário são muitos os fatores que podem causar variabilidade nas pontuações atribuídas, podendo comprometer a justiça da avaliação. O objetivo deste estudo é analisar os efeitos de um caderno de prova que na avaliação considera uma grande área de concentração, agregando diversas disciplinas - são criadas as subescalas para contemplar as disciplinas agrupadas nos blocos - e como esses efeitos podem influenciar a mensuração do desempenho do examinando nos testes, como podem auxiliar na análise pedagógica fornecida aos educadores. Desta forma, com a aplicação do Modelo Multifacetas de Rasch com três facetas (habilidade do examinando, dificuldade do item e subescalas) em um bloco de questões de uma avaliação educacional em larga escala é possível estimar os efeitos das facetas de forma direta e individual. Neste trabalho foi feita a aplicação da metodologia no bloco de Ciências da natureza e suas tecnologias do ENEM 2016, nesse caderno estão contempladas as disciplinas de Física, Química e Biologia. Mesmo o caderno contendo três disciplinas distintas ele é unidimensional, o construto medido é único, garantido pelas análises de dimensionalidade. A contribuição do MFRM está relacionada à possibilidade de aplicação prática de metodologia para análise e estimação dos principais efeitos da aplicação desse tipo de prova, determinando o quanto esses traços influenciam o resultado do desempenho dos examinandos, podendo também subsidiar outras áreas da avaliação, como a análise pedagógica que serve de insumo para os educadores na melhoria da qualidade do processo de ensino-aprendizagem. A análise da amostra de respostas foi realizada no software *FACETS* com a aplicação do modelo com três facetas. As principais estatísticas, *MQ-Infit* e *MQ-Outfit*, apresentaram resultados muito satisfatórios para todas as facetas estudadas. Tais resultados demonstram que um bom ajuste do modelo, desta forma os resultados se mostraram favoráveis para a medição. Os resultados da aplicação do modelo têm qualidade, apontando ser favorável para utilização destes conjuntamente com análises pedagógicas.

Palavras-chave: subescalas; Modelo Multifacetas de Rasch; avaliação educacional; análise pedagógica e qualidade da educação.

ABSTRACT

Large-scale evaluation is an important instrument to measure the characteristics of different actors involved in the educational process. Their results can help in the creation of educational public policies for the improvement of educational quality. In addition, there are many factors that can cause variability in the scores attributed to the examinee, and may compromise the fairness of the evaluation. The Many-Facets Rasch Measurement (MFRM) stands out because it allows the inclusion in the model of variables that can generate bias in the evaluation, called facets, allowing the analysis of the effects caused by each element of the evaluation to be done individually. In the educational area, facets usually consist of variations of examinees, tasks / items (difficulty), criteria, scoring scales (understanding and use) and evaluators (severity, tendencies to systematic judgments and personal characteristics). In addition, in this scenario there are many factors that can cause variability in the assigned scores, which may compromise the fairness of the evaluation. The objective of this study is to analyze the effects of a test book that in the evaluation considers a large area of concentration, adding several disciplines - the subscales are created to contemplate the disciplines grouped in the blocks - and how these effects can influence the measurement of the performance of the examining in the tests, how they can assist in the pedagogical analysis provided to educators. Thus, with the application of the Many-Facets Rasch Measurement with three facets (examining ability, item difficulty and subscales) in a block of questions of a large scale educational evaluation, it is possible to estimate the effects of the facets in a direct and individual way. In this work the application of the methodology was done in the block of the Natural Sciences and technologies of the ENEM (National High School Examination) 2016 in this notebook are included the disciplines of Physics, Chemistry and Biology. Even the one containing three distinct disciplines is one-dimensional, the measured construct is unique, guaranteed by dimensionality analyzes. The contribution of the MFRM is related to the possibility of practical application of methodology for analysis and estimation of the main effects of the application of this type of test, determining how these traits influence the performance of the examinees, and can also subsidize other areas of evaluation, such as pedagogical analysis that serves as input for educators in improving the quality of the teaching-learning process. The analysis of the response sample was performed in the FACETS software with the application of the three-facets model. The main statistics, MQ-Infit and MQ-Outfit, presented very satisfactory results for all facets studied. These results demonstrate a good fit of the model, in this way the results are favorable for the measurement. The results of the application of the model have quality, pointing to be favorable for their use conjoin with pedagogical analysis.

Keywords: subscales; Many-Facets Rasch Measurement; educational evaluation; pedagogical analysis and quality of education.

LISTA DE ILUSTRAÇÕES

Figuras

Quadros

Tabelas

Gráficos

LISTA DE SIGLAS

ADES - Avaliação Discente da Educação Superior;

ANA – Avaliação Nacional da Alfabetização

ANEB - Avaliação Nacional da Educação Básica;

ANRESC - Avaliação Nacional do Rendimento Escolar (Prova Brasil)

Caed/UFJF – Centro de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora;

CEPAL - *Comisión Económica para América Latina y el Caribe;*

DELE - Diploma de Espanhol como Língua Estrangeira;

ENADE - Exame Nacional de Desempenho dos Estudantes;

ENEM - Exame Nacional do Ensino Médio;

FIES - Fundo de Financiamento Estudantil;

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira;

JLME - *Joint Maximum Likelihood Estimation;*

MEC – Ministério da Educação;

MFRM - Modelo Multifacetas de Rasch (*Many-Facets Rasch Measurement*);

NAEP - *National Assessment Educational Progress;*

PISA - *Programme for International Student Assessment*;

ProUni - Programa Universidade para Todos;

SAEB - Sistema de Avaliação da Educação Básica;

SINAES - Sistema Nacional de Avaliação da Educação Superior;

Sisu - Sistema de Seleção Unificada

TCT - Teoria Clássica dos Testes;

TOEFL - *Test of English as a Foreign Language;*

TRI - Teoria de Resposta ao Item;

UFJF – Universidade Federal de Juiz de Fora;

UnB – Universidade de Brasília.

SUMÁRIO

1. INTRODUÇÃO

Os números sobre educação no Brasil começaram a ser registrados de forma sistemática no século passado. No entanto, apenas o registro das informações sobre educação não garante a sua utilização para implementação de políticas públicas ou de melhorias. Como exemplos de levantamentos de dados educacionais temos o Censo Escolar - levantamento de informações estatístico-educacionais relativas à Educação Básica, em seus diferentes níveis (educação infantil, ensino fundamental e ensino médio) e modalidades (ensino regular, educação especial e educação de jovens e adultos) e o Censo do Ensino Superior - levantamento de dados e informações estatístico-educacionais junto às instituições de ensino superior.

Como afirma Horta Neto (2007), o caminho entre um sistema de medições para levantar dados sobre a educação e a construção de um sistema de avaliação da educação básica no Brasil foi longo. E esse caminho não para de ser construído, pois, um sistema de avaliação é algo complexo, composto de diversas etapas e que exige a seleção de atributos importantes acerca do que será avaliado, com a descrição desses de maneira objetiva e precisa e a síntese das evidências alcançadas.

O processo de avaliação educacional é recente no Brasil, se comparado a outros países da Europa ou aos Estados Unidos. Os primeiros relatos robustos de avaliação em larga escala referenciam um grande levantamento de informações educacionais dos Estados Unidos que gerou o chamado Relatório Coleman[1]. A preocupação com a qualidade da educação e de que maneira se distribuíam as oportunidades educacionais dentro do seu território despertaram o interesse por informações que pudessem explicar essas questões. O contexto político era da ocorrência da Guerra Fria, discussões sobre a democracia no ocidente e sobre a legislação americana acerca da igualdade de direitos civis para negros e brancos.

O governo norte-americano encomendou a pesquisa, que tinha por objetivo conhecer as razões da diferença de disponibilidades educacionais em razão da raça, cor, religião ou naturalidade em instituições públicas. O Relatório Coleman apresentou, no contexto de sua aplicação, que as diferenças de conhecimento encontradas nos resultados dos testes aplicados

[1] O Relatório Coleman desenvolvido nos Estados Unidos na década de 1960, encomendado pelo governo americano após a aprovação da Lei de Direitos Civis. Tinha como objetivo analisar a diferença de atendimento educacional no país, e constatou que a condição socioeconômica era determinante nas diferenças do desempenho estudantil.

deviam-se muito mais a fatores socioeconômicos dos discentes que a outros fatores apresentados para análise, e muitos dos seus resultados foram amplamente divulgados, chegando ao Brasil. Nos anos de 1990, o Brasil alcançou notável progresso tanto na área de avaliação quanto na produção de informação educacional que ainda hoje pode ser visto.

No Brasil, foram postas algumas considerações sobre a importância da avaliação na Portaria 1.795/94, tais como: a necessidade de assegurar uma educação básica de qualidade com equidade e eficiência, em Soares (2005, p.90) "o legislador deixa claro que o ensino deve propiciar ao estudante o domínio de determinados conteúdos. Quando isso ocorre, assume-se que o aluno teve acesso a uma *educação de qualidade*". Para o governo, a verificação dessa educação de qualidade é feita mediante avaliação em larga escala.

O permanente monitoramento de execução e avaliação de resultados das políticas públicas; a necessidade de uma organização sistêmica dos processos de monitoramento e avaliação - envolvendo órgãos governamentais, universidades e centros de pesquisa; a necessidade de que a disseminação das informações geradas pelas avaliações seja de domínio público, de forma a haver um controle social de seus resultados (HORTA NETO, 2007). Esses são apenas alguns exemplos da dimensão que uma avaliação educacional pode abranger como objetivo.

Avaliações como a Prova Brasil (ANRESC - Avaliação Nacional do Rendimento Escolar), Avaliação Nacional da Educação Básica (ANEB) e Avaliação Nacional da Alfabetização (ANA) juntas constituem o Sistema de Avaliação da Educação Básica (SAEB) que tem na sua concepção diretrizes de como essas avaliações devem contribuir para o desenvolvimento de uma cultura de avaliação que favoreça a melhoria da qualidade da educação; aplicar e desenvolver processos permanentes de avaliação; mobilizar recursos para as políticas públicas educacionais; e proporcionar à sociedade informações sobre o desempenho e os resultados dos sistemas avaliados (BRASIL, 2013a) e também o Programa Internacional de Avaliação de Estudantes - *Programme for International Student Assessment* - (PISA) que tem por objetivo produzir indicadores pertinentes a qualidade da educação dos países participantes, subsidiando possíveis políticas públicas de melhoria da educação (OCDE, 2015).

A avaliação educacional em larga escala tem, primeiramente, caráter diagnóstico da educação ofertada no Brasil, com finalidade de acompanhar a evolução da qualidade e da equidade da educação. Outro fator importante dentro do processo de avaliação é a divulgação dos resultados que deve contemplar os principais interessados, ou seja, as equipes escolares, os pais e a comunidade, favorecendo a articulação para a melhoria da educação bem como a

influência da escola na busca por aprimoramento dos métodos de ensino e conteúdos ensinados (BONAMINO & SOUSA, 2012).

O Brasil tem avançado e consolidado os métodos e as técnicas de avaliação (educacional, de sistemas e testes psicológicos) com a introdução de metodologias modernas. Principalmente em avaliações somativas, cuja principal função é, ao final de um ciclo, verificar a existência de ganho de conhecimento, tomando como base competências e habilidades avaliadas. Em contrapartida, as avaliações formativas, que têm como principal objetivo contribuir para uma boa regulação do processo de ensino (ou da formação) houve pouco avanço nas pesquisas e análises. A combinação entre os dois tipos de avaliação é cada vez mais importante para o sucesso do implementação, desenvolvimento e continuidade de políticas públicas (CHUIEIRE, 2008; DA SILVA & GOMES, 2018).

O maior destaque tem sido o ganho alcançado nas análises dos resultados de avaliações em larga escala com questões objetivas de múltipla escolha com 2 ou mais opções. Um exemplo é o que ocorre na Prova Brasil, que faz parte do SAEB, pois ela oferece escalas de proficiências em Língua Portuguesa e em Matemática para os anos 5º e 9º do Ensino Fundamental, comparáveis ao longo de todos esses anos, permitindo assim o acompanhamento da evolução do desempenho dos alunos brasileiros. Tal comparação é possível porque na Prova Brasil é feito o uso de itens comuns entre as séries avaliadas e entre anos de aplicação, permitindo que os alunos de todas as séries e de todos os anos sejam postos em uma mesma escala de proficiência. Para tanto, foi necessário que houvesse uma definição adequada do construto a ser avaliado para o desenvolvimento de uma matriz de referências de qualidade, assim como a construção de itens e provas com garantias psicométricas.

E ainda, visando à melhoria da avaliação, em 2009, o Exame Nacional do Ensino Médio - ENEM também foi reestruturado passando a adotar matriz de referência para cada uma das áreas avaliadas, a elaboração e a correção das provas seguem premissas psicométricas já implementadas em outras avaliações brasileiras.

Com a evolução dos processos de avaliação educacional no Brasil, houve um considerável aprofundamento em estudos que ultrapassam a avaliação educacional, um exemplo são os estudos sobre a relevância de fatores associados e sua influência na explicação do desempenho dos estudantes nos exames padronizados aplicados com questões objetivas e questões discursivas (SOARES NETO; JESUS; KARINO; ANDRADE, 2013).

Na contramão da evolução das metodologias de análise das avaliações em larga escala está o fato de que a educação brasileira vem sofrendo profundas mudanças nas últimas décadas, a expansão das redes de ensino gerou carência de profissionais, que por sua vez afeta o cumprimento do currículo determinado para cada fase do ensino. Prejudicando a realização de avalições com contexto formativo de maneira ampla. Gatti (2003) afirma que além da carência de profissionais falta a qualificação desses e como ponto fundamental, os administradores públicos não têm contemplado a educação com políticas públicas coerentes. As consequências desencadeadas pelos fatos mencionados podem ser detectadas nos resultados das avaliações educacionais aplicadas na Educação Básica e no Ensino Superior.

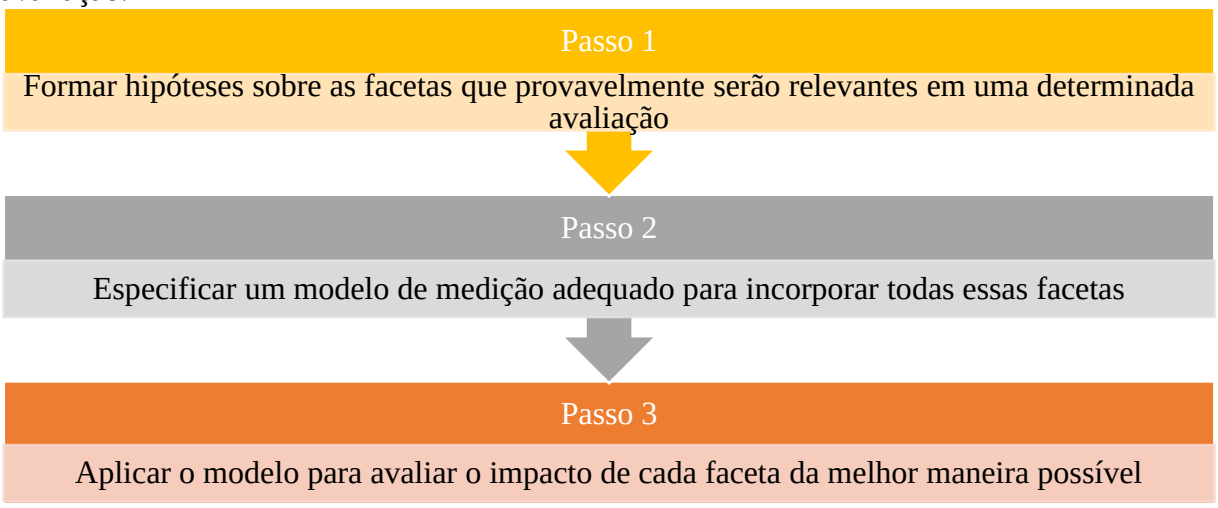
Dentro dessa evolução e aprimoramento dos processos de avaliação educacional, também vale salientar a utilização de uma diversidade de modelos para analisar os fatores que são relevantes e oferecem informações que podem auxiliar na implementação de melhorias para a educação. Uma vez que, os processos avaliativos são caracterizados por conjuntos de fatores que estão envolvidos direta ou indiretamente na obtenção de resultados de medição do desempenho, o desenvolvimento e o aprimoramento de metodologias para seu estudo devem ser objetivos dos pesquisadores da área.

A importância de um estudo que contemple técnicas modernas e consistentes para colaborar com a atividade docente e com a interpretação dos resultados das avaliações deve contribuir para um cenário de mudança e evolução no ensino brasileiro, trazendo consigo uma articulação entre avalição somativa e formativa. A interpretação pedagógica dos resultados traz consigo a possibilidade de melhoria do ensino dos conteúdos na prática, ou seja, em sala de aula. A criação de subescalas, com a utilização do Modelo Multifacetas de Rasch ou *Many-facets Rasch Measurement* (MFRM), nas provas que contemplam vários conteúdos que podem ser interpretados de forma mais específica, é uma contribuição válida para os resultados pedagógicos das avaliações.

Os diversos modelos utilizados para analisar avaliações educacionais têm suas nomenclaturas específicas e de forma geral qualquer fator ou variável que supostamente afete a avaliação e/ou os escores de teste de maneira sistemática podem ser definidos como uma faceta, designação utilizada nos Modelos Multifacetas de Rasch. As facetas que supostamente contribuem com erros sistemáticos de medição criando viés nos resultados estão ligadas principalmente aos avaliadores, às tarefas, aos critérios de correção, aos entrevistadores, ao tempo de teste, entre outros (LINACRE, 1994; ECKES, 2011; TOFFOLI, 2015).

Para Eckes (2011, p.17) a construção de medidas confiáveis, válidas e justas da capacidade do examinando – entenda-se examinando como o sujeito que participa de forma ativa, respondendo ao exame - está ligada à implementação de métodos de análise que possam trabalhar com múltiplas fontes de variabilidade. Do ponto de vista da análise, uma abordagem adequada para dados com múltiplas facetas envolveria três etapas gerais, conforme Figura 1. Essas etapas formam a base metodológica de uma abordagem de medição para a análise e medição de avaliações de desempenho.

Figura 1 – Os três passos básicos para análise e medição do desempenho em processos de avaliação.



Fonte: Adaptado pela autora (Eckes, 2011).

Diante disto, para a contínua evolução das técnicas de análise das avaliações trazemos a proposta da utilização de novos procedimentos estatísticos para testes compostos por avaliações nas quais são utilizados cadernos que agregam muitas disciplinas para uma análise dos itens separados nas respectivas disciplinas, criando assim as subescalas que representam uma das facetas do modelo. Os modelos multifacetas de Rasch são, atualmente, os procedimentos modernos aplicados nas avaliações que têm a necessidade de análises simultâneas de múltiplas variáveis interferentes nos resultados, ou seja, o MFRM incorpora mais parâmetros (facetas) ao modelo; pois, existem diversos fatores que podem interferir na medida de desempenho dos alunos em avaliações que vão além da habilidade ou do conhecimento do examinando, tais como: a dificuldade da tarefa, a ordem da prova, o formato da questão, o tema abordado, os critérios de correção e a escala de pontuação utilizada, entre outras variáveis (LINACRE, 1994; MYFORD & WOLFE, 2000; ECKES, 2011; TOFFOLI, 2015; GOODWIN, 2016).

A utilização de modelos robustos de análise em avaliações deste tipo pode contribuir para a resolução de um dos principais desafios encontrados nas avaliações, a explicação dos fatores que influenciam no desempenho dos alunos e a possibilidade de determinar quantitativamente essa influência. A utilização e o aprimoramento de metodologias como a MFRM ampliam as possibilidades de análise.

Uma vez já conhecido que com a utilização da Teoria de Resposta ao Item (TRI) nas análises de avaliações educacionais é possível comparar testes e indivíduos mesmo eles participando de testes diferentes, a comparabilidade em avaliações com itens objetivos de múltipla escolha é assegurada pela TRI, quando da introdução de itens comuns nas provas aplicadas. A TRI é um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e da habilidade (ou habilidades) do respondente. A TRI qualifica o item de acordo com três parâmetros: a discriminação, a dificuldade e o acerto ao acaso (chute)[2].

Atualmente a TRI é a metodologia mais amplamente difundida e aplicada nas avaliações em larga escala brasileiras.

Na maioria dos casos, as provas são apresentadas com suas questões por disciplina e o fator que pode interferir no desempenho é a ordem de apresentação das disciplinas e das questões. Nesse processo pode surgir a dúvida de que há similaridade das dificuldades tanto quanto às questões quanto às disciplinas. Uma prova que se inicia com matemática e outra com português podem apresentar desempenhos diferentes para seus examinandos. Em testes com cadernos que abrangem áreas de conhecimento, como no ENEM, o ordenamento das questões pode se constituir um fator que cria viés no resultado, o qual poderia ser explicado pela análise das disciplinas envolvidas. Com a aplicação da MFRM, podemos estudar mais eficientemente o viés da influência desses fatores, ajustar o desempenho do aluno em função dos fatores detectados, trazendo, assim justiça ao resultado da medida.

Os pesquisadores estão buscando entender o que acontece em correções de testes com essas características, estudando os fatores, ou seja, as facetas que podem influenciar os resultados da avaliação, com o auxílio de uma metodologia, que vem sendo aplicada em diversos campos de estudo como medicina, educação, psicologia, entre outros (ENGELHARD, 1994; WANG & STAHL, 2012; PARRA-LÓPEZ & OREJA-RODRÍGUEZ, 2014), e que traz consigo maior transparência no processo avaliativo. O principal ganho é assegurar maior justiça

[2] Maiores detalhes podem ser encontrados em Andrade, Tavares & Valle, 2000.

para os alunos avaliados e, indiretamente, para os testes.

No Brasil, ainda são raras pesquisas que analisam os resultados das avaliações em larga escala sob essa perspectiva. Dentre as aplicações, destacam-se as realizadas para analisar a influência do corretor/avaliador na correção de itens abertos, incorporando ao modelo a faceta de severidade do avaliador (TOFOLLI & SIMON, 2018).

Desta maneira, analisar os fatores que influenciam e criam viés no desempenho das avaliações educacionais, principalmente efeitos que podem ser controlados pelos coordenadores da avaliação, aplicando uma metodologia quantitativa e objetiva de análise, tornou-se um desafio para o progresso da área de avaliação educacional.

Com a aplicação do Modelo Multifacetas de Rasch, é possível quantificar o viés criado pela utilização de caderno que abrange várias disciplinas ou assuntos e apresentar o desempenho do aluno de acordo com cada subescala criada, ou seja, cada disciplina, buscando, assim, equalizar o resultado para uma análise pedagógica eficiente. Essa análise detalhada da prova pode viabilizar mudanças na forma de ensinar conteúdos que necessitem maior atenção por parte dos educadores.

Deste modo, as avaliações em larga escala com testes que apresentam agrupamento de disciplinas em áreas ou domínios, com ordenamento de cadernos e itens, têm a necessidade de serem desenvolvidas quanto à análise dos resultados, buscando sempre aprimorar as metodologias, garantindo maior justiça e confiabilidade. Neste contexto, algumas questões e inquietações são colocadas nesta pesquisa:

- É possível construir um diagnóstico sobre os vieses encontrados devido ao agrupamento de disciplinas em áreas ou domínios e qual o efeito disto no desempenho dos examinandos?
- É possível o tratamento dos resultados de testes com cadernos de grandes áreas com a criação de subescalas?
- É possível tornar os resultados das avaliações educacionais mais palatáveis para os educadores, com análises pedagógicas que facilitem o trabalho desenvolvido em sala de aula?

Dessas inquietações emergiu a questão central à qual constituiu a pergunta de pesquisa que o referido estudo buscará responder: Como analisar testes com disciplinas agrupadas em áreas ou domínios, criar modelo que avalie as subescalas de disciplinas e analisar os resultados

para auxiliar o processo de ensino-aprendizagem com informações relevantes para os educadores?

O objetivo geral da pesquisa é analisar as subescalas construídas na área de Ciências da natureza e suas tecnologias da prova do ENEM 2016, estimando, com o Modelo Multifacetas de Rasch, o efeito dessas subescalas nos resultados da avaliação. Para o alcance do objetivo geral, foram definidos os seguintes objetivos específicos:

a) Analisar as facetas do Modelo Multifacetas de Rasch desenhado (item, examinando e subescalas), gerando resultados de ajuste e estatísticas de análise da qualidade dos dados e do modelo;
b) Calcular a proficiência dos alunos utilizando a técnica de Multifacetas de Rasch considerando as subescalas de disciplinas construídas;
c) Propor a utilização dos resultados da metodologia para a análise pedagógica da avaliação.

1.1 Justificativa

Dentre os objetivos das avaliações educacionais em larga escala, estão a mensuração do desempenho individual dos examinandos e a análise diagnóstica de programas e políticas educacionais, sendo responsáveis por nortear decisões importantes, visando melhorias para os atores envolvidos ou sanar eventuais problemas encontrados. Estudar os principais pressupostos que aferem a qualidade de uma prova e a proposição de novos métodos de definição da proficiência dos examinandos é um assunto que sempre pode ser abordado e desenvolvido nas pesquisas relacionadas às técnicas de análise dos processos. Nessas avaliações, são muitos os fatores que podem afetar a proficiência, ou seja, o resultado do desempenho dos examinandos ao responder esses itens.

A busca por metodologias que permitem melhorar o processo de avaliação, com resultados que ampliem as possibilidades das técnicas de ensino-aprendizagem dentro de sala de aula justifica um trabalho como apresentado nesta dissertação. Trazer para a linguagem do educador as análises dos resultados das avaliações educacionais e facilitar o entendimento e a aplicação prática dentro do contexto da análise pedagógica tornam-se um dos principais

objetivos das medidas nas avaliações para melhorar a qualidade do ensino oferecido. Esses atores (diretores, coordenadores educacionais, professores, entre outros) apresentam certa dificuldade em lidar, analisar e interpretar os dados pedagogicamente (GATTI, 2009). Ainda segundo Gatti (2009), o grande desafio é a apropriação por parte das escolas dos resultados e sua utilização para orientar as atividades de ensino.

A relevância deste trabalho reside na aplicação prática de mecanismos modernos que auxiliem nas análises das avaliações com cadernos de múltiplas disciplinas, nas quais a definição da proficiência considera como se todos os itens do caderno avaliassem um único construto. A necessidade da elaboração de modelos práticos e objetivos, como o Modelo Multifacetas de Rasch, que possam ser aplicados de forma transparente no processo avaliativo, justifica o estudo, dada a importância da avaliação tanto para um examinando quanto para a sociedade.

2. AVALIAÇÃO EDUCACIONAL

O século XX foi marcado por grandes mudanças políticas, tecnológicas e econômicas que refletiram na organização e no papel dos Estados-Nação ao redor do mundo. O desenvolvimento e crescimento gerado também trouxeram desigualdades, pois o progresso não ocorreu da mesma forma para todos, sejam eles países, estados ou até mesmo pessoas. Tais mudanças levaram também a um questionamento sobre a função da escola para a sociedade, a qualidade de ensino oferecida à população e todos os processos educacionais empregados pelos gestores (BAUER, 2017).

Após a Revolução Industrial o mundo passou por profundas modificações e em muitos países surgiu a preocupação com a criação de programas sociais de diferentes áreas, incluindo a educação. Nos Estados Unidos, os primeiros registros de testes avaliativos estruturados são datados do final do século XIX, sua metodologia quantitativa e o caráter experimental se mostraram como uma grande contribuição para a área de avaliação. Muitos procedimentos da época ainda hoje são empregados. O principal propósito no início da criação da avaliação educacional era a busca por eficiência na educação e o alcance de uma educação de qualidade (VIANNA, 2014).

Diante disso, aumentou o interesse de se avaliar os sistemas envolvidos no processo de educação, transformando a avaliação em larga escala em um ponto de destaque na criação de políticas públicas educacionais em vários países, a fim de superar a crise instalada no Estado, diminuir a desigualdade e a exclusão social e possibilitar o crescimento econômico interno mais justo.

Nos Estados Unidos, destacam-se os estudos de Ralph W. Tyler que foram disseminados em todo o mundo, que segundo críticos da época suas análises estariam limitando os currículos e desencadeando falta de autonomia das escolas e dos professores para definir conteúdos e abordagens adequadas ao contexto dos alunos (HORTA NETO, 2007). Na Inglaterra, importantes pesquisas foram feitas, os trabalhos realizados por F. Galton, K. Pearson, C. Spearman e C. Burt, entre outros, contribuíram para que a psicometria tivesse influência considerável na avaliação educacional, também se destacaram os estudos avaliativos dos Projetos *Nuffield* e estudos que interligaram análises qualitativas e quantitativas (VIANNA, 2014).

Ainda na década de 70, surge em diferentes partes do mundo outro conceito importante para a área de educação e avaliação educacional, o termo *accountability* - processo que visa ajudar os atores envolvidos no processo educacional a cumprir responsabilidades e alcançar metas. Nesse processo, indivíduos e/ou instituições são obrigados, com base em uma justificativa legal, política, social ou moral, a fornecer uma explicação de como eles cumpriram responsabilidades claramente definidas (UNESCO, 2017).

O termo *accountability* apresenta-se como forma de responsabilização em educação na tentativa de evitar perdas de investimentos financeiros com políticas públicas feitos em programas curriculares nas escolas e nas avaliações ligadas a eles. Paralelamente, também se conectam os conceitos de avaliação e de busca por qualidade do ensino.

Na década de 1990, nos Estados Unidos, teve início a política de *School accountability* que foi adotada por vários estados americanos e apresentava os seguintes parâmetros: (i) estabelecimento de padrões educacionais mínimos para cada ano escolar; (ii) realização de testes de proficiência para averiguar os conhecimentos adquiridos pelos alunos; (iii) tornar público os resultados das escolas nestes testes; (iv) adotar como objetivo explícito de política a melhoria no desempenho dos estudantes nos testes; (v) responsabilizar os professores/diretores da escola pelo resultado dos alunos (ANDRADE, 2009).

Já na Europa, alguns países adotaram e estão consolidando mecanismos de *accountability* no setor educacional, esse termo passou a ser recorrente nos textos oficiais orientadores do governo (AFONSO, 2012). No Brasil, não existe uma política nacional vigente de responsabilização, como a implantada nos Estados Unidos, pois em 1990 o governo federal implantou os pontos (i) e (ii), ou seja, estabeleceu-se um sistema de avaliação nacional que possibilitou a criação de padrões desejáveis de conhecimento para cada ano escolar. E, somente em 2006, houve a implantação do ponto (iii), com a divulgação dos resultados por escola (ANDRADE, 2009).

Para Gatti (2009), o impacto das avaliações começa a ser sentido na educação básica e espera-se que elas sejam vistas como estímulos às mudanças em processos educacionais, e, não como forma de punição. O entendimento dessa cultura de responsabilização ou de *accountability* deve ser apropriado pelos atores envolvidos na análise pedagógica, com a possibilidade de implementar políticas que busquem a melhoria da qualidade do processo de ensino-aprendizagem. Isso é possível com estudos e metodologias de análise que aproximem os resultados das avaliações à realidade dos educadores, daí a importância de integrar avaliações de caráter somativo com avaliações formativas.

Com a crescente utilização de *accountability* em todos os níveis do sistema de ensino, faz-se necessária a melhoria das informações prestadas aos educadores, que fazem parte de um grupo grande e que é atingindo pelas políticas de responsabilização. Diante disso, a proposição da introdução de análises pedagógicas das avaliações com objetivo de suportar escolas e educadores com relatórios consistentes e claros que possam detectar oportunidades de melhorias e trazer mudanças positivas ao processo de ensino em sala de aula é cada vez mais um ponto a ser discutido.

Outro fato importante ocorrido nos anos 1960 é a criação do *National Assessment Educational Progress* (NAEP) nos Estados Unidos o qual desde 1969 tem sido periodicamente realizado. Nessa mesma época, conforme anteriormente citado, teve início o primeiro grande levantamento educacional em larga escala que deu origem ao chamado Relatório Coleman.

O desenvolvimento da área de avaliação educacional no Brasil iniciou-se no século passado, muito influenciado pelas práticas e metodologias utilizadas na Europa e Estados Unidos. A evolução do processo avaliativo da educação brasileira ainda é recente, situar historicamente o desenvolvimento da avaliação brasileira colabora com o entendimento de questões relevantes ao estudo, a seguir uma seção destinada aos fatos que delimitam a criação e o desenvolvimento da avaliação educacional no Brasil.

2.1 Avaliação educacional em larga escala no Brasil e os principais instrumentos

Conforme mencionado, é recente o desenvolvimento da avaliação educacional no Brasil, principalmente quando ligada à análise e medição da qualidade e do desempenho de sistemas educacionais. Uma das razões foi a escassez de especialistas na área, pois a formação de especialistas se deu, primeiramente, por meio de funcionários dos diversos níveis da administração e também de profissionais formados no exterior, sendo que os pesquisadores em avaliação educacional se concentraram em centros de pesquisa e em universidades. A não existência de cursos específicos na área de educação é um fato que dificulta o desenvolvimento da área, o que há são poucos casos de pós-graduação (VIANNA, 2003; GATTI, 2014).

Os primeiros passos da avaliação eram voltados para o acesso ao ensino superior como forma de seleção de estudantes, não era foco a avaliação de desempenho escolar voltada à

análise de sistemas escolares. Para Gatti (2014, p. 11) “foi nesse movimento que alguns profissionais começaram a receber formação mais aprofundada na área de avaliação de rendimento escolar, vinculada à teoria da medida e aos conhecimentos sobre elaboração de testes objetivos, sua validade e fidedignidade”.

O começo da formação da base de conhecimentos na área de avaliação educacional no Brasil retratou dois panoramas de trabalho, foi quando surgiram: pesquisas avaliativas que tratavam do desempenho escolar de alunos e também trabalhos de avaliação de políticas e programas educacionais. Brotava tanto o interesse por avaliação baseada em modelos específicos quanto a formação de pesquisadores avaliadores (GATTI, 2014).

Porém, a preocupação em nível nacional se inicia ainda de forma insipiente a partir do final da década de 1980, com o desenvolvimento de estudos exploratórios. A avaliação em larga escala de abrangência nacional se desdobra em múltiplas modalidades na década de 1990. Houve a implementação efetiva em 1993 do Sistema Nacional de Avaliação da Educação Básica (SAEB), que é uma avaliação focada em competências em leitura e matemática (BRASIL, 2013a; BRASIL, 2013b).

O SAEB sofreu reestruturações e atualmente escolas públicas e privadas do Ensino Fundamental e de Ensino Médio, passaram a ter resultados no SAEB. E também, foi criado em 1998 outro instrumento, o ENEM, instituído com o objetivo de verificar o comportamento de saída dos alunos do ensino médio. Este último, posteriormente, foi além da avaliação do Ensino Médio quando passou a ser utilizado por instituições de Ensino Superior como critério de ingresso em 2009 (COELHO, 2008; WERLE, 2011; GATTI, 2014).

De modo a consolidar os processos avaliativos no Brasil, no final do ano de 1996, é promulgada a Lei Nº. 9.394 (BRASIL, 1996) que reafirma a importância do papel da avaliação externa e a coloca como imprescindível, bem como a sua universalização. Nesta lei está explícito que os entes federativos devem integrar todos os estabelecimentos de ensino fundamental ao sistema nacional de avaliação do rendimento escolar. Para Werle (2011) o sistema de avaliação estava se desenvolvendo ao longo da década de 1990, com o advindo de leis e, paralelamente, os mecanismos de financiamento se consolidavam na legislação educacional, contribuindo para a sua concretização.

Nos primeiros anos do século XXI, com o crescimento de demandas em relação à avaliação do processo de alfabetização nos anos iniciais do Ensino Fundamental, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) buscou disponibilizar

instrumentos de avaliação, com o objetivo de subsidiar a obtenção de dados sobre alfabetização, leitura e escrita, qualidade do ensino e das escolas, desde as etapas mais precoces da escolarização (OLIVEIRA & ROCHA, 2007).

Já no ano de 2005, o desenvolvimento da avaliação em larga escala contou com a criação da Prova Brasil, uma avaliação censitária, visando à obtenção de dados mais detalhados sobre a realidade educacional nacional, e que tem por objetivo avaliar em Língua Portuguesa e Matemática, alunos da 4ª e 8ª séries do Ensino Fundamental (respectivamente, 5º e 9º anos de escolaridade) contemplando as redes de ensino que ampliaram o Ensino Fundamental para 9 anos.

Por enquanto, tratou-se de avaliações educacionais para os ensinos básico, mas o Ensino Superior também tem histórico de programas de avaliação. Atualmente, para o Ensino Superior existe uma avaliação em larga escala, o Exame Nacional de Desempenho dos Estudantes (ENADE). Ele faz parte do Sistema Nacional de Avaliação da Educação Superior (SINAES) e foi estabelecido pela Lei 10.861/04 de 14 de abril de 2004 no seu "Art. 5 - A avaliação do desempenho dos estudantes dos cursos de graduação será realizada mediante aplicação do Exame Nacional de Desempenho dos Estudantes - ENADE" (BRASIL, 2004).

O ENADE é composto por uma prova, pelo questionário de Avaliação Discente da Educação Superior (ADES) (antigo questionário socioeconômico), o questionário dos coordenadores de curso e pela percepção do aluno sobre a prova. O objetivo da prova é de aferir as habilidades acadêmicas, as competências profissionais básicas das áreas, o conhecimento sobre conteúdos básicos e profissionalizantes, além de questões transdisciplinares (DE BRITO, 2008, p. 842).

O desenvolvimento da questão avaliativa no Brasil tem focado, principalmente, em avaliações somativas, que vinculam-se à disposição pelos responsáveis de continuar, alterar ou até mesmo de encerrar um programa. A parte formativa, que ocorrem ao logo da evolução de um programa, fornecendo informações relevantes aos responsáveis pela implementação do objeto da avaliação é, por vezes, negligenciada. Sendo necessário a combinação dos dois tipos com objetivo de melhor entender a qualidade da educação ofertada, viabilizando informações úteis para serem usadas em sala de aula e por fim, propiciando papel de destaque aos educadores, que poderão utilizar tais recursos para melhorar o processo de ensino-aprendizagem.

Com os avanços, nos últimos anos, dos estudos na área de avaliação educacional em larga escala, os processos que têm seus resultados analisados com a metodologia da Teoria de Resposta ao Item se consolidaram. E surgiram outras inquietações dos pesquisadores quanto aos modelos utilizados nos últimos anos, levantando a questão de que é preciso cada vez mais aprimorar e diversificar para acrescentar mais informação aos resultados apresentados, quanto ao seu conteúdo, formato e finalidade.

Como parte do estudo foi necessária a apropriação de base de dados de uma avaliação educacional em larga escala. O estudo utilizou os dados do ENEM 2016. Diante disto, cabe a apresentação das características desta avaliação e alguns números sobre a aplicação, a seguir serão elencados fatores históricos, descrição dos objetivos e especificidades dos resultados do bloco de Ciências da natureza e suas tecnologias.

2.1.1 O objeto do estudo: Exame Nacional do Ensino Médio - ENEM

Historicamente, a partir da década de 1970, os países capitalistas mobilizam-se para restabelecer o desenvolvimento econômico, seguindo uma postura voltada para as premissas básicas do liberalismo, ou seja, a livre concorrência e a mínima intervenção do Estado. Essa reconfiguração do Estado, a partir das últimas décadas do século XX no Brasil, apresentou como uma das principais consequências, o fato de evidenciar a função de avaliar os serviços oferecidos pelas instituições ligadas ou mantidas pelo Estado Brasileiro.

Diante disto, foi a partir dos anos de 1990, que tornou pungente a busca por se alcançar equidade, qualidade e eficiência na educação brasileira. Em 1992, a *Comisión Económica para América Latina y el Caribe* (CEPAL) produziu um importante documento que apresentava diretrizes para políticas que favorecessem a educação, o conhecimento e o desenvolvimento nos países da América Latina e Caribe, o *Educación y conocimiento: eje de la transformación productiva com equidade,* em parceria com a UNESCO. Nesse documento estava destacada a necessidade de reformas administrativas que possibilitassem reformular o papel do Estado, que passaria de uma postura de administrador e provedor à função de Estado avaliador (MACHADO & LIMA, 2014).

Para Machado e Lima (2014) outro fato é que em 1995, no governo de Fernando Henrique Cardoso, sob os efeitos de uma crise internacional, foi proposta uma reforma do Estado Brasileiro, na qual a educação era apresentada como essencial para o desenvolvimento.

Já o tema avaliação aparecia como parte de uma administração pública gerencial. A avaliação exigida pela reforma do Estado em 1995 era um reflexo das recomendações das agências multilaterais para os países em desenvolvimento, como o Banco Mundial, que vincula as suas linhas de crédito à necessidade da existência de indicadores que possam sinalizar a melhoria na qualidade da educação. A avaliação, neste contexto de reforma do Estado é encontrada no artigo 9º, parágrafo VI, da LDB 9.394/96, que diz ser incumbência do Estado: *assegurar processo nacional de avaliação do rendimento escolar no ensino fundamental, médio e superior, em colaboração com os sistemas de ensino, objetivando a definição de prioridades e a melhoria da qualidade do ensino* (BRASIL, 1996).

É nesse contexto que surge o Exame Nacional do Ensino Médio (ENEM), criado em 1998 pelo governo federal, mais especificamente pelo INEP, do Ministério da Educação (MEC). Até 2008, o ENEM era um exame de caráter individual e voluntário, sendo ofertado anualmente aos alunos concluintes do ensino médio do ano de realização e também aos concluintes de anos anteriores. Era voltado para as competências e habilidades que o estruturavam. O Exame era formado por uma prova de 63 questões objetivas e uma redação que tinham como objetivo avaliar "as competências e habilidades desenvolvidas pelos participantes ao longo da escolaridade básica, a partir de uma Matriz de Competências especialmente desenvolvida para estruturar o exame" (BRASIL, 2002, p. 9).

Ao longo da primeira década do século XXI, o ENEM foi utilizado como instrumento de certificação de conclusão do Ensino Médio, bem como de critério de acesso ao Ensino Superior, através de programas como o Programa Universidade para Todos (ProUni) e o Fundo de Financiamento Estudantil (FIES).

Em 2009, medidas governamentais estimularam o uso do ENEM como forma de acesso ao Ensino Superior no Brasil. O Sistema de Seleção Unificada (Sisu) passou a operar em larga escala no processo de alocação dos candidatos às vagas e ser aceito como exame de seleção para a maioria das universidades federais do Brasil.

O novo ENEM manteve-se como uma avaliação individual, mas estruturada a partir de uma matriz de referência própria (BRASIL, 2009). O modelo de avaliação é composto de uma redação dissertativa acerca de temas atuais e socialmente relevantes, em conjunto com quatro blocos: Linguagens e códigos e suas tecnologias, Ciências da natureza e suas tecnologias, Ciências humanas e suas tecnologias e Matemática e suas tecnologias, de 45 questões objetivas cada. A partir de 2009, cada uma das quatro provas objetivas passou a contar com uma matriz de referência e uma escala própria.

Quadro 1: Quadro relacional das áreas do conhecimento e componentes curriculares das provas do Enem.

Áreas de Conhecimento	**Componentes Curriculares**
Ciências Humanas e suas Tecnologias (CH)	História, Geografia, Filosofia e Sociologia
Ciências da Natureza e suas Tecnologias (CN)	Química, Física e Biologia
Linguagens, Códigos e suas Tecnologias (LC) e Redação (RD)	Língua Portuguesa, Literatura, Língua Estrangeira (Inglês ou Espanhol), Artes, Educação Física e Tecnologias da Informação e Comunicação
Matemática e suas Tecnologias (MT)	Matemática

Fonte: BRASIL, 2009.

O novo ENEM ainda é pautado em competências e habilidades (STADLER & HUSSEIN, 2017). O sucesso efetivo deste sistema depende de formulação eficiente das provas, apresentando questões consistentes com a avaliação das habilidades e competências recomendadas para o Ensino Médio, em alinhamento com as diretrizes curriculares.

Nos editais do novo ENEM o INEP elenca as atuais finalidades dos resultados das provas. Que são elas: compor a avaliação de medição da qualidade do Ensino Médio no País; subsidiar a implementação de políticas públicas; criar referência nacional para o aperfeiçoamento dos currículos do Ensino Médio; desenvolver estudos e indicadores sobre a educação brasileira; estabelecer critérios de acesso do participante a programas governamentais; e, constituir parâmetros para a autoavaliação do participante, com vista à continuidade de sua formação e à sua inserção no mercado de trabalho.

O estudo está em consonância às finalidades estabelecidas para o novo ENEM no que se refere ao desenvolvimento de estudos sobre a educação brasileira e a criação de referência para o aperfeiçoamento do currículo do Ensino Médio. O desenvolvimento de metodologias de análise que possibilitem o melhor aproveitamento dos resultados, gerando relatórios com informações para os diversos atores envolvidos no processo educacional é parte importante da avaliação educacional.

Para a área Ciências da natureza e suas tecnologias (CN), foco do estudo e das análises apresentadas, em que estão inseridas as disciplinas de Química, Física e Biologia, a Matriz de Referência elenca oito competências que englobam, no total, 30 habilidades, evidenciando a

possibilidade de contextualização e interdisciplinaridade entre as disciplinas envolvidas, uma vez que 19 habilidades são consideradas não específicas para nenhuma delas. As outras habilidades estão divididas de acordo com a disciplina, constituindo-se a parte específica da prova (STADLER & HUSSEIN, 2017).

Quadro 2 - Matriz de Referência de Ciências da Natureza e suas Tecnologias.

Competência de área 1 – Compreender as ciências naturais e as tecnologias a elas associadas como construções humanas, percebendo seus papéis nos processos de produção e no desenvolvimento econômico e social da humanidade.	
H1	Reconhecer características ou propriedades de fenômenos ondulatórios ou oscilatórios, relacionando-os a seus usos em diferentes contextos.
H2	Associar a solução de problemas de comunicação, transporte, saúde ou outro, com o correspondente desenvolvimento científico e tecnológico.
H3	Confrontar interpretações científicas com interpretações baseadas no senso comum, ao longo do tempo ou em diferentes culturas.
H4	Avaliar propostas de intervenção no ambiente, considerando a qualidade da vida humana ou medidas de conservação, recuperação ou utilização sustentável da biodiversidade.
Competência de área 2 – Identificar a presença e aplicar as tecnologias associadas às ciências naturais em diferentes contextos.	
H5	Dimensionar circuitos ou dispositivos elétricos de uso cotidiano.
H6	Relacionar informações para compreender manuais de instalação ou utilização de aparelhos, ou sistemas tecnológicos de uso comum.
H7	Selecionar testes de controle, parâmetros ou critérios para a comparação de materiais e produtos, tendo em vista a defesa do consumidor, a saúde do trabalhador ou a qualidade de vida.
Competência de área 3 – Associar intervenções que resultam em degradação ou conservação ambiental a processos produtivos e sociais e a instrumentos ou ações científico-tecnológicos.	
H8	Identificar etapas em processos de obtenção, transformação, utilização ou reciclagem de recursos naturais, energéticos ou matérias-primas, considerando processos biológicos, químicos ou físicos neles envolvidos.
H9	Compreender a importância dos ciclos biogeoquímicos ou do fluxo energia para a vida, ou da ação de agentes ou fenômenos que podem causar alterações nesses processos.
H10	Analisar perturbações ambientais, identificando fontes, transporte e(ou) destino dos poluentes ou prevendo efeitos em sistemas naturais, produtivos ou sociais.
H11	Reconhecer benefícios, limitações e aspectos éticos da biotecnologia, considerando estruturas e processos biológicos envolvidos em produtos biotecnológicos.
H12	Avaliar impactos em ambientes naturais decorrentes de atividades sociais ou econômicas, considerando interesses contraditórios.
Competência de área 4 – Compreender interações entre organismos e ambiente, em particular aquelas relacionadas à saúde humana, relacionando conhecimentos científicos, aspectos culturais e características individuais.	
H13	Reconhecer mecanismos de transmissão da vida, prevendo ou explicando a manifestação de características dos seres vivos.

H14	Identificar padrões em fenômenos e processos vitais dos organismos, como manutenção do equilíbrio interno, defesa, relações com o ambiente, sexualidade, entre outros.
H15	Interpretar modelos e experimentos para explicar fenômenos ou processos biológicos em qualquer nível de organização dos sistemas biológicos.
H16	Compreender o papel da evolução na produção de padrões, processos biológicos ou na organização taxonômica dos seres vivos.
Competência de área 5 – Entender métodos e procedimentos próprios das ciências naturais e aplicá-los em diferentes contextos.	
H17	Relacionar informações apresentadas em diferentes formas de linguagem e representação usadas nas ciências físicas, químicas ou biológicas, como texto discursivo, gráficos, tabelas, relações matemáticas ou linguagem simbólica.
H18	Relacionar propriedades físicas, químicas ou biológicas de produtos, sistemas ou procedimentos tecnológicos às finalidades a que se destinam.
H19	Avaliar métodos, processos ou procedimentos das ciências naturais que contribuam para diagnosticar ou solucionar problemas de ordem social, econômica ou ambiental.
Competência de área 6 – Apropriar-se de conhecimentos da física para, em situações problema, interpretar, avaliar ou planejar intervenções científicotecnológicas.	
H20	Caracterizar causas ou efeitos dos movimentos de partículas, substâncias, objetos ou corpos celestes.
H21	Utilizar leis físicas e (ou) químicas para interpretar processos naturais ou tecnológicos inseridos no contexto da termodinâmica e(ou) do eletromagnetismo.
H22	Compreender fenômenos decorrentes da interação entre a radiação e a matéria em suas manifestações em processos naturais ou tecnológicos, ou em suas implicações biológicas, sociais, econômicas ou ambientais.
H23	Avaliar possibilidades de geração, uso ou transformação de energia em ambientes específicos, considerando implicações éticas, ambientais, sociais e/ou econômicas.
Competência de área 7 – Apropriar-se de conhecimentos da química para, em situações problema, interpretar, avaliar ou planejar intervenções científicotecnológicas.	
H24	Utilizar códigos e nomenclatura da química para caracterizar materiais, substâncias ou transformações químicas.
H25	Caracterizar materiais ou substâncias, identificando etapas, rendimentos ou implicações biológicas, sociais, econômicas ou ambientais de sua obtenção ou produção.
H26	Avaliar implicações sociais, ambientais e/ou econômicas na produção ou no consumo de recursos energéticos ou minerais, identificando transformações químicas ou de energia envolvidas nesses processos.
H27	Avaliar propostas de intervenção no meio ambiente aplicando conhecimentos químicos, observando riscos ou benefícios.
Competência de área 8 – Apropriar-se de conhecimentos da biologia para, em situações problema, interpretar, avaliar ou planejar intervenções científicotecnológicas.	
H28	Associar características adaptativas dos organismos com seu modo de vida ou com seus limites de distribuição em diferentes ambientes, em especial em ambientes brasileiros.
H29	Interpretar experimentos ou técnicas que utilizam seres vivos, analisando implicações para o ambiente, a saúde, a produção de alimentos, matérias primas ou produtos industriais.

H30	Avaliar propostas de alcance individual ou coletivo, identificando aquelas que visam à preservação e a implementação da saúde individual, coletiva ou do ambiente.

Fonte: BRASIL, 2009.

2.2 Características importantes de uma Avaliação Educacional

As avaliações em larga escala necessitam atender a certos requisitos para que seus dados e resultados sejam validados e possam ser considerados quando da análise ou implementação de políticas educacionais.

Existem requisitos como: (i) validade – capacidade de diferenciação dos construtos que se pretende medir pelo instrumento criado; (ii) confiabilidade – consistência ou estabilidade de uma medida; (iii) comparabilidade – processo de ajustar a estimativa dos parâmetros dos indivíduos para uma escala única; (iv) justiça – busca por equidade no teste, que podem ser medidos por estatísticas específicas, que não se resumem apenas a princípios de medição; são valores sociais com significado, e devem ser considerados sempre que decisões de valores são tomadas com base nas avaliações (MESSICK, 1989).

As avaliações em larga escala tornaram-se instrumentos importantes que subsidiam os gestores na criação, manutenção e reformulação de políticas públicas, dada sua importância é preciso garantir que os aspectos relevantes, para o processo que é avaliar, sejam garantidos. Alguns desses aspectos são definidos a seguir. Vale ressaltar que existem grupos de pesquisadores que tratam esses conceitos, já citados, de formas distintas. As definições apresentadas neste estudo são as mais aderentes ao método de análise que é proposto para o atingimento do objetivo deste trabalho.

A definição de validade sofreu transformações ao longo dos anos, as mais relevantes e citadas por pesquisadores dizem respeito à validade de construto, que se refere ao grau que o instrumento criado é capaz de diferenciar os construtos que ele se propõe a medir. A validade de conteúdo demonstra se a amostra do conteúdo abordado no teste é relevante e representativa. A validade de critério de um teste consiste no grau de eficácia que ele tem em predizer um desempenho específico de um sujeito (PASQUALI, 2009).

A medida de confiabilidade está associada à premissa se é possível medir instrumento aplicado em uma avaliação de forma consistente (YANCEY, 1999 *apud* TOFFOLI, 2015). Neste sentido, independentemente do momento em que o indivíduo respondeu a um teste ele deve alcançar o mesmo resultado, de forma similar as pontuações registradas por dois avaliadores a uma mesma questão não devem ser excessivamente diferentes entre si. Para Martins (2006, p. 2), a confiabilidade de um instrumento para coleta de dados, teste, técnica de aferição é sua coerência, determinada através da constância dos resultados. Ela refere-se à consistência ou estabilidade de uma medida.

Alguns índices podem ser calculados para medir a confiabilidade, por exemplo, o alfa de Cronbach, o Coeficiente KR-20 e também os coeficientes de correlação de Pearson, Spearman e Kendall no caso de medições, julgamentos ou correções feitas por avaliadores.

Confiabilidade e validade são requisitos que devem se aplicar em diversas situações, nas medições de testes, nos instrumentos de coletas de dados e técnicas de aferição. Os dois requisitos são importantes, mas a relação entre eles não é de suficiência, ou seja, nem todo instrumento que apresenta confiabilidade é necessariamente válido (GOLAFSHANI, 2003; MARTINS, 2006).

A comparabilidade também é dos elementos essenciais nas avaliações, pois sem garantias de equivalência entre diferentes instrumentos e resultados que são usados para ajudar a aprendizagem futura, a base da tomada de decisão dos formuladores de políticas públicas é falha (TATTERSALL, 2007).

A comparação entre os instrumentos e os participantes de testes diferentes é uma característica ambicionada pelos pesquisadores da área da avaliação. Busca-se a equalização, que é o processo de ajustar a estimativa dos parâmetros dos indivíduos para uma escala única. Nesse processo, a TRI apresenta muitos benefícios se comparada a outras técnicas de análise, pois a equalização é um dos fundamentos centrais da teoria.

Vale lembrar, a comparação é importante de acordo com o objetivo da avaliação, se o objetivo da avaliação é acompanhar a evolução de um quesito, como qualidade, ao longo do tempo, ou verificar na mesma escala a evolução dos examinandos de quaisquer séries, a comparabilidade dos instrumentos é imprescindível. No caso de avaliações que têm como objetivo a classificação dos examinandos, ou seja, a formação de um *ranking* para selecionar candidatos, a comparabilidade do teste pode ser uma premissa dispensável.

Com a equivalência entre as provas garantida pela TRI, consequentemente também é assegurada a justiça quanto ao escore dos participantes e à classificação dos mesmos. O mais comum em avaliações objetivas, que asseguram a comparabilidade, consiste em manter itens comuns de uma prova para a outra. As questões relacionadas à justiça versam sobre o conceito de equidade do teste, ou seja, a possibilidade de que todos os participantes tenham oportunidades iguais, e, também garantindo a imparcialidade de modo que permitam inferências corretas do desempenho em relação ao construto avaliado.

A objetividade do modelo Rasch garante a apresentação dos resultados de forma direta, tornando-os de fácil explicação. Logo, o modelo Rasch pode ser mais justo a depender do objetivo da avaliação, ser mais fácil de se entender seu resultado e a comparação pode ser realizada diretamente.

Essa busca por um teste não tendencioso que garanta condições similares aos participantes, passa por dificuldades que vão além de uma avaliação por meio de testes, elas vêm das condições sociais, econômicas e educacionais às quais os avaliados estão submetidos.

Para Karino (2016) um sistema igualitário, garante aprendizado a todos, sem distinção, permitindo que todos alcancem os níveis adequados esperados. Sendo necessário que seja dada a mesma oportunidade de acesso a recursos e processos escolares. O termo equidade se apresenta relacionado ao senso de justiça e à inclusão.

3. PROCEDIMENTOS METODOLÓGICOS

Este estudo utilizou um instrumento de avaliação com itens dicotômicos que possibilitaram a avaliação do traço latente, que é a habilidade do examinando considerando subescalas construídas a partir do caderno de Ciências da natureza e suas tecnologias, prova aplicada no ENEM 2016, por meio de ferramentas estatísticas para análise dos dados, de tal forma que se caracteriza como uma pesquisa de abordagem quantitativa.

Esta pesquisa tem como objetivo o detalhamento das proficiências em ciências dos examinandos, por meio de uma análise que considera as disciplinas contidas no caderno, bem como a apresentação dos resultados e parâmetros das facetas elencadas no modelo, sendo considerada uma pesquisa que fez uso da aplicação de metodologias quantitativas, gerando assim conhecimentos que podem ser aplicados em avaliações desta natureza.

Outra característica deste estudo é a dupla composição, o trabalho pode ser dividido em duas partes: uma teórica – com levantamento de referências bibliográficas; e uma prática – na qual foi utilizada base de dados amostral de uma avaliação em larga escala para aplicação do Modelo Multifacetas de Rasch.

3.1 Os Modelos de Rasch

As bases do modelo Rasch foram inicialmente publicadas por Georg Rasch no livro *Probabilistic models for some intelligence and attainment tests* de 1960. O modelo proposto trouxe uma nova perspectiva, diferente das bases da Teoria Clássica dos Testes (TCT), propondo novos métodos para o desenvolvimento e a análise dos testes. Nela, os itens não dependeriam mais da amostra e a habilidade não dependeria dos itens aplicados (ANDRICH, 1988; BAKER, 2001; CHACHAMOVICH, 2007).

O modelo Rasch contribuiu com o entendimento de que a probabilidade maior de acertar determinado item pertence a um aluno que possui maior habilidade em comparação a um aluno de menor habilidade. E quanto ao item, os de menor dificuldade devem ter maior probabilidade de acerto por um aluno de habilidade θ, que os itens com maior dificuldade.

Como resultado do modelo Rasch temos o produto entre a habilidade do aluno e a dificuldade do item.

O modelo assume a existência de um traço latente unidimensional, baseado no padrão ideal de respostas que considera a relação entre habilidade do aluno e dificuldade do item. No entanto, existem fatores aleatórios que podem influenciar os resultados e que não seguem ao padrão determinístico, o modelo verifica a adequação dos dados observados considerando as probabilidades.

O modelo Rasch por incluir apenas um parâmetro é bastante utilizado na Teoria de Resposta ao Item como o modelo de 1 parâmetro, a variável dependente é a probabilidade de acerto e as variáveis independentes são o traço latente da habilidade do aluno (θ) e a dificuldade do item (b). As variáveis independentes são combinadas, ou seja, a dificuldade do item é subtraída da habilidade do aluno. O modelo Rasch foi bastante difundido devido a sua parcimônia de medida e da simplicidade de sua lógica, ele é um modelo objetivo. Esse modelo é representado pela função logística de um parâmetro, a qual considera que as respostas de um sujeito a um conjunto de itens dependem apenas de sua habilidade e da dificuldade dos respectivos itens (BAKER, 2001; LINACRE & WRIGHT, 2002; COUTO & PRIMI, 2011).

$$\ln\left[\frac{P_{is}}{(1-P_{is})}\right] = \theta_s - b_i \qquad (1)$$

Onde,

P_{is} – Probabilidade do sujeito s acertar o item i
$1 - P_{is}$ - Probabilidade do sujeito s não acertar o item i
θ_s - habilidade do sujeito s
b_i – dificuldade do item i

O modelo é simples e conforme a organização de suas variáveis pode ser interpretado de tal forma que, conhecendo a dificuldade do item e a habilidade do sujeito, é possível predizer qual é a probabilidade desse sujeito acertar o item. O valor de b é dado pelo valor de θ no qual a probabilidade de acertar o item é de 50% (COUTO & PRIMI, 2011).

Como o modelo de Rasch estima de forma independente a habilidade do sujeito e a dificuldade. Rasch descreveu isso em 1977, afirmando que as comparações entre itens não dependem dos examinandos que os respondem e também, as comparações obtidas por sujeitos

são independentes dos itens aplicados. Gerando assim, que pelo conceito de "especificidade objetiva" o modelo Rasch é considerado como um modelo que não depende de amostras, ou seja, as comparações entre os itens e a determinação das dificuldades são independentes dos alunos, e simultaneamente, a comparação entre alunos e a determinação de suas habilidades independe dos itens a que são submetidos. Desta forma, essa objetividade do teste não gera necessidade de calibrá-lo para cada amostra na qual é realizado (CHACHAMOVICK, 2007, p. 63).

O modelo Rasch, assim como os modelos TRI, apresenta suposições necessárias para sua credibilidade, a unidimensionalidade – o teste deve medir apenas um traço latente, e também, a ausência de dependência local entre os itens – a probabilidade de acerto ou erro em um determinado item devem ser independentes entre si e com as probabilidades dos demais itens do teste (ANDRICH, 1988; PALLANT & TENANT, 2007; COUTO & PRIMI, 2011).

As principais contribuições deste modelo para a análise de instrumentos de medida e testes são a aferição de estrutura intervalar da medida, a consistência interna da medida, o exame do comportamento de resposta (*thresholds*) – parâmetro "*b*", o comportamento invariante dos itens (*differential item functioning*) – DIF e a calibragem dos itens (CHACHAMOVICK, 2007, p. 69).

Dentre os modelos Rasch tem se destacado o uso do Modelo Multifacetas de Rasch (MFRM), diversos pesquisadores de fora do Brasil publicaram estudos em forma de artigos e livros sobre o MFRM aplicado em diversos campos de pesquisa e em tipos de avaliações variadas. Algumas aplicações feitas do MFRM são na avaliação de teste de língua estrangeira como o americano TOEFL[3], o espanhol DELE[4], também em pesquisas de satisfação como os SET[5]. O MFRM tem contribuído com a evolução das análises das avaliações em larga escala, quando em seus testes apresentam fontes de variação que criam viés nos resultados e que podem ser consideradas como facetas dentro do modelo, sendo mensuradas e explicadas de forma consistente.

Contudo, antes de discutirmos o MFRM faz-se necessária a contextualização de uma das principais metodologias de análise de avaliações em larga escala, a Teoria de Resposta ao

[3] *Test of English as a Foreign Language* (TOEFL) é um teste que avalia sua capacidade de usar e compreender o inglês no nível universitário. Ele avalia também sua capacidade de combinar as habilidades de *Listening*, *Reading*, *Speaking* e *Writing* para realizar tarefas acadêmicas. https://www.ets.org/pt/toefl/ibt/about

[4] Diploma de Espanhol como Língua Estrangeira) DELE – sistema de provas da Espanha para obter o Diploma de Espanhol como Língua Estrangeira.

[5] SET (Student Evaluations of Teaching) são avaliações de satisfação dos alunos que fornecem de maneira estruturada a coleta da opinião dos alunos sobre o curso e a eficácia do professor.

Item (TRI), que possibilitou o avanço nas análises das avaliações educacionais, sempre alavancado por pesquisadores interessados no desenvolvimento da área.

Além de uma breve explicação teórica sobre a TRI, na próxima seção encontram-se mais informações sobre a TRI na avaliação educacional, a sua relevância para a confiabilidade dos resultados e o ganho nas análises das provas aplicadas em avaliações em larga escala.

3.2 A Teoria de Resposta ao Item (TRI) na avaliação educacional

Segundo Andrade, Tavares & Valle (2000) o envolvimento brasileiro com a TRI inicia-se em 1996 com análises dos dados gerados em pesquisas e em sistemas de avaliação. Todavia, foi na década de 50 que surgiram os primeiros modelos de resposta ao item. Esses modelos consideravam que uma única habilidade, de um único grupo, estava sendo avaliada por um teste onde os itens eram dicotômicos, ou seja, corrigidos como ou certos ou errados. Posteriormente, foram desenvolvidos outros modelos importantes na TRI. As análises e interpretações avaliam a prova como um todo, diferentemente da Teoria Clássica dos Testes, caracterizada pelos resultados encontrados que dependem do particular conjunto de itens (questões) que compõem o instrumento de medida, ou seja, na TRI as análises e interpretações estão sempre associadas à prova como um todo, gerando escores brutos ou padronizados (ANDRADE, TAVARES & VALLE, 2000).

Klein (2013) avalia que a introdução da TRI na avaliação brasileira trouxe muitas vantagens sobre o método tradicional de avaliação. Pois ela coloca os itens em uma mesma escala e reforça a característica da comparabilidade, "a TRI permite estimar e comparar os resultados dos alunos, mesmo que eles respondam a itens diferentes" (KLEIN, 2013, p. 40).

Na Teoria de Resposta ao Item o foco é a análise de cada item, onde os modelos apresentam a probabilidade de resposta a um item como função da proficiência do aluno (habilidade - variável latente, não observável) e de parâmetros que expressam certas propriedades dos itens. Quanto maior a proficiência do aluno maior a probabilidade de ele acertar o item.

Uma das propriedades importantes da TRI é o fato de os parâmetros dos itens e as proficiências dos indivíduos serem invariantes. Para Couto e Primi (2011) "sua principal

contribuição do ponto de vista teórico é a invariância dos parâmetros de medida, além de apresentar inovações técnicas como as funções de informação dos itens e do teste". Tanto os parâmetros dos itens obtidos de grupos diferentes de alunos testados quanto os parâmetros de proficiência baseados em grupos diferentes de itens são invariantes, exceto pela escolha de origem e escala.

Graças à propriedade da invariância, a TRI, associada a outros procedimentos estatísticos, permite comparar alunos, estimar a distribuição de proficiências da população e subpopulações e ainda monitorar os progressos de um sistema educacional.

O que esta metodologia, de correção das provas, sugere são formas de representar a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item e seus traços latentes, proficiências ou habilidades na área de conhecimento avaliada. Permitindo a comparação entre populações, desde que submetidas a provas que tenham alguns itens comuns, ou ainda, a comparação entre indivíduos da mesma população que tenham sido submetidos a provas totalmente diferentes (ANDRADE, TAVARES & VALLE, 2000). Pois a utilização de itens comuns nas provas permite a comparabilidade dos alunos que fazem provas diferentes, gerando uma mesma escala de proficiência para o desempenho dos alunos.

Ainda segundo Andrade, Tavares e Valle (2000), os diversos modelos propostos na literatura dependem essencialmente de três fatores: (i) a natureza do item – dicotômicos ou não dicotômicos; (ii) o número de populações envolvidas – apenas uma ou mais de uma; e, (iii) a quantidade de traços latentes que está sendo medida – apenas um ou mais de um.

No Brasil, o modelo de três parâmetros da TRI é amplamente aplicado para a correção de testes realizados nas avaliações em larga escala. A determinação do desempenho dos examinandos, considerando a dificuldade do item, a discriminação e o acerto ao acaso, outro fator do modelo é a habilidade do examinando, compõem o modelo logístico de três parâmetros, conforme apresentado:

$$P\left(U_{ij}=1\middle|\theta_j\right)=c_i+\left(1-c_i\right)\frac{1}{1+e^{-Da_i\left(\theta_j-b_i\right)}} \qquad (2)$$

com i = 1, 2, · · · , n, e j = 1, 2, · · · , m.

Uij é uma variável dicotômica (igual a 1, quando o indivíduo j acerta o item i, ou 0 quando o indivíduo j erra o item i.
θj representa a habilidade (traço latente) do *j-ésimo* indivíduo.
P(Uij = 1|θj) é a probabilidade de um indivíduo *j* com habilidade *θ* responder corretamente o item *i* e é chamada de Função de Resposta do Item – FRI.
bi é a dificuldade do item *i*, medido na mesma escala da habilidade.
ai é a discriminação do item *i*.
ci é a probabilidade de acerto ao acaso (chute).
D constante e igual a 1 para o modelo logístico e assume o valor 1,7 quando deseja-se que a função logística forneça resultados semelhantes ao da função ogiva normal.

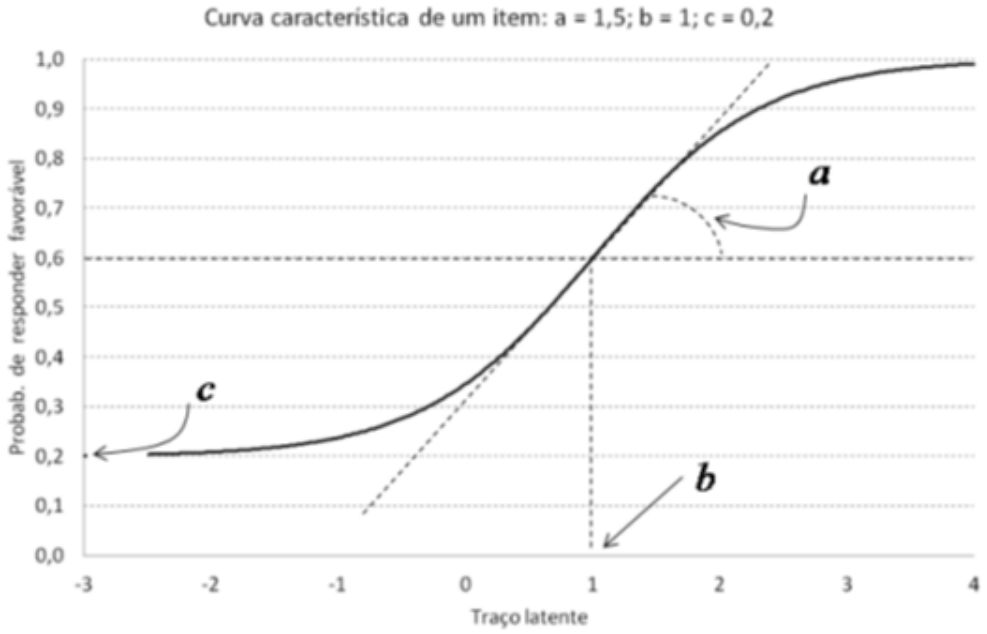
De maneira geral, a escala de medida do traço latente *θ* segue uma distribuição normal padrão, ou seja, possui média 0 e desvio padrão 1. Como o parâmetro de dificuldade do item *b* é expresso na mesma escala de *θ* (proficiência do aluno), um item com *b* = - 2, por exemplo, pode ser considerado fácil para um examinando de proficiência mediana. De forma análoga, um examinando com *θ* = 2 tem alta proficiência enquanto um com *θ* = - 2 tem baixa proficiência (BARBETTA et al., 2014).

Quanto ao parâmetro *a*, ele não assume valores negativos, pois isso significaria que a probabilidade de responder corretamente a um item diminuiria com o aumento da habilidade do examinando. Valores baixos para o parâmetro *a* indicam que o item não consegue diferenciar corretamente os respondentes. E, valores muito altos de *a* indicam itens que discriminam os alunos apenas em dois grupos: os que possuem habilidades acima do valor do parâmetro *b* e os que possuem habilidades abaixo do valor de *b* (ANDRADE, TAVARES & VALLE, 2000).

Ainda existe o parâmetro *c*, que representa o acerto ao acaso - uma probabilidade, ou seja, um valor entre 0 e 1. Quanto mais próximo de 0, menor é a chance de um indivíduo com baixa habilidade acertar um item considerado difícil para ele.

Temos a representação gráfica (Figura 2) de um item no qual é possível observar os parâmetros descritos anteriormente com os seguintes valores: a = 1,5; b = 1; e c = 0,2.

Figura 2: Curva Característica de um Item com Parâmetro a = 1,5, b = 1 e c = 0, 2.



Fonte: (BARBETTA et al., 2014)

O Modelo Dicotômico de Rasch é também conhecido por modelo logístico da TRI de 1 parâmetro (ML1). Nele, quanto maior for a habilidade do examinando em relação à dificuldade do item maior será a probabilidade de responder ao item corretamente. Os modelos apresentam pressupostos importantes que devem ser atendidos quando da sua utilização, como a invariância das medidas ou objetividade das medidas; a unidimensionalidade; e a independência local.

Com o avanço dos métodos de avaliação e também das metodologias de correção dessas avaliações, os estudos estão extrapolando a avaliação do conhecimento e englobando a medição do construto aprendizagem. Neste sentido, é importante desenvolver estudos com análises que contemplem tal construto e que possam gerar relatórios com definições e resultados sobre a parte pedagógica do processo de aprendizagem.

A busca por modelos robustos de análise dos vieses que afetam os resultados dos examinandos encontrou nos modelos Rasch, principalmente nos Modelos Multifacetas de Rasch, uma alternativa para o desenvolvimento de metodologias que têm se mostrado capazes de garantir a compreensão dos fatores causadores de variabilidade nos resultados das correções, chamados de facetas.

3.2.1 O pressuposto da unidimensionalidade da prova

O conceito de unidimensionalidade já foi abordado anteriormente que tratou-se de pressupostos dos modelos utilizados nas análises de avaliações (modelos de TRI e Rasch), a busca pela unidimensionalidade foi, desde sempre, almejada nos testes, até mesmo na Teoria Clássica dos Testes (TCT), pois conceitos como o de homogeneidade e de dificuldade só fazem sentido quando um único atributo é avaliado.

Considerando que o desempenho de um examinando é sempre determinado e motivado por mais de um fator, é possível que existam "traços latentes" presentes em qualquer testagem. Assim, o desempenho num teste pode ser influenciado por outras variáveis cognitivas, ou, mesmo, por fatores ligados à própria aplicação do teste, como por exemplo a motivação, a ansiedade, o uso correto das folhas de resposta, entre outros (VITORIA, ALMEIDA & PRIMI, 2006). E, Pasquali e Primi (2003, p. 104) defendem que a unidimensionalidade deve ser tratada como uma questão de grau, ou seja, que para ser satisfeita é suficiente considerar que os dados possuam um fator ou traço dominante responsável pelo desempenho num conjunto de itens de um teste.

Ainda segundo Vitoria, Almeida e Primi (2006) não existe uma definição de unidimensionalidade comum a todos os autores, pois a depender de como o teste é analisado e de como os resultados serão usados ela pode não ser alcançada.

Uma das premissas da TRI é que o instrumento de avaliação tenha um traço latente dominante, o que no extremo seria a unidimensionalidade (ANDRADE, TAVARES & VALLE, 2000; DE AYALA, 2009). Quanto aos métodos de análise da dimensionalidade, pode-se considerar que se um teste é unidimensional, quando submetido à análise fatorial ou à análise em componentes principais emergirá um único fator. No estudo para verificar este pressuposto foi utilizada a técnica mais comum que consiste em realizar a análise fatorial restrita com análise de componentes principais sobre a matriz de correlação tetracórica dos itens.

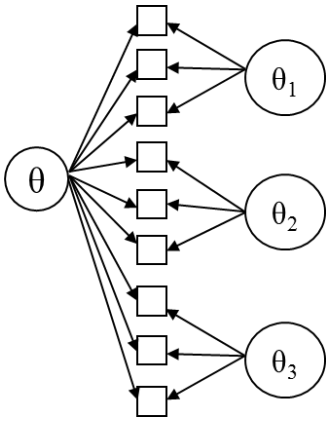
No caso de variáveis psicológicas a obtenção de um único fator é praticamente impossível, conforme já discutido. Levando assim, para a questão do grau de subjetividade associado ao se definir sobre a unidimensionalidade de um teste. Logo, a melhor alternativa é que na análise feita, o fator principal explique a maior quantidade de variância possível. Sendo que essa dimensão pode representar uma composição de traços latentes (RECKASE *apud* BARBETTA et al., 2014).

Considerando o caderno de Ciências da natureza e suas tecnologias com 45 itens, a Figura 3 apresenta um resumo dessa análise apresentando na ordem os autovalores da matriz de correlação, que podem ser interpretados como variâncias explicadas por cada dimensão considerada. Graficamente, verifica-se que o maior autovalor é bem superior aos demais, caracterizando uma dimensão dominante, a unidimensionalidade da amostra da prova do ENEM 2016 analisada foi comprovada.

3.2.2 O modelo Bifatorial

Os dados do caderno de Ciências da natureza e suas tecnologias quando ajustados pelo modelo Bifatorial possui um traço latente dominante no teste e três subdomínios ou subdimensões. Nesse modelo existe um fator comum que explica as intercorrelações de todos os itens, e possui também fatores de cada grupo que tentam capturar a covariância dos itens que é independente da covariância devida ao fator comum (REISE, MORIZOT & HAYS, 2007).

Figura 3 – Esquema do modelo Bifatorial



Fonte: Adaptado de Reise, Morizot e Hays (2007)

Em Gibbons e Hedeker (1992) tem-se que o modelo Bifatorial força cada item *i* a ter um índice de discriminação diferente de zero na dimensão principal, *ai1*, e um segundo índice, *aik*, *k = 2, ..., s*, em apenas um dos *s - 1* fatores. Supondo que os dois primeiros estejam na primeira subdimensão, os dois segundos estejam na segunda subdimensão e os dois últimos na terceira subdimensão, para seis itens, a matriz dos índices de discriminação seria:

$$a = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ a_{31} & 0 & a_{33} & 0 \\ a_{41} & 0 & a_{43} & 0 \\ a_{51} & 0 & 0 & a_{54} \\ a_{61} & 0 & 0 & a_{64} \end{bmatrix}$$

Um modelo Bifatorial hipotetiza que (a) há um fator geral que explica a comunalidade compartilhada pelas facetas, e (b) existem múltiplos fatores específicos, cada um dos quais explica a influência única do componente específico para além do fator geral, ou seja, explica as facetas.

Argumentamos que o modelo Bifatorial tem ampla aplicabilidade na medida em que fornece informações mais ricas e conceitualmente menos ambíguas do que a abordagem de escore total ou individual. A análise com MFRM traz essa apresentação de resultados.

Assim como na abordagem do escore total, o modelo Bifatorial estima uma variável latente geral com maior validade de conteúdo do que qualquer uma de suas facetas. Já na abordagem de pontuação individual, o modelo Bifatorial testa as contribuições exclusivas das facetas. No entanto, ele supera a principal desvantagem da abordagem de escore individual, pois elimina a semelhança entre as facetas ao testar a associação única entre cada faceta e uma variável externa.

Mais especificamente, o modelo Bifatorial tem duas vantagens centrais. Primeiro, testa simultaneamente a associação de uma variável de resultado com o fator latente geral e a única contribuição dos fatores específicos que são distintos do construto geral. Em segundo lugar, o modelo Bifatorial pode ser usado para identificar uma faceta que pode deixar de ser um contribuinte único, depois de levar em consideração a variância comum compartilhada com outras facetas (VIEIRA & BARBETTA, 2016).

Juntas, as vantagens do modelo Bifatorial podem levar a uma maior clareza conceitual do que a abordagem de pontuação total ou individual. O modelo Bifatorial distingue claramente as variâncias explicadas pelo fator comum e os fatores específicos, enquanto as outras duas abordagens não fazem tais distinções.

3.3 A prova de Ciências da natureza e suas tecnologias do ENEM

Na construção de avaliações de larga escala, o INEP busca elaborar matrizes de referência específicas para cada avaliação. Dado que a aprendizagem não pode ser medida de maneira direta, necessita-se identificar quais são as características ligadas à aprendizagem diretamente observáveis, valendo-se de um construto e de teorias que o embasam. O objetivo das matrizes de referência é agrupar características que analisadas conjuntamente gerem informações sobre esse construto.

Desse modo, a matriz retrata a escolha por determinados saberes e informações que representam o construto examinado em detrimento de outros, mas não é negada a existência de outros saberes ou informações significativas que podem contribuir para a visão dele. Tal fato ocorre por limitações dos instrumentos elaborados para a avaliação em larga escala. Por essa razão, a matriz de referência é um recorte de determinada realidade, ou seja, a escolha dos saberes e eixos analisados deriva de opções fundamentadas em pareceres técnicos, políticos e pedagógicos. A justificativa de tal recorte se ampara tanto pelas limitações dos instrumentos quanto por uma opção política sobre o que deve ser analisado em um dado construto (BRASIL, 2009).

Diante do exposto, os saberes elencados no teste avaliado são agrupados em áreas de conhecimentos, que podem não contemplar todos os assuntos referentes às disciplinas agrupadas. Esse fato reforça a necessidade de análises mais detalhadas dos itens. A aplicação de metodologia quantitativa com resultados objetivos pode colaborar com as análises qualitativas tornando-se instrumento para os responsáveis pela tomada de decisão sobre o processo ensino-aprendizagem e para educadores na criação de melhorias no ensino.

Para a aplicação do Modelo Multifacetas de Rasch neste estudo foi utilizada base de dados amostral, com as respostas das 45 questões do caderno de Ciências da natureza e suas tecnologias, essa amostra é de 100.000 examinandos que realizaram a prova do ENEM no ano de 2016. A amostra foi extraída da base ENEM 2016 e disponibilizada para a realização da pesquisa sem as variáveis que poderiam identificar os examinandos.

Foi realizada a análise descritiva das variáveis constantes na base de dados. Os resultados dessa análise estão nas tabelas a seguir. A distribuição da amostra representa bem todos as unidades da federação, conforme Tabela 1. Já a representatividade da variável sexo

também se comportou de forma semelhante ao que ocorre na população brasileira, com um percentual levemente superior de mulheres (Tabela 2). Quanto à faixa etária da amostra, a maior parte dos participantes tem entre 16 e 20 anos de idade (51,0%). Essa faixa de idade é a mais provável de abarcar o principal público do ENEM, a saber, os formandos do Ensino Médio.

Tabela 1 – Frequência e percentual da amostra de examinandos do ENEM 2016 por Unidade da Federação.

UF	Frequência	Percentual
AC	818	0,82
AL	1.763	1,76
AM	2.057	2,06
AP	709	0,71
BA	7.341	7,34
CE	6.153	6,15
DF	1.889	1,89
ES	1.805	1,81
GO	3.279	3,28
MA	4.154	4,15
MG	10.602	10,60
MS	1.538	1,54
MT	1.838	1,84
PA	5.276	5,28
PB	2.862	2,86
PE	5.024	5,02
PI	2.354	2,35
PR	4.486	4,49
RJ	6.226	6,23
RN	2.440	2,44
RO	1.244	1,24
RR	251	0,25
RS	4.915	4,92
SC	2.225	2,23
SE	1.413	1,41
SP	16.430	16,43
TO	908	0,91
Total	100.000	100,00

Fonte: Dados da amostra

Tabela 2 – Frequência e percentual da amostra de examinandos do ENEM 2016 por sexo

SEXO	Frequência	Percentual
F	57.863	57,86
M	42.137	42,14
Total	100.000	100,0

Fonte: Dados da amostra

Tabela 3 – Frequência e percentual da amostra de examinandos do ENEM 2016 por faixa etária.

Faixa Etária	Frequência	Percentual
Até 16	10.841	10,84
De 16 a 20	51.001	51,00
De 20 a 25	19.187	19,19
De 25 a 35	12.255	12,26
De 35 a 45	4.643	4,64
Mais de 45	2.073	2,07
Total	100.000	100,00

Fonte: Dados da amostra

3.4 O Modelo Multifacetas de Rasch (MFRM)

O Modelo Multifacetas de Rasch (MFRM) desde a primeira demonstração teórica apresentou um crescimento rápido na sua aplicação em estudos com testes de linguagem, avaliações educacionais e psicológicas, pesquisa de satisfação, estudos de saúde, entre outros. O MFRM é adequado quando são necessárias análises simultâneas de múltiplas variáveis que são fontes responsáveis pela ocorrência de erros nas avaliações, com a sua utilização incorpora-se mais parâmetros ao modelo, as facetas. (LINACRE, 1994; ECKES, 2011).

No caso das avaliações educacionais com itens abertos, além da habilidade do examinando e da dificuldade das tarefas ou dos itens, existem outros fatores que influenciam o processo de avaliação, por exemplo, podemos acrescentar fatores ligados ao formato da prova, ao avaliador, entre outros. A inovação desse modelo é, justamente, sua capacidade de estimar as facetas e com isso estimar de forma mais precisa a habilidade do examinando. Apresentando um resultado de medida mais objetivo e que pode ser interpretado de forma direta, sendo

apresentados para os atores do processo conjuntamente com as análises das provas e dos vieses encontrados.

Para Toffoli e Simon (2018, p. 3) "a possibilidade de obter informações que possam servir de diagnóstico, em nível individual, sobre o funcionamento de cada elemento é considerada valiosa e torna a utilização do modelo MFRM muito vantajosa".

O modelo apresentado é um modelo linear aditivo baseado numa transformação logística com as classificações em escala *logito*. A transformação logística de razões de probabilidades, ou seja, a *log odds* é a nossa variável dependente com várias facetas, no caso: examinandos, itens e cadernos de questões, consideradas como as variáveis independentes. As facetas presentes no modelo são dadas na mesma escala *logito* que pode variar no intervalo $(-\infty,\infty)$. Empiricamente, valores mais comuns pertencem ao intervalo $(-5,5)$. A definição de *logito* é a distância ao longo da escala da variável latente que aumenta a probabilidade de observar o evento especificado no modelo por um fator de aproximadamente 2,7178, o valor de *e* (ECKES, 2011).

O MFRM de três facetas conforme apresentado por Linacre (1994) também consta em Eckes (2011) – com nomenclatura diferente, para ser aplicado a testes que possuem uma única escala de classificação para todas as subescalas em todos os itens, ou seja, escala gradual – e é dado pela seguinte fórmula:

$$\ln\left[\frac{P_{sih}}{1-P_{sih}}\right] = \theta_s - b_i - d_h \qquad (3)$$

Onde,
$P_{sih}(\theta_s)$ é a probabilidade do indivíduo s, acertar o item i, da subescala h.
$1-P_{sih}(\theta_s)$ é a probabilidade do indivíduo s, não acertar o item i, da subescala h.
θ_s é a habilidade do indivíduo s.
b_i é a dificuldade do item i.
d_h é a subescala h.

A equação (3) refere-se ao Modelo Multifacetas de Rasch de escala gradual de três facetas: habilidade do examinando, dificuldade do item e subescala de disciplinas, esse modelo de escala gradual (ANDRICH, 1978 apud ECKES, 2011) é adequado para itens com escalas de pontuação ordenadas igualmente espaçadas. No caso, pode ser considerada como o parâmetro

de transição que separa o 0 (acerto) do 1 (não acerto). Quanto maior for b_i, maior a dificuldade do item e, portanto, menor a probabilidade de acertá-lo (atribuir resposta 1).

As propriedades e os recursos do MFRM são os mesmos dos modelos Rasch: medição conjunta, estatísticas suficientes, nível de intervalo, objetividade específica, quantificação da precisão em nível local e análise do ajuste ao modelo das pessoas, dos itens, dos avaliadores e das categorias de avaliação (ECKES, 2011; ADÁNEZ, 2011).

Com a utilização dos modelos MFRM é possível estimar subescalas para os cadernos de questões possibilitando uma interpretação pedagógica valiosa para a análise dos atores envolvidos no processo educacional. Com essa premissa, é possível assegurar inferências sobre a qualidade da educação e medidas assertivas sobre o desempenho em relação ao construto medido, garantindo, assim, o aumento da justiça, que está relacionada com a igualdade de condições a todos os seus participantes.

3.4.1 As Facetas

A definição de faceta segundo autores como Linacre (1994), Eckes (2011) e Engelhard (1994) pode ser apresentada como qualquer variável ou fator da avaliação que cria viés nos resultados da avaliação de modo sistemático. Podendo, as facetas incluírem tanto as variáveis de interesse direto, como a habilidade, quanto as que contribuem indiretamente para o aparecimento de erros nas medições, tais como os fatores ligados aos avaliadores, às formas das tarefas, o meio disponível para as provas, entre outros.

As facetas podem ser diversas, pois na aplicação de uma prova de múltipla escolha temos a habilidade dos alunos e a dificuldade dos itens do teste incorporadas ao modelo como facetas. Já em testes escritos, podemos acrescentar a severidade do corretor como terceira faceta. E ainda mais, como exemplificado por Eckes (2011), em uma avaliação com entrevista face a face podemos considerar cinco facetas: alunos, tarefas/questões/itens, entrevistadores, critérios de pontuação e avaliadores e também, em um teste que apresenta as questões em cadernos que abrangem várias dimensões/disciplinas. Desta forma, uma faceta pode ser considerada como qualquer variável ou componente que afete a medição na avaliação.

O Modelo Multifacetas de Rasch (MFRM) vem com a proposta de ser uma ferramenta útil para aferir a qualidade das avaliações nas quais se verifica a existência de fatores que geram variabilidade nos resultados e que podem ser mensurados. Tal modelo permite também aos pesquisadores análises para os efeitos individuais causados pelos elementos que fazem parte da avaliação, ou seja, cada examinando, cada avaliador, cada uma das tarefas, cada critério de pontuação utilizado, cada tipo de prova e etc.

O MFRM tem gerado avanços na teoria de medição e está no centro dessa discussão desse progresso, com a aplicação desse modelo em diversos campos de estudo. O surgimento e o desenvolvimento de novas ferramentas e técnicas sempre desperta o interesse dos pesquisadores por estudos que garantam a aplicabilidade da metodologia.

3.4.2 As subescalas e o Modelo Multifacetas de Rasch (MFRM)

No ENEM, para cada área avaliada, existe uma escala de proficiência geral baseada nos resultados combinados para todos os itens dentro dessa área. Além disso, na estrutura dentro das áreas existem disciplinas que podem suportar a construção de subescalas com base nas várias dimensões da estrutura. Com a intervenção de especialistas do Caed/UFJF foi possível realizar o processo de construção das subescalas utilizadas no Modelo Multifacetas de Rasch desse estudo.

Em seu artigo Brinthaupt e Kang (2014) construíram subescalas para escala original do *Self-Talk Scale* (STS) de aprendizado da fala o analisaram com um Modelo Multifacetas de Rasch de três parâmetros. A STS possui 16 itens aferidos numa escala Likert ((1 = nunca, 2 = raramente, 3 = às vezes, 4 = frequentemente, 5 = muito frequentemente), para esses itens foram construídas quatro subescalas. Apesar do tipo de questão ser diferente, no artigo de Brinthaupt e Kang (2014) os itens são politômicos, a metodologia mostrou-se adequada.

O PISA utiliza subescalas nas suas análises e quando as subescalas são incluídas, a metodologia preconiza que elas devem surgir de maneira clara da estrutura do domínio, devem ser significativas e úteis para objetivos de *feedback* e para a confecção de relatórios com análises pedagógicas e também, garantir os pressupostos das propriedades de medição. Desta forma, a primeira etapa do processo envolve a análise de especialistas para a criação de possíveis subescalas baseadas na estrutura mais recente da avaliação (OCDE, 2017).

O que ocorreu no PISA 2015 ilustra bem o processo de criação de subescalas. Primeiramente, houve o trabalho na identificação de possíveis subescalas para a Ciência, além da escala geral de alfabetização científica. O trabalho iniciou-se com uma revisão das subescalas utilizadas no ciclo de 2006, quando a Ciência era o domínio principal. No PISA 2006 as subescalas selecionadas para inclusão no banco de dados foram as três subescalas nas seguintes dimensões científicas: *explicando fenômenos cientificamente, identificando questões científicas e usando evidências científicas*. Após a análise documental, o grupo de especialistas de 2015 recomendou a elaboração de relatórios sobre as três competências científicas: *explicar os fenômenos cientificamente, avaliar e conceber a investigação científica e interpretar dados e provas cientificamente*. Além disso, também foi recomendada a adição de duas subescalas de conhecimento: *conhecimento de conteúdo e conhecimento procedural / epistêmico* (OCDE, 2017).

A criação de subescalas nos domínios do PISA acontece quando o domínio é o principal do ano avaliativo. Segundo o Relatório Técnico no seu Capítulo 15, para leitura no ciclo do PISA 2000, além da escala global de alfabetização em leitura, foram consideradas duas subescalas baseadas: *no tipo de tarefa de leitura e na forma do material de leitura*. Para o PISA 2012, quando a leitura reverteu para o status de domínio menor, uma única escala de leitura de impressão foi relatada. No caso da matemática, uma única escala de alfabetização matemática foi desenvolvida para o PISA 2000. Com os dados adicionais disponíveis no ciclo de pesquisa de 2003, quando a matemática era o principal domínio de teste, foram criadas subescalas em torno de quatro ideias gerais: *espaço e forma, mudança e relações, quantidade* e *incerteza*. No PISA 2006 e no PISA 2009, quando a matemática era novamente um domínio menor, apenas uma única escala foi relatada (OCDE, 2017).

Os resultados do PISA acima citados ilustram como é possível trabalhar com subescalas em provas de avaliação em larga escala que contemplam conhecimentos considerados um único construto, pois a prova é unidimensional, mas possui dentro dela domínios mais específicos. Estes domínios podem ser estudados de forma separada, garantindo assim o melhor aproveitamento dos resultados. Quanto ao ENEM, a matriz de referência já apresenta uma estrutura esperada de subescalas para ser analisada. As questões são separadas por conteúdos distintos que formam as disciplinas, que após a avaliação dos especialistas do Caed/UFJF geraram as subescalas do modelo e por fim o caderno de área.

3.5 A ferramenta de análise - Software *FACETS*

Quando se trata de automação das análises de TRI, existem muitos *softwares* que executam as aplicações da TRI e de modelos Rasch, como *ConQuest, BilogMG, Multilog, Parscale e FACETS*. Para a análise do Modelo Multifacetas de Rasch no referido trabalho, utilizou-se o programa *FACETS* (Version 3.71.4; LINACRE, 2014), pois ele compreende vários modelos de TRI e Rasch. Especificamente, a partir do *FACETS*, pode-se trabalhar com itens dicotômicos ou politômicos em modelos multifacetados e a partir desses modelos, pode-se utilizar o *FACETS* para estimar os parâmetros do modelo especificado. Sendo que um dos principais interesses são os efeitos causados pelas facetas incorporadas ao modelo proposto. No estudo, a faceta estimada de maior interesse é a das subescalas para as disciplinas presentes no caderno de Ciências da natureza e suas tecnologias.

A apropriação do conhecimento para utilização do software *FACETS* tornou-se imprescindível para a continuidade do estudo. Os manuais e os artigos com aplicação do MFRM contribuíram para alcançar tal objetivo, uma vez que foi disponibilizada uma licença do software para as análises dos dados desta pesquisa[6].

O programa foi utilizado para estimar todas as facetas, ou seja, a proficiência individual de cada examinando (faceta 1), a dificuldade dos itens (faceta 2) e as proficiências específicas das subescalas (faceta 3). O critério de convergência utilizado foi o padrão do programa, ou seja, o procedimento de estimação de JMLE (*Joint Maximum Likelihood Estimation*). A seguir é apresentada a programação feita no software *FACETS* para analisar o modelo proposto.

[6] Para maiores esclarecimentos ver os manuais em Linacre (2012).

Figura 4 – Programação implementada no *FACETS*.

```
Title = ENEM CN 100000 ALUNOS ;
Score file = ENEMSC,sav    ; score files ENEMSC.1, ENEMSC.2 produced
Facets = 3 ; three facets, children, items, subscales  ; was: 2 ; two facets: children and items
Positive = 1 ; for facet 1, children, higher score = higher measure
Noncenter = 1 ; only facet 1, children, does not have mean measure set to zero
Pt-biserial = Yes ; report the point-biserial correlation
Yard = 112,4 ; Vertical rulers 112 columns wide, with 4 lines per logit
Vertical = 1*,2*,2A,3A
Model =
?,?,?,D  ; 3 facets ; was ?,?,D ; elements of the two facets interact to produce dichotomous responses
*
; log(Pni1/Pni0) = Bn - Di
; Bn = ability of child n, Di = Difficulty of item i,
; Pni1 = probability that child n on item i is scored 1.
Labels =
1,Children ; Children are facet 1
1-100000
* ; end of child labels for facet 1
2,items ; Items are facet 2
1-16 = F,,1 ; FÍSICA, in group 1, are numbered 1 through 16.
17-32 = Q,,2; QUÍMICA, in group 2, are numbered 17 through 32.
33-45 = B,,3; BIOLOGIA, in group 3, are numbered 33 through 45.
3, subscales, D ; dummy facet for subscale reporting - all elements anchored at 0
1 , FÍSICA,
2 , QUÍMICA,
3 , BIOLOGIA
*
Dvalues = 3, 2, $GROUP ; element number for Facet 3 is the group number of the element in Facet 2
Data = DATA  ENEM  NS.sav
```

Fonte: Elaboração da autora

3.6 Definição do Modelo Multifacetas de Rasch – 3 facetas

Conforme mostrado anteriormente, o modelo Rasch é uma abordagem de medição avançada objetiva. Inicialmente, temos o modelo Rasch de um parâmetro, considerado relativamente simples, que também pode ser chamado de modelo de duas facetas, pois caracteriza a relação entre o item e a habilidade do sujeito. Sendo que, após a calibração, ambos, itens e sujeitos, são colocados na mesma métrica, ou seja, em *logitos*; tal fato permite que os itens possam ser analisados conjuntamente.

O modelo Rasch de duas facetas pode ser estendido a um Modelo Multifacetas de Rasch (LINACRE, 1994) quando são acrescentadas outras variáveis que podem causar viés no resultado de uma avaliação. No estudo, as subescalas criadas na prova de Ciências da natureza

e suas tecnologias (Física, Química e Biologia) foram incluídas na modelagem, o modelo de duas faces torna-se um modelo de três facetas (ou seja, item, examinando e subescala) (BRINTHAUPT & KANG, 2014). A construção das subescalas contou com a participação de especialistas do Caed/UFJF que avaliaram a prova e determinaram qual a melhor divisão das questões pelas disciplinas contidas no caderno. O uso de um Modelo Multifacetas de Rasch permite estimar as dificuldades da subescala através da calibração de Rasch e "controlar" a dificuldade da subescala, na estimação dos parâmetros de dificuldade do item e habilidade do sujeito.

Embora o modelo Rasch tenha várias vantagens, também existem desafios que precisam ser superados para a aplicação do modelo. Primeiro, o modelo requer amostras maiores para obter estimativas com maior precisão possível, no estudo não foi problema pelo fato de analisar uma amostra de 100.000 respostas. Em segundo lugar, a abordagem do modelo de multifacetas Rasch é baseada na suposição de unidimensionalidade, portanto, os dados devem se encaixar no modelo unidimensional. Desta forma, o modelo não pode ser usado com medidas de natureza multidimensional.

A abordagem do modelo Rasch vem sendo aplicada com sucesso a uma série de estudos já mencionados no Capítulo 3. O objetivo do presente estudo foi usar o modelo Rasch para calibrar a prova de Ciências da natureza e suas tecnologias em uma abordagem mais detalhada dos itens, com a utilização das facetas no modelo, com objetivo de subsidiar futuros trabalhos que visem uma investigação pedagógica das provas de avaliação em larga escala.

Conforme mencionado anteriormente, a análise do Modelo Multifacetas de Rasch foi realizada usando o programa *FACETS*, no qual foi implementado o Modelo de Multifacetas de Rasch (LINACRE, 1994) utilizado para calibrar a escala com as três facetas: habilidade do sujeito (θ), dificuldade do item (b) e a subescala da prova de Ciências da natureza e suas tecnologias (d). O modelo de três facetas foi definido da seguinte forma:

$$\ln\left[\frac{P_{sih}}{1-P_{sih}}\right] = \theta_s - b_i - d_h \qquad (4)$$

Onde,
$P_{sih}(\theta_s)$ é a probabilidade de o indivíduo s acertar o item i da subescala h.
$1-P_{sih}(\theta_s)$ é a probabilidade de o indivíduo s não acertar o item i da subescala h.
θ_s é a habilidade do indivíduo s.

b_i é a dificuldade do item i.
d_h é a subescala h.

O processo de estimação dos parâmetros é estatístico e dele obtém-se as estimativas dos parâmetros do modelo utilizado. Os modelos Rasch podem empregar métodos iterativos e não iterativos. A análise do Modelo Multifacetas de Rasch, pelo software *FACETS*, utilizou método iterativo JLME (*Joint Maximum Likelihood Estimation*). Os métodos iterativos são definidos como métodos numéricos que iniciam o processo de cálculo com um valor inicial determinado (um chute) e as iterações vão ajustando esse valor até se alcançar um critério definido e aceitável de parada. Por esse motivo, a dificuldade dos itens, a habilidade dos examinandos e a dificuldade das subescalas somente podem ser estimadas de acordo com todas as respostas dos estudantes para cada questão.

No modelo desenhado cada subescala foi parametrizada também como um coeficiente aplicado a uma variável *dummy*. A variável *dummy* assume o valor "1" se a subescala participa da observação e "0" caso contrário. Em princípio, isso poderia ser estimado com um software estatístico padrão, mas raramente um software desse permite estimar as centenas ou milhares de parâmetros que podem ser encontrados em apenas uma análise Rasch. As "facetas *dummy*" são facetas destinadas apenas a investigar interações e ajustes, não para medir efeitos principais. Todos os elementos de uma faceta falsa estão ancorados em 0 (LINACRE, 2012).

3.6.1 Estatísticas de ajuste e análise do modelo

A análise do Modelo Multifacetas de Rasch gera estatísticas de ajustes dos dados ao modelo, essas estatísticas são muito importantes e devem ser obtidas para cada um dos parâmetros (facetas) presentes na definição do modelo. As estatísticas de ajuste estão ligadas à qualidade dos dados utilizados, elas são representadas pelas médias quadráticas residuais entre respostas observadas e esperadas: *MQ-Infit* e *MQ-Outfit* e pelas médias quadráticas padronizadas: *MQZ-Infit* e *MQZ-Outfit*. Consequentemente, o primeiro passo é avaliar o ajuste de dados pelo exame das estatísticas *MQ-Infit* e *MQ-Outfit* para cada faceta (habilidade, item e subescala). A seguir é apresentado quadro com a interpretação para os resultados das estatísticas de ajuste.

Quadro 3 - Interpretação do nível dos elementos das estatísticas de ajuste do quadrado médio

Escala ***Infit* e *Outfit***	**Interpretação**
>2.0	Distorce ou degrada o sistema de medição.
1.5 - 2.0	Improdutivo para construção de medição, mas não degradante.
0.5 - 1.5	Produtivo para medição.
<0.5	Menos produtivo para medição, mas não degradante. Pode produzir confianças e separações enganosamente boas.

Fonte: Adaptado de Linacre (2012)

A medida *Infit* é a média quadrática ponderada baseada no quadrado dos resíduos padronizados entre os dados observados e os esperados. Outro fato importante é que *Infit* é sensível a padrões de valores inesperados. Já *Outfit* é uma estatística de ajuste que é sensível a valores discrepantes individuais de itens ou examinandos. Ambas as estatísticas variam de zero a infinito positivo, onde um valor próximo a 1,0 indica um ajuste adequado e valores menores que 0,5 (pouca variação nas respostas) e maior que 1,5 (grande variação) indicam um ajuste ruim (LINACRE, 2012; TOFOLLI, 2014).

Ademais, a resposta para algumas perguntas pode ajudar a avaliar a qualidade do ajustamento do modelo aos dados da amostra, como: as estatísticas *Infit* e *Outfit* apresentam residual quadrado adequado para cada item e/ou subescala? Pois, valores superiores a 2 indicam que existem respostas mais inesperadas do que esperadas.

Em seguida, é apresentado o *facets map*. O *facets map* é a representação gráfica que liga a dificuldade do item, a subescala da prova e as estimativas de habilidade de sujeitos na escala comum em *logitos* (LINACRE, 2012). Ilustrando a distribuição da dificuldade dos itens, da subescala, dos níveis de intercomunicação dos participantes e a posição relativa do desempenho de um indivíduo para as subescalas e itens da prova, o que permite a comparação entre a dificuldade dos itens, a dificuldade das subescalas e as medidas de habilidade.

As medidas de dificuldade do item e a dificuldade de subescala foram estimadas durante o processo de calibração de Rasch em *logitos*. E, quanto maior o valor *logito*, menor a probabilidade de os participantes acertarem esse item. O nível individual de desempenho também foi estimado por meio da análise de Rasch.

Outro índice importante para a interpretação do modelo é o índice de separação, ele indica quão bem a escala separa os itens e subescalas e identifica as diferenças individuais ao

longo da escala de medição. Para as análises das localizações na escala de mensuração dos componentes envolvidos no modelo MFRM, são apresentadas as estatísticas de separação para cada uma das facetas. A saber: (1) a taxa de separação, que fornece a dispersão das medidas dos elementos da faceta em relação à precisão dessas medidas; (2) a confiabilidade do índice de separação - calculada como a razão da variância verdadeira das medidas pela variância observada dessas medidas (3) o estrato - número de níveis estatisticamente diferentes das medidas dos elementos da faceta; e, (4) o índice de homogeneidade - indica que as medidas de, pelo menos, dois elementos da faceta são significativamente diferentes (teste *Qui-Quadrado*) (MYFORD & WOLFE, 2004).

Em outra palavras, as medidas de ajuste e as estatísticas de separação são utilizadas com o objetivo de avaliar a adequação dos dados aos modelos e à qualidade das pontuações (WRIGHT & STONE, 1999; TOFOLLI & SIMON, 2018).

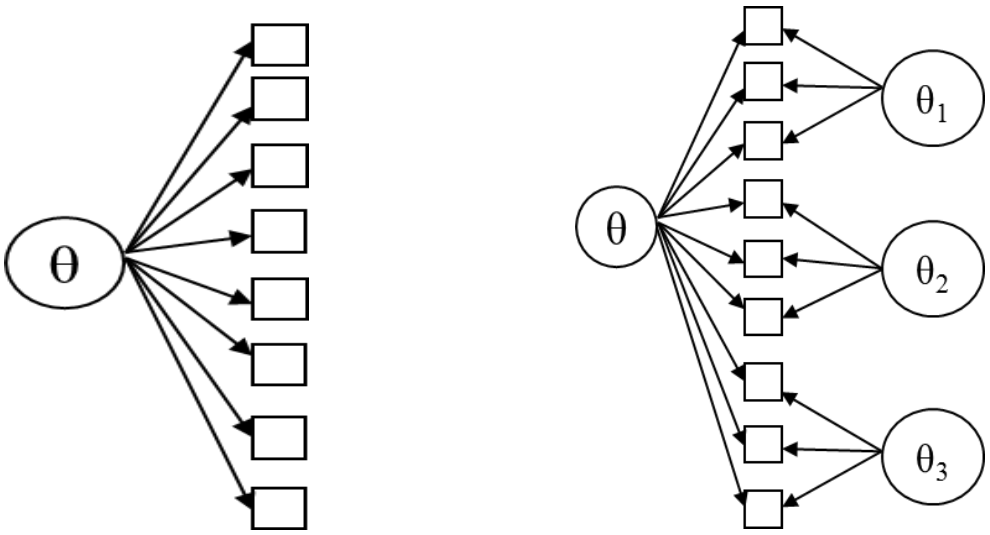
Um alto índice de separação e estatísticas de confiabilidade de separação próximas a 1.00 indicam uma boa discriminação para uma faceta (ou seja, item, subescala ou sujeito) ao longo da escala de medição com um alto grau de confiança (WRIGHT & STONE, 1999).

4. RESULTADOS DA APLICAÇÃO DO MODELO MULTIFACETAS DE RASCH E DA CONSTRUÇÃO DE SUBESCALAS

4.1 A dimensionalidade da prova de Ciências da natureza e suas tecnologias

O Modelo Multifacetas de Rasch com 3 facetas considerando a habilidade dos alunos, a dificuldade dos itens e as subescalas, foi aplicado considerando a premissa de unidimensionalidade do teste para a amostra e respostas do caderno de Ciências da natureza e suas tecnologias. O modelo (a) representa um modelo unidimensional em que existe um traço latente dominante.

Figura 5 – Esquema dos modelos de análise de dimensionalidade



(a) Modelo unidimensional (b) Modelo bifatorial

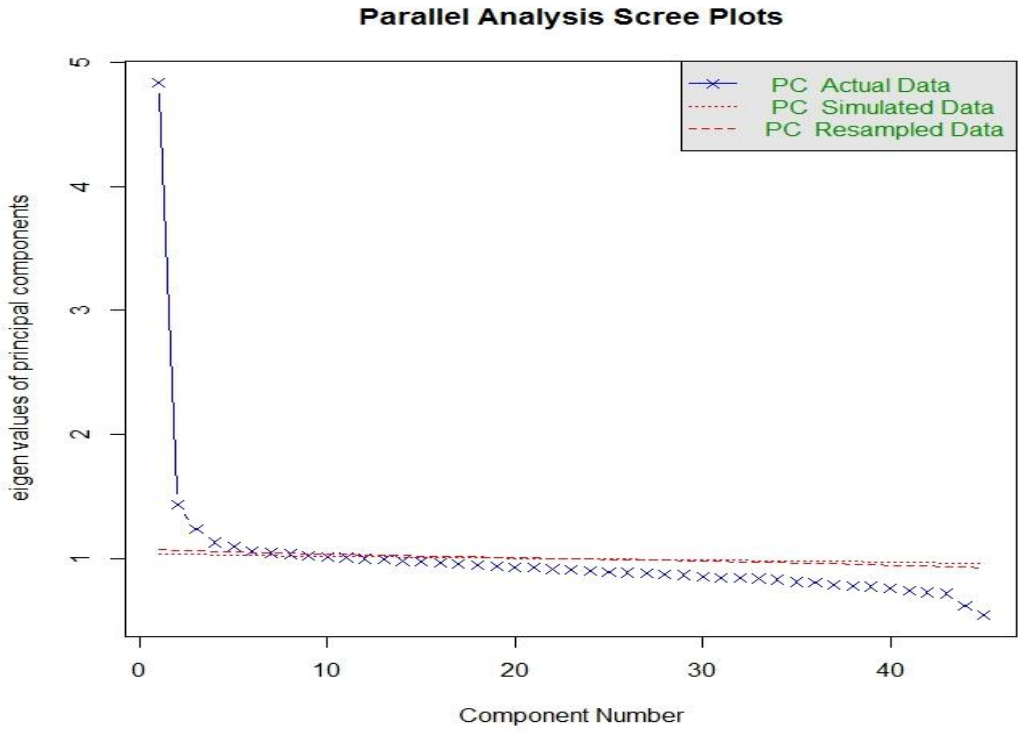
O modelo Bifatorial, ou seja, há um fator geral que explica a proficiência e as intercorrelações, mas, além disso, há também os fatores chamados "grupo", no estudo “subescalas”, que tentam capturar a covariação do item que é independente da covariação devido ao fator geral. Em outras palavras, o modelo (b) afirma que itens podem ser correlacionados porque compartilham um traço comum e uma fonte adicional de variação comum, possivelmente devido ao conteúdo dos itens compartilhados.

Sobre a análise de dimensionalidade da prova, ela foi testada utilizando a técnica de Análise Fatorial Confirmatória – AFC com o pacote MIRT do software R (os resultados estão no Anexo 2). Com a AFC buscou-se definir os fatores (variáveis latentes) e se os dados apresentavam um fator predominante com objetivo de validar o construto avaliado pela prova de Ciências da natureza e suas tecnologias.

Onde verificou-se que o modelo unidimensional, considerando a representação matemática, teve um bom ajustamento aos dados. Outro fato que deve ser considerado é que o modelo bifatorial se ajusta bem aos dados e ele é útil para avaliar a possibilidade de existir subescalas, pois considera que a estrutura dos dados apresenta um fator geral que explica a interrelação entre os itens e fatores dos grupos, que no estudo são as disciplinas que formam as subescalas. No modelo Bifatorial foi utilizada a procedure PROC CALIS do software SAS (os resultados estão no Anexo 2).

A representação gráfica dos fatores mostra a predominância de um fator sobre os demais, conforme afirmado por Pasquali e Primi (2003), para ser satisfeito o pressuposto da unidimensionalidade é suficiente considerar que os dados possuam um fator dominante.

Figura 6 - Screeplot variância explicada pelo número de dimensões



Fonte: Elaboração da autora.

4.2 Resultados das medidas de ajuste qualidade do modelo

A análise dos resultados da aplicação do modelo no software *FACETS* (LINACRE, 2012; LINACRE, 2014) apresenta as medidas de ajuste de *MQ-Infit* e *MQ-Outfit* para as facetas do modelo. A faceta dos itens teve um bom ajustamento dos dados com todos os resultados na faixa de produtivos para a medição, ou seja, entre 0,5 e 1,5 para os quadrados-médios. Na Tabela 4 com as estatísticas descritivas dos ajustes *Infit* e *Outfit* para os itens é possível observar que o valores mínimos e máximos estão dentro da faixa de boa adequação do ajustamento, conforme já explicado.

No geral, o MFRM se ajusta bem aos dados, uma vez que as estatísticas *MQ-Infit* e *MQ-Outfit* para os itens variaram de 0,84 a 1,16, indicando que todas as pontuações de 45 itens estavam dentro da faixa aceitável entre 0,5 e 1,5. Esta constatação indica que os 45 itens do caderno de Ciências da natureza e suas tecnologias se encaixam na construção unidimensional que a escala se destina a medir (isto é, proficiência em ciências). E também, para as três subescalas os índices estão bem ajustados para a quantidade de itens de 16, 16 e 13 respectivamente das subescalas de Física, Química e Biologia.

Tabela 4 - Estatísticas descritivas das estatísticas *MQ-Infit* e *MQ-Outfit* dos itens

Subescalas		N	Mínimo	Máximo	Média	Desvio padrão
Itens Física	*MQ-Infit*	16	0,90	1,08	1,0131	0,05056
	MQ-Outfit	16	0,88	1,16	1,0275	0,07371
Itens Química	*MQ-Infit*	16	0,94	1,06	1,0094	0,03255
	MQ-Outfit	16	0,92	1,10	1,0181	0,04778
Itens Biologia	*MQ-Infit*	13	0,89	1,05	0,9731	0,03924
	MQ-Outfit	13	0,84	1,06	0,9662	0,05440
Geral	*MQ-Infit*	45	0,89	1,08	1,0002	0,04429
	MQ-Outfit	45	0,84	1,16	1,0064	0,06425

Fonte: Dados da amostra

Na Tabela 5 apresentam-se os valores detalhados das estatísticas resumo para os 45 itens. Esse resumo é importante para identificar os itens de maior e de menor medida e assim realizar um detalhamento com objetivo de diagnosticar possíveis motivos que levaram ao

resultado obtido na modelagem. Colaborando assim com a parte pedagógica da avaliação e com análises qualitativas dos itens.

As dificuldades dos itens de ciências, os erros padrão e as estatísticas *Infit* e *Outfit* são relatadas na Tabela 5. Quanto maior a medição *logito* menor a probabilidade de os participantes acertarem o item. Embora os resultados suportem a estrutura unidimensional do teste, os dados de dificuldade dos itens também sugerem que subescalas criadas são importantes quando se considera a proficiência dos examinandos.

Tabela 5 – Resumo da calibração pelo modelo multifacetas de Rasch para a escala dos itens

(continua)

Número do Item	Grupo Subescala	Medida da dificuldade (*logitos*)	Erro Padrão (*logitos*)	*MQ-Infit*	*MQ-Outfit*
1 F	1	-0,58	0,01	1,03	1,03
2 F	1	0,29	0,01	1,08	1,13
3 F	1	-0,11	0,01	1,04	1,04
4 F	1	0,20	0,01	1,02	1,04
5 F	1	0,14	0,01	0,95	0,93
6 F	1	0,01	0,01	1,05	1,07
7 F	1	0,14	0,01	1,08	1,10
8 F	1	0,80	0,01	1,02	1,06
9 F	1	-0,16	0,01	0,94	0,94
10 F	1	0,02	0,01	0,97	0,97
11 F	1	-0,06	0,01	1,04	1,05
12 F	1	0,01	0,01	1,03	1,03
13 F	1	-0,47	0,01	1,02	1,03
14 F	1	1,16	0,01	1,05	1,16
15 F	1	0,16	0,01	0,99	0,98
16 F	1	-0,38	0,01	0,90	0,88
17 Q	2	-0,28	0,01	0,94	0,92
18 Q	2	-0,17	0,01	0,96	0,94
19 Q	2	0,15	0,01	1,00	1,01
20 Q	2	0,31	0,01	1,01	1,02
21 Q	2	0,01	0,01	1,02	1,03
22 Q	2	-0,71	0,01	1,00	1,00
23 Q	2	-0,08	0,01	1,03	1,04
24 Q	2	0,27	0,01	0,98	0,97
25 Q	2	0,06	0,01	1,03	1,06
26 Q	2	0,20	0,01	0,99	0,99
27 Q	2	0,59	0,01	1,06	1,10
28 Q	2	0,65	0,01	1,00	1,02
29 Q	2	0,30	0,01	1,04	1,05
30 Q	2	0,30	0,01	1,06	1,08

Tabela 5 – Resumo da calibração pelo modelo multifacetas de Rasch para a escala dos itens (conclusão)

Número do Item	Grupo Subescala	Medida da dificuldade (logitos)	Erro Padrão (logitos)	MQ-Infit	MQ-Outfit
31 Q	2	0,36	0,01	1,02	1,05
32 Q	2	-0,27	0,01	1,01	1,01
33 B	3	-0,31	0,01	0,97	0,97
34 B	3	0,06	0,01	1,01	1,02
35 B	3	-0,21	0,01	0,97	0,96
36 B	3	-0,15	0,01	0,94	0,94
37 B	3	-0,27	0,01	1,05	1,06
38 B	3	-0,63	0,01	0,97	0,96
39 B	3	-0,79	0,01	1,01	1,02
40 B	3	0,10	0,01	0,99	0,99
41 B	3	-0,98	0,01	0,96	0,95
42 B	3	0,95	0,01	0,95	0,92
43 B	3	-0,19	0,01	0,99	0,99
44 B	3	-0,53	0,01	0,95	0,94
45 B	3	0,10	0,01	0,89	0,84

Fonte: Dados da amostra

Ademais, a faceta dos examinandos apresentou valores bem ajustados de *Outfit* o percentual foi elevado na faixa de 0.5 a 1.5, ou seja, produtivo para medição, mas existem valores nas outras faixas da escala, esses valores estão na Tabela 6. Para a estatística *Infit*, 100% dos valores estão na faixa considerada ideal, os dados foram produtivos para a medição.

Tabela 6 – Distribuição da frequência por faixa de ajuste do *Outfit* dos examinandos.

Escala Ajuste	Frequência	Porcentual
< 0,5	3	0,003%
0,5 - 1,5	99.690	99,690%
1,5 - 2,0	299	0,299%
> 2,0	8	0,008%
Total	100.000	100,000%

Fonte: Dados da pesquisa

Na Tabela 7 temos os valores das estatísticas de ajuste para as subescalas, o fato de os valores das medidas de dificuldade estarem todos zerados é explicado pelo fato de as subescalas serem consideradas como variáveis *dummy* ancoradas em zero quando da calibração. Os valores de *Infit* e *Outfit* para as três subescalas demonstram o bom ajustamento dos dados ao modelo e podemos inferir que não foram encontrados valores inesperados ou discrepantes para as

subescalas. Também teve bom ajustamento a categoria de respostas das questões (0 e 1), a estatística *Outfit* foi igual a 1,00 para ambas as categorias.

Todas as facetas obtiveram valores de *Infit* e *Outfit* dentro do nível produtivo para a medição. Demostrando assim o bom ajustamento dos dados ao modelo, o que garante a possibilidade de utilização das análises em posterior relatório pedagógico das disciplinas de forma separada.

Tabela 7 – Resumo da calibração pelo modelo multifacetas de Rasch para as subescalas.

Nome subescala	Medida (*logitos*)	Erro Padrão (*logitos*)	*MQ-Infit*	*MQ-Outfit*
1 FÍSICA	0,00	0,00	1,01	1,03
2 QUÍMICA	0,00	0,00	1,01	1,02
3 BIOLOGIA	0,00	0,00	0,97	0,97

Fonte: Dados da pesquisa

As estatísticas de separação indicam o quanto os elementos da avaliação estão separados entre si (examinandos, itens e subescalas, etc.). Sendo que, as estatísticas de separação dos sujeitos indicam o quanto um conjunto de itens é apropriado para separar as habilidades das pessoas que estão sendo avaliadas. Já as estatísticas de separação dos itens indicam o quanto uma amostra de respondentes é capaz de separar os itens utilizados no teste quanto às suas dificuldades. As variações dessas estatísticas estão entre 0 e 1, e quanto mais próximas de 1 melhor a separação e mais precisa é a medição (LINACRE, 2014; TOFOLLI, 2015).

As estatísticas de separação apresentaram resultados considerados significativos, o que indica que existem diferenças entre os elementos dentro de cada uma das facetas. As medições realizadas neste estudo apresentaram valores altos para a confiabilidade de separação das facetas: subescalas de Física, Química e Biologia (1,00) e Itens (1,00). No entanto, a faceta Examinandos apresenta a confiabilidade de separação com valor um pouco menor, 0,58. Quanto aos valores de *Qui-quadrado* ($\chi 2$), eles são significativos com probabilidade $p < 0,05$. Esses índices indicam que os itens do caderno de Ciências da Natureza e suas tecnologias tiveram boa variabilidade com um alto grau de confiança na replicação da colocação dos itens dentro do erro de medição para outra amostra.

O índice estrato que representa a variação em unidades da variância do erro, teve valor para examinados de 1,91, sugerindo que há cerca de dois estratos estatisticamente diferentes

para o grupo, o que era esperado dado que as questões apresentadas são dicotômicas, ou o sujeito acerta ou erra a questão. Já para itens o valor do estrato foi 74,58 sugerindo que há mais de 70 estratos estatisticamente diferentes de níveis de desempenho para as questões.

O teste *Qui-Quadrado,* com hipótese nula de que todos os examinandos possuem o mesmo nível de desempenho, tem valor de 231.892 com 99.999 graus de liberdade e é uma medida estatisticamente significativa ($p < 0,05$), indicando que a habilidade dos examinandos varia entre os níveis da escala de pontuação.

Tabela 8 - Estatísticas de separação

Estatísticas	Examinandos	Itens	Subescala		
			Física	Química	Biologia
Taxa de separação	1,18	55,68	53,25	43,08	62,89
Confiabilidade de separação	0,58	1,00	1,00	1,00	1,00
Estrato	1,91	74,58	71,34	57,78	84,19
Qui-quadrado ($\chi 2$)	231.892*	130.554*	37.138*	32.015*	43.595*
Graus de liberdade	99.999	44	15	15	12

$*$ $p<0,05$

Fonte: Dados da pesquisa

Na Figura 7 mostra-se o *facets map* do Modelo Multifacetas de Rasch (examinando, item e subescala). A escala é apresentada em *logitos*, mostrados no lado esquerdo do mapa. A distribuição dos participantes (indicada por * e pontos, para examinandos cada ponto corresponde a 2.146 sujeitos e para item * = 2 itens) é mostrada ao lado esquerdo do mapa. A distribuição das subescalas do caderno de Ciências da natureza e suas tecnologias está localizada no lado direito do mapa. A distribuição de itens é colocada no centro do mapa com base nos níveis de dificuldade do item (indicados por B para itens de Biologia, F pra Física e Q para Química).

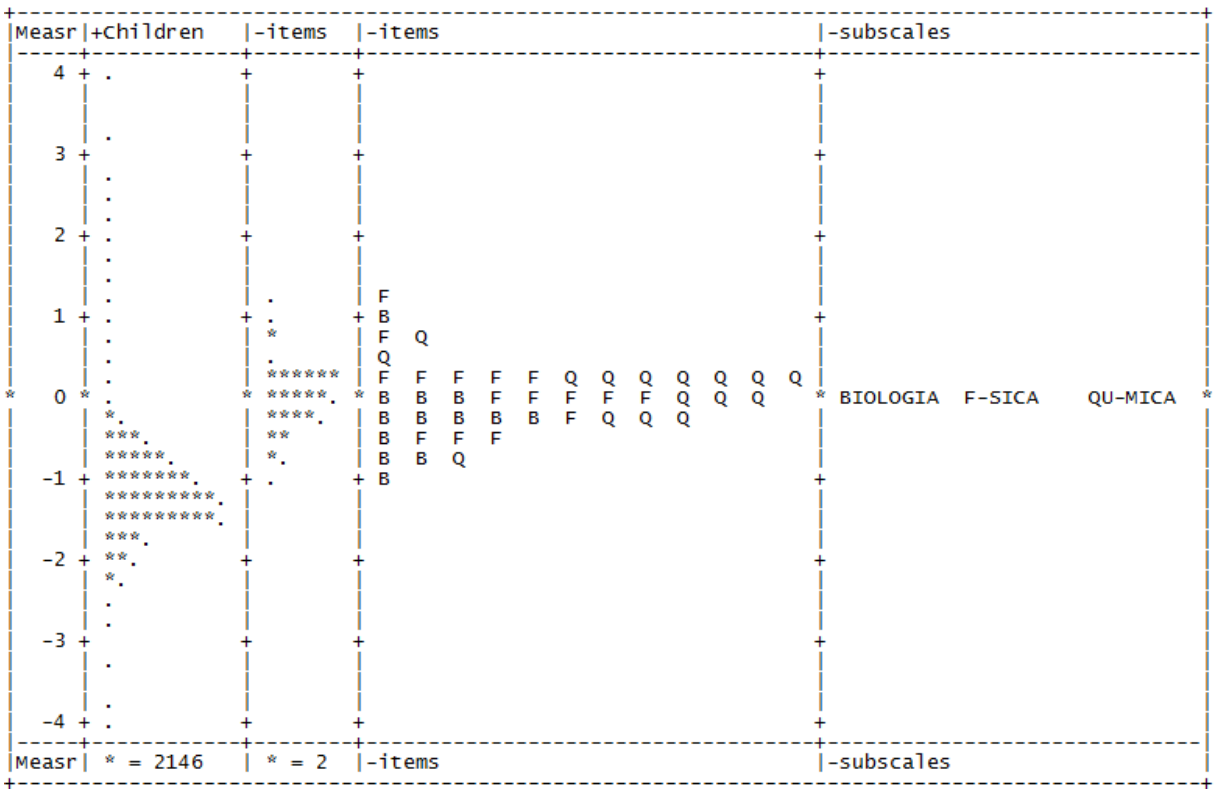
As medidas de todas as facetas são dadas na mesma escala em *logitos,* a primeira coluna. Por meio da figura mostra-se que os níveis de habilidade dos examinandos estão distribuídos na parte inferior ao longo da escala de *logitos*; a orientação dessa faceta é positiva, ou seja, quanto maior a pontuação (escore), maior é a medida, nesse caso, a habilidade do indivíduo, que apresenta em sua maioria valores negativos. Nesta prova, o desempenho dos examinandos foi baixo, temos mais sujeitos com menor habilidade em Ciências. As medidas dos desempenhos dos examinandos variam entre -3,87 e 3,88 *logitos,* embora a maior

concentração de indivíduos ocorra entre -2,0 e 0,0 *logitos*. E, a média das medidas da habilidade dos participantes é de M = -1,17 e o desvio-padrão é de SD = 0,57.

Os itens que são apresentados na terceira e quarta coluna possuem orientação negativa, o que significa que um escore maior corresponde a uma medida menor. Neste estudo, os itens obtiveram níveis de dificuldade próximas de zero. Demonstrando, assim, uma dificuldade dos itens dentro das subescalas com uma dispersão aceitável. Entretanto, os itens do caderno não são apropriados para grande parte desses indivíduos, pois não fornecem cobertura de conteúdo para indivíduos com os níveis mais altos (isto é, localizados em *logitos* >1). Já para os níveis com *logitos* <−1 a cobertura é boa. Apesar de existirem diferenças entre a dificuldade dos itens, estes não ocupam um intervalo amplo na escala de habilidades e estão todos localizados perto da origem.

A quinta coluna, apresenta a distribuição do desempenho das subescalas, e essa faceta possui orientação negativa, maior escore implica em menor medida. Todas as subescalas ficaram na mesma localização, com média 0,0 e com desvio-padrão de 0,0 *logitos*.

Figura 7 – *Facets Map* para item, examinando e subescala



Fonte: Dados da pesquisa. Software: *FACETS*.

4.3 Proficiências dos examinandos no Modelo Multifacetas de Rasch

Para comparar as proficiências dos estudantes da amostra analisada optou-se por realizar uma transformação na escala *logito*, que é apresentada nas análises do *FACETS*, e elas foram colocadas em uma mesma escala, com média 50 e desvio 10. A nota do ENEM também foi transformada para a mesma escala, isso torna de fácil entendimento as comparações entre os desempenhos. No Quadro 5 são apresentadas as estatísticas descritivas das proficiências, sendo possível avaliar que as médias das disciplinas são parecidas, já os desvios padrão apresentam pequenas diferenças.

Com os intervalos de confiança construídos notamos que os desempenhos sofrem pouca variação, o que é comprovado pelo cálculo dos Coeficientes de Variação[7], eles mostram que não existe grande variabilidade nos grupos. Outro ponto interessante é o fato das proficiências calculadas para as facetas estarem todos abaixo da média da escala escolhida para comparar os resultados (média=50 e desvio padrão = 10), o que corrobora com os resultados da aplicação do MFRM, pois os valores das dificuldades dos itens apresentados Tabela 5 e Figura 7 demonstram que a prova foi realmente difícil.

Tabela 9 – Estatísticas descritivas das proficiências.

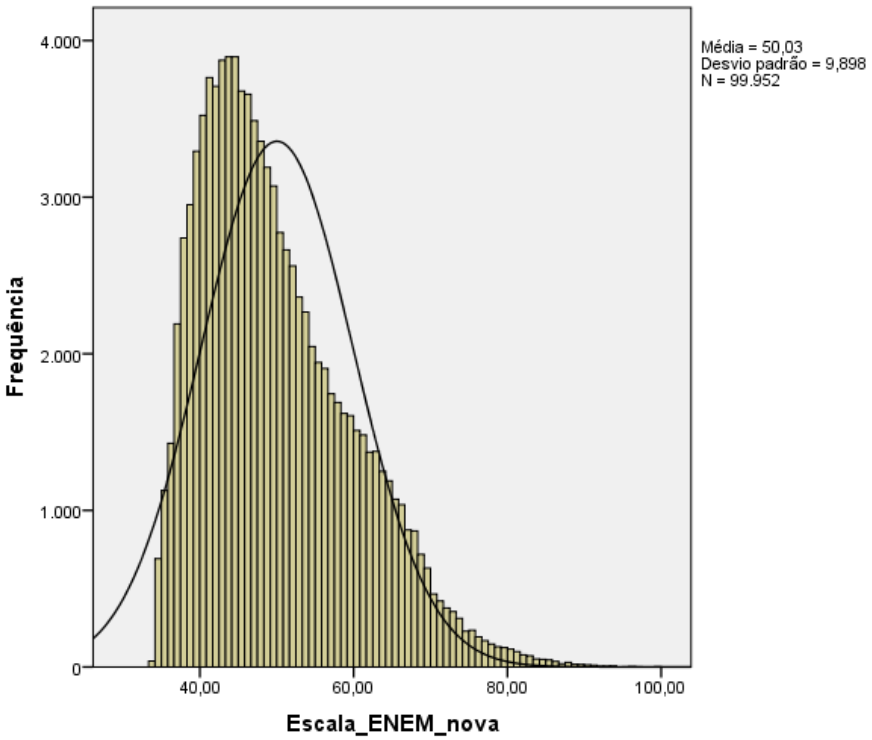
Estatísticas		Proficiência Geral	Proficiência Física	Proficiência Química	Proficiência Biologia	Nota ENEM
Média		38,31	37,63	37,61	37,43	50,03
Intervalo de confiança de 95% para média	LI	38,28	37,58	37,56	37,37	49,97
	LS	38,35	37,68	37,66	37,49	50,09
Mediana		38,30	39,30	39,80	39,30	48,00
Variância		32,21	63,30	65,66	94,18	97,95
Desvio padrão		5,68	7,96	8,10	9,70	9,90
Mínimo		11,30	10,80	11,30	10,20	33,96
Máximo		88,80	89,10	90,00	88,80	99,73
Coeficiente de Variação		15%	21%	22%	26%	20%

Fonte: Dados da pesquisa

[7] Coeficiente de Variação - É uma medida de dispersão relativa que elimina o efeito da magnitude dos dados. É dada pela divisão do desvio padrão pela média.

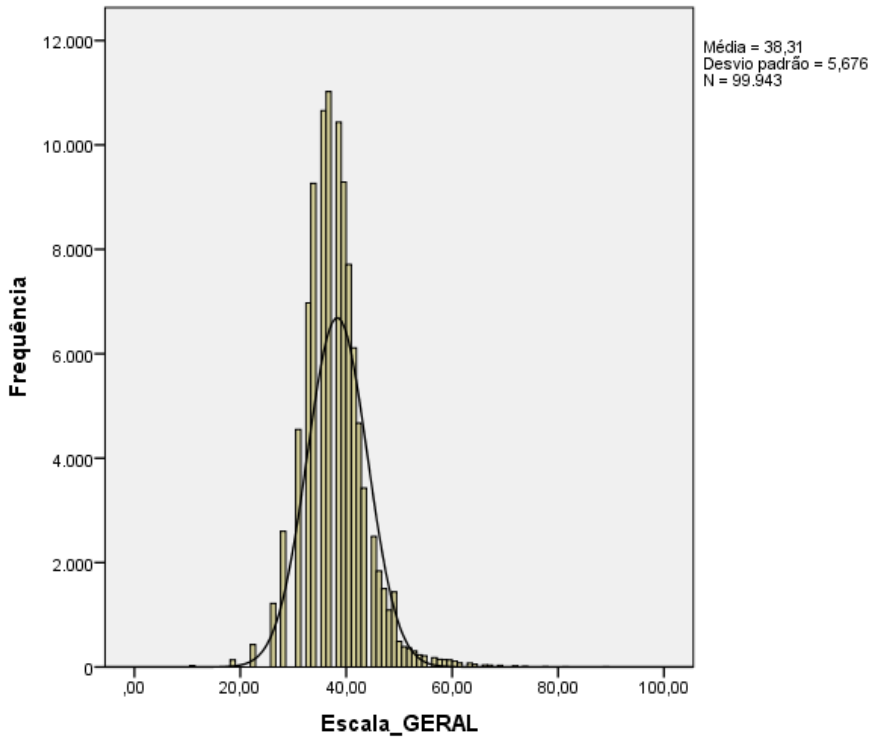
Os gráficos a seguir são a representação do desempenho dos examinandos por nota no ENEM, desempenho geral, isto é, a proficiência geral fornecida pela análise MFRM e desempenho por subescala do modelo de três facetas elaborado no estudo. No Gráfico 1 nota-se que as proficiências no ENEM têm assimetria positiva.

Gráfico 1 – Distribuição das notas de proficiência no ENEM (escala M=50, DP=10)



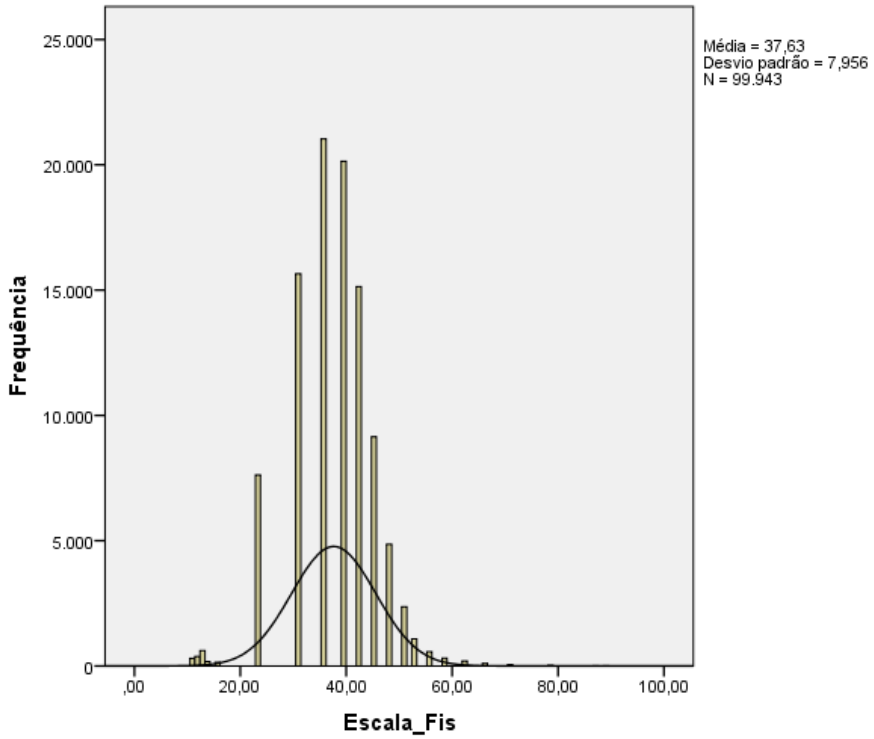
Fonte: Dados da amostra

Gráfico 2 – Distribuição das notas de proficiência GERAL do Modelo Multifacetas de Rasch (escala M=50, DP=10)



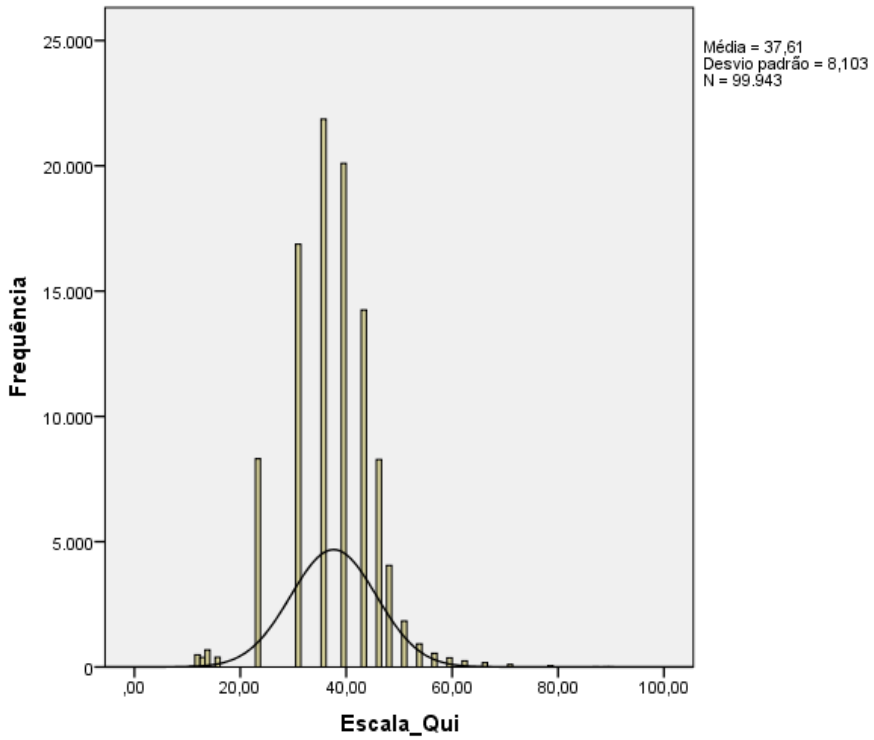
Fonte: Dados da pesquisa

Gráfico 3 – Distribuição das notas de proficiência em Física do Modelo Multifacetas de Rasch (escala M=50, DP=10)



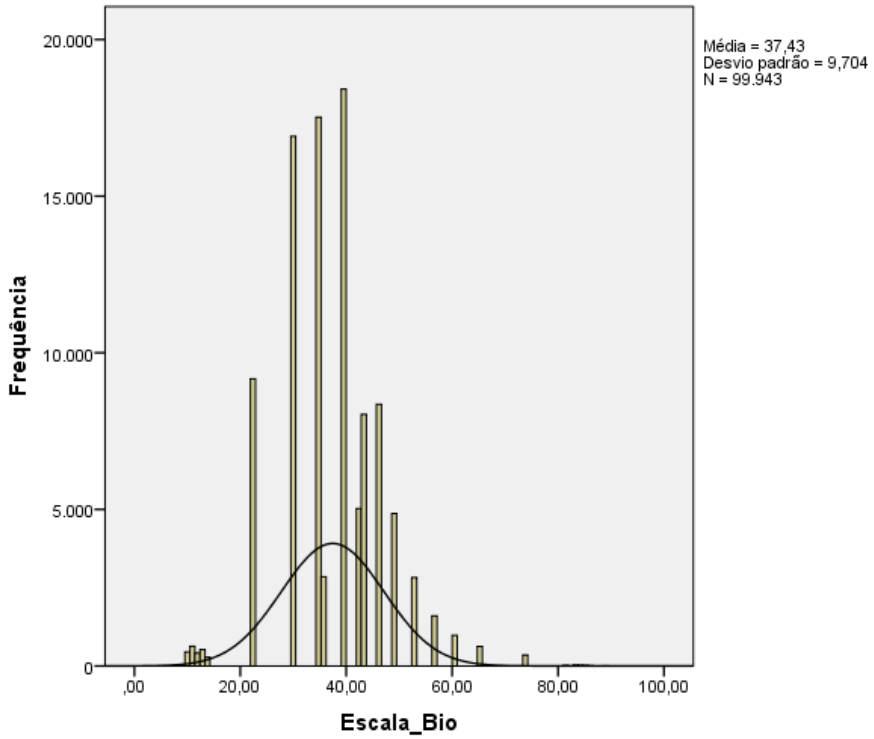
Fonte: Dados da pesquisa

Gráfico 4 – Distribuição das notas de proficiência em Química do Modelo Multifacetas de Rasch (escala M=50, DP=10)



Fonte: Dados da pesquisa

Gráfico 5 – Distribuição das notas de proficiência em Biologia do Modelo Multifacetas de Rasch (escala M=50, DP=10)



Fonte: Dados da pesquisa

De acordo com a definição de coeficiente de correlação, ele mede o grau de associação linear entre duas variáveis aleatórias X e Y. Quanto à análise da correlação entre as proficiências dos estudantes apresentada na Tabela 10, o seu comportamento foi conforme o esperado, ou seja, os dados apresentaram uma correlação elevada da Proficiência Geral com as demais proficiências e a Nota do ENEM. Já as proficiências das disciplinas tiveram baixa correlação entre si. Também neste caso, os dados apresentam comportamento que é aguardado de itens que avaliam conteúdos diferentes uns dos outros.

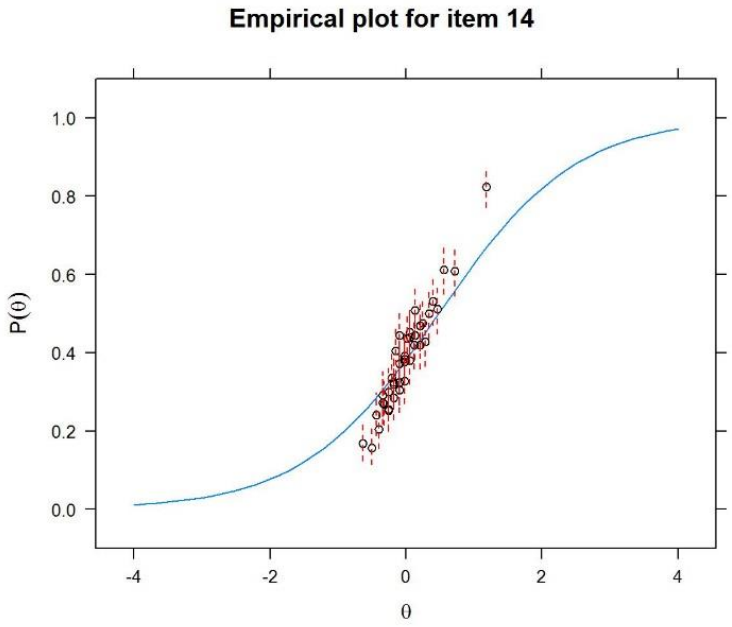
Tabela 10 – Correlação entre as proficiências do Modelo Multifacetas Rasch (Geral, Física, Química e Biologia) e a nota do ENEM.

	Proficiência Geral	Proficiência Física	Proficiência Química	Proficiência Biologia	Nota ENEM
Proficiência Geral	1	0,695**	0,696**	0,748**	0,873**
Proficiência Física		1	0,242**	0,294**	0,544**
Proficiência Química			1	0,301**	0,549**
Proficiência Biologia				1	0,772**
Nota ENEM					1

**. A correlação é significativa no nível 0,01 (2 extremidades).

Fonte: Dados da pesquisa.

Figura 8 – Gráfico do item com maior dificuldade na análise multifacetas de Rasch.



Fonte: Dados da pesquisa

Figura 9 – Questão com maior dificuldade (medição em *logitos*)

QUESTÃO 84

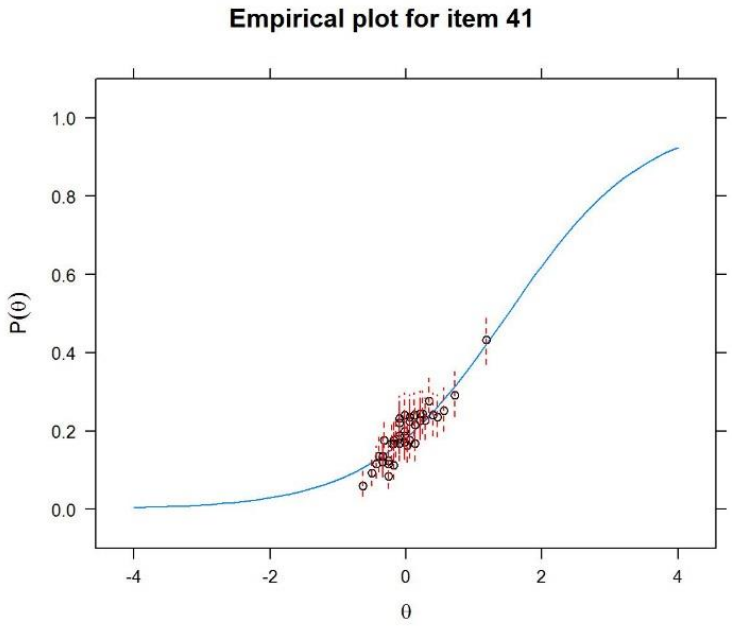
Num experimento, um professor deixa duas bandejas de mesma massa, uma de plástico e outra de alumínio, sobre a mesa do laboratório. Após algumas horas, ele pede aos alunos que avaliem a temperatura das duas bandejas, usando para isso o tato. Seus alunos afirmam, categoricamente, que a bandeja de alumínio encontra-se numa temperatura mais baixa. Intrigado, ele propõe uma segunda atividade, em que coloca um cubo de gelo sobre cada uma das bandejas, que estão em equilíbrio térmico com o ambiente, e os questiona em qual delas a taxa de derretimento do gelo será maior.

O aluno que responder corretamente ao questionamento do professor dirá que o derretimento ocorrerá

A) mais rapidamente na bandeja de alumínio, pois ela tem uma maior condutividade térmica que a de plástico.
B) mais rapidamente na bandeja de plástico, pois ela tem inicialmente uma temperatura mais alta que a de alumínio.
C) mais rapidamente na bandeja de plástico, pois ela tem uma maior capacidade térmica que a de alumínio.
D) mais rapidamente na bandeja de alumínio, pois ela tem um calor específico menor que a de plástico.
E) com a mesma rapidez nas duas bandejas, pois apresentarão a mesma variação de temperatura.

Fonte: Prova ENEM 2016 – Caderno Azul.

Figura 10 – Gráfico do item com menor dificuldade na análise multifacetas de Rasch.

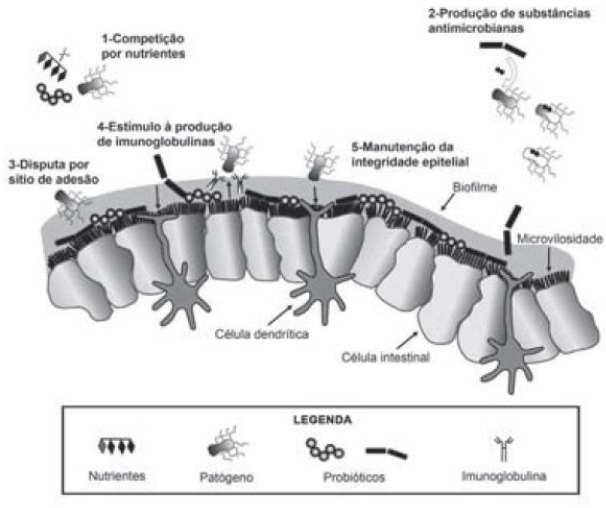


Fonte: Dados da pesquisa

Figura 11 – Questão com menor dificuldade (medição em *logitos*)

QUESTÃO 79

Vários métodos são empregados para prevenção de infecções por microrganismos. Dois desses métodos utilizam microrganismos vivos e são eles: as vacinas atenuadas, constituídas por patógenos avirulentos, e os probióticos que contêm bactérias benéficas. Na figura são apresentados cinco diferentes mecanismos de exclusão de patógenos pela ação dos probióticos no intestino de um animal.



McALLISTER, T. A. et al. Review: The use of direct fed microbials to mitigate pathogens and enhance production in cattle. **Can. J. Anim. Sci.**, jan. 2011 (adaptado).

Qual mecanismo de ação desses probióticos promove um efeito similar ao da vacina?

A 5
B 4
C 3
D 2
E 1

Fonte: Prova ENEM 2016 – Caderno Azul.

5. CONSIDERAÇÕES FINAIS

As avaliações em larga escala devem ter por objetivo diagnosticar problemas relevantes para a sociedade e para a educação, sendo relevante que tais problemas sejam solucionados. Estudos desenvolvidos pelos pesquisadores devem fazer uso de metodologias que possibilitem entender a demanda com o objetivo de solucionar os problemas; de tal forma que os estudos forneçam à sociedade resultados, tanto quanto possível, corretos, válidos e justos.

A utilização de novas técnicas que possibilitem maior abrangência de análise das provas e dos conteúdos abordados em avaliações educacionais deve ser sempre foco de pesquisas, numa perspectiva positiva da democratização dos resultados para os atores envolvidos no processo avaliativo. A necessidade de transparência e divulgação de resultados quantitativos e qualitativos para desenvolver a educação e as diversas formas de ensinar, gerando maior envolvimento e conhecimento sobre os aspectos pedagógicos que podem ser abordados nos relatórios das avaliações educacionais, devem ser contempladas com o emprego de modelos de análise como o proposto neste estudo.

Segundo Gatti (2013), a interpretação, análise e crítica do processo avaliativo requerem não só conhecimento e domínio de técnicas de medidas e modelos de avaliação, mas também conhecimento dos conteúdos, dos pressupostos, da realidade a que se reportam e dos objetivos de ensino. Sem isso, temos a não integração dos processos avaliativos com o quotidiano das decisões, por vezes, negando os resultados obtidos ou até, simplesmente, ignorando todo o processo de avaliação.

O Modelo Multifacetas de Rasch apresenta resultados diretos das variáveis apresentadas no modelo, possibilitando uma interpretação acessível. Tal fato é importante para a implementação de políticas públicas em diversos níveis. Por exemplo, na formação de professores com a especificação de disciplinas ou conteúdos nos quais os estudantes estão com maior dificuldade. Desta forma, a preparação para os professores torna-se mais direcionada, buscando uma maior qualidade na educação ofertada.

A utilização dos resultados de avaliações educacionais em larga escala para o planejamento de ações voltadas para os docentes, que são os viabilizadores da transmissão do conhecimento para os alunos, é tema relevante e deve sempre que possível ser abordado em pesquisas. No estudo, a utilização do MFRM para entender cadernos de questões que

contemplam múltiplas disciplinas traz informações sobre como se apresenta o desempenho dos estudantes em cada disciplina (subescala), tal dado pode ser trabalhado por uma análise pedagógica consistente que gere relatórios acessíveis aos docentes.

Bauer (2012) afirma que o discurso sobre uma melhor formação docente influenciando a qualidade de ensino tem aumentado sua importância nas discussões a respeito dos fatores que afetam a qualidade escolar e isso se traduz na forma em que são apresentados os resultados dos alunos nas avaliações educacionais em larga escala. Um bom desempenho dos estudantes se deveria às políticas de formação docente e consequentemente ao ensino de qualidade.

Com os resultados obtidos na aplicação do MFRM é possível verificar que as questões das disciplinas contidas no caderno de Ciências da natureza e suas tecnologias apresentaram nível de dificuldade alto. Historicamente o desempenho dos alunos na parte de Ciências da natureza é de mediano a baixo e as questões são consideradas difíceis pelos alunos. Mas, a possibilidade de analisar separadamente as disciplinas de Física, Química e Biologia traz luz para os educadores, com a possibilidade de avaliar conteúdos específicos.

O objetivo de apresentar outras metodologias de análise de avaliações educacionais vem como forma de agregar mais informação para o público interessado nos resultados. Ademais, uma metodologia alternativa não inviabiliza os resultados apresentados pela Teoria de Resposta ao Item, que é a principal fonte de informação disponibilizada pelos gestores das principais avaliações educacionais em larga escala brasileiras. O MFRM é uma proposta para acréscimo de resultados e geração de análises que relacionam diversas variáveis com vieses que podem ser medidos, fornecendo assim mais detalhes sobre as questões aplicadas.

O Modelo Multifacetas de Rasch mostrou-se eficiente para analisar dados que apresentam diversas fontes de variação, as quais são especificadas no modelo como as facetas. Neste estudo as informações sobre a qualidade e o ajuste do modelo construído de três facetas (habilidade, item e subescalas) foram todas consideradas adequadas, ou seja, apresentaram valores de referência (*Infit*, *Outfit*, Taxa de separação, entre outros) dentro dos limites aceitáveis. Temos um bom modelo para avaliar as questões apresentadas aos estudantes que realizaram o ENEM no ano de 2016, avaliadas a partir das subescalas. Um posterior aprofundamento na técnica faz-se necessário, tanto replicando a metodologia em outras bases de dados quanto disponibilizando os resultados para os educadores e demais sujeitos envolvidos no processo de ensino-aprendizagem com o objetivo de entender o comportamento do desempenho específico dos estudantes por disciplina, realizando análises pedagógicas dos resultados. Desta forma, buscar maior qualidade para o processo educacional.

É importante ressaltar que o modelo e o método descritos neste trabalho fornecem uma base teórica e prática consistente para análises de avaliações em que podem ser construídas subescalas, principalmente no nível individual dos elementos que formam o modelo especificado. As vantagens na utilização do MFRM nessas avaliações são inúmeras e muitos estudos podem ser desenvolvidos com foco nos mais variados elementos, detalhando os resultados e proporcionando maior conhecimento para grupos de atores importantes na avaliação educacional. Abre-se a possibilidade de integrar os resultados da avaliação de maneira formativa e somativa. A aplicação do MFRM no *FACETS* possibilita análises que utilizam índices quantitativos, gráficos e tabelas que ajudam a determinar indícios que permitam a verificação sobre a qualidade das avaliações e um maior detalhamento dos resultados de acordo com as facetas utilizadas no modelo.

Este trabalho traz como proposta demonstrar como o Modelo Multifacetas de Rasch pode colaborar para a determinação de análises detalhadas em provas construídas com diversos assuntos em um mesmo caderno ou bloco de questões. Estas provas fazem parte das mais variadas avaliações educacionais. Entretanto, os procedimentos e resultados estabelecidos neste trabalho, delimitam-se ao estudo das subescalas formadas pelas disciplinas do caderno de Ciências da natureza e suas tecnologias do ENEM 2016. Para uma consolidação da metodologia e sua aplicação em análises de avaliações, seria ideal replicar este estudo com dados provenientes de avaliações de outras avaliações e de outras áreas de conhecimento.

6. REFERÊNCIA BIBLIOGRÁFICA

ADÁNEZ, G. P. *Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement*. Psicothema, v. 23, n. 2, p. 233-238, 2011.

AFONSO, A. J. *Para uma conceitualização alternativa de accountability em educação*. Educação & Sociedade, v. 33, n. 119, p. 471-484, 2012.

ANDRADE, E. C. *Alternativa de política educacional para o Brasil: School Accountability*. Brazilian Journal of Political Economy, v. 29, n. 4, p. 454-472, 2009.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da Resposta ao Item: Conceitos e Aplicações*. São Paulo: Associação Brasileira de Estatística, 2000.

ANDRICH, D. *Rasch models for measurement*. Sage, 1988.

BAKER, F. B. *The basics of item response theory*. Retirado em 15/07/2018 do Ericae (Eric clearinghouse assessment and evaluation), http//ericae.net/ftlib.htm. 2001.

BARBETTA, P. A.; TREVISAN, L. M.; TAVARES, H.; AZEVEDO, T. C. A. M. *Aplicação da Teoria da Resposta ao Item uni e multidimensional*. Estudos em Avaliação Educacional, v. 25, n. 57, p. 280-302, 2014.

BAUER, A. *É possível relacionar avaliação discente e formação de professores? A experiência de São Paulo*. Educação em Revista, 28(2), p. 61-82. 2012.

_____. *Estudos sobre Sistemas de Avaliação Educacional no Brasil: um retrato em preto e branco.* Revista @mbienteeducação, [S.l.], v. 5, n. 1, p. 7-31, dez. 2017. ISSN 1982-8632. Disponível em: http://publicacoes.unicid.edu.br/index.php/ambienteeducacao/article/view/115. Acesso em: 7 mai. 2018.

BRASIL. *Lei de Diretrizes e Bases da Educação Nacional*. Lei nº 9.394, 20 de dezembro de 1996. Brasília, 1996. Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/l9394.htm. Acesso em: 15 mai. 2018.

_____. Ministério da Educação (MEC). Secretaria de Educação Média e Tecnológica. *Exame Nacional do Ensino Médio-ENEM: documento básico*. Brasília: INEP, 2002.

_____. Lei Nº 10.861, de 14 de abril de 2004. - *Institui o Sistema Nacional de Avaliação da Educação Superior – SINAES – e dá outras providências*. Brasília, 2004. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.861.htm>. Acesso em: 10 jun. 2018.

_____. Ministério da Educação (MEC). Secretaria de Educação Média e Tecnológica. *Matriz de referência para o ENEM*. Brasília: INEP, 2009.

_____. Ministério da Educação (MEC). Instituto Nacional de estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Portaria nº 482, de 7 de junho de 2013. *Dispõe sobre o Sistema de Avaliação da Educação Básica*. Brasília, 2013a. Disponível em: < http://download.inep.gov.br/educacao_basica/prova_brasil_saeb/legislacao/2013/portaria_n_482_07062013_mec_inep_saeb.pdf>. Acesso em: 02 jun. 2018.

_____. Ministério da Educação (MEC). Instituto Nacional de estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Portaria nº 304, de 21 de junho de 2013. *Dispõe sobre o*

Sistema de Avaliação da Educação Básica – SAEB. Brasília, 2013b. Disponível em: http://download.inep.gov.br/educacao_basica/prova_brasil_saeb/legislacao/2013/portaria_n304_saeb_RevFC.pdf.. Acesso em: 02 jun. 2018.

BONAMINO, A.; SOUSA, S. Z. *Three generations of assessments of basic education in Brazil: Interfaces with the curriculum in/of the school*. Educação e Pesquisa, v. 38, n. 2, p. 373-388, 2012.

BRINTHAUPT, T. M.; KANG, M. *Many-faceted rasch calibration: An example using the self-talk scale*. Assessment, v. 21, n. 2, p. 241-249, 2014.

CHACHAMOVICK, E. *Teoria da Resposta ao Item: Aplicação do Modelo de Rasch em Desenvolvimento e Validação de Instrumentos em Saúde Mental*. 2007, 288 f. Tese (Doutorado em Ciências Médicas) - Programa de Pós-Graduação em Ciências Médicas: Psiquiatria, Faculdade de Medicina Universidade Federal do Rio Grande do Sul. Porto Alegre, 2007.

CHACHAMOVICH, E.; FLECK, M. P; TRENTINI, C. & POWER, M. *Brazilian WHOQOL-OLD Module version: a Rasch analysis of a new instrument*. Rev. Saúde Pública, São Paulo v. 42, n.2, p.308-316, Abr. 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102008000200017&lng=en&nrm=iso>. Acesso em: 02 Jan. 2019. http://dx.doi.org/10.1590/S0034-89102008000200017.

COELHO, M. I. M. *Vinte anos de avaliação da educação básica no Brasil: aprendizagens e desafios*. Ensaio: Avaliação e políticas públicas em Educação, v. 16, n. 59, 2008.

COUTO, G.; PRIMI, R. *Teoria de resposta ao item (TRI): conceitos elementares dos modelos para itens dicotômicos*. Boletim de Psicologia, v. 61, n. 134, p. 1-15, 2011.

DE AYALA, R. J. *The theory and practice of item response theory*. Guilford Publications, 2009.

DE BRITO, M. R. F. *O SINAES e o ENADE: da concepção à implantação*. Avaliação: Revista da Avaliação da Educação Superior, v. 13, n. 3, 2008.

ECKES, T. *Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis*. Language Assessment Quarterly: An International Journal, v. 2, n. 3, p. 197-221, 2005.

_____. *Rater types in writing performance assessments: A classification approach to rater variability*. Language Testing, v. 25, n. 2, p. 155-185, 2008.

_____. Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages:* Learning, teaching, assessment, p. 1-52, 2009.

_____. *Introduction to Many-Facet Rasch Measurement: Analyzing and evaluating rater-mediated assessment*. Frankfurt: Peter Lang, 2011.

ENGELHARD, G. *Examining rater errors in the assessment of written composition with a many-faceted Rasch model*. Journal of Educational Measurement, v. 31, n. 2, p. 93-112, 1994.

GATTI, B. A. *A formação de professores: seus desafios, a pesquisa e seus contornos sociais*. Educação e filosofia, v. 17, n. 34, p. 241-252, 2003.

_____. *Avaliação de sistemas educacionais no Brasil*. Revista de ciências da educação, v. 9, p. 7-18, 2009.

_____. *Testes e avaliações do ensino no Brasil*. Educação e Seleção, (16), 33-42. 2013.

_____. *Avaliação: contexto, história e perspectivas*. Olh@ res, v. 2, n. 1, p. 8-26, 2014.

GIBBONS, R. D.; HEDEKER, D. R. *Full-information item bi-factor analysis*. Psychometrika, v. 57, n. 3, p. 423-436, 1992.

GOLAFSHANI, N. *Understanding reliability and validity in qualitative research*. The qualitative report, v. 8, n. 4, p. 597-606, 2003.

GOODWIN, S. *A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes*. Assessing Writing, v. 30, p. 21-31, 2016.

HORTA NETO, J. L. *Um olhar retrospectivo sobre a avaliação externa no Brasil: das primeiras medições em educação até o SAEB de 2005*. Revista Iberoamericana de Educación, Madrid, v. 42, p. 1-14, 2007. Disponível em:<http://www.rieoei.org/deloslectores/1533Horta.pdf>. Acesso em: 15 mai. 2018.

KARINO, Camila Akemi. Avaliação da igualdade, equidade e eficácia no sistema educacional brasileiro. Tese de Doutorado, Brasília, 2016.

KLEIN, R. *Alguns aspectos da teoria de resposta ao item relativos à estimação das proficiências*. Ensaio: avaliação e políticas públicas em educação, Rio de Janeiro, v. 21, n. 78, p. 35-55, 2013.

KLEIN, R.; FONTANIVE, N. *Uma nova maneira de avaliar as competências escritoras na redação do ENEM*. Ensaio: Avaliação e Políticas Públicas em Educação, v. 17, n. 65, 2009.

LINACRE, J. M. *Many-facet Rasch measurement*, 2nd ed. Chicago: MESA Press, 1994.

_____. *Many-facet Rasch measurement: Facets tutorial*. Retrieved December, v. 19, p. 2018, 2012.

_____. *Facets Rasch measurement computer program* (Version 3.71.4) [Computer software]. Chicago: Winsteps.com. 2014.

LINACRE, J. M.; WRIGHT, B. D. *Construction of measures from many-facet data*. Journal of Applied Measurement, 2002.

MACHADO, P. H. A.; LIMA, E. G. S. *O ENEM no contexto das políticas para o Ensino Médio*.Perspectiva, v. 32, n. 1, p. 355-373, 2014.

MARTINS, G. A. *Sobre confiabillidade e validade*. Revista Brasileira de Gestão de Negócios, v. 8, n. 20, 2006.

MESSICK, S. *Validity*. In: LINN, R.(Ed.). Educational Measurement. 3nd ed. New York: Macmillan, p. 13-103, 1989.

MYFORD, C. M.; WOLFE, E. W. *Monitoring sources of variability within the Test of Spoken English assessment system*. ETS Research Report Series, v. 2000, n. 1, 2000.

_____. *Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II*. journal of applied measurement, v. 5, n. 2, p. 189-227, 2004.

OECD. PISA 2015 Technical Report: *Chapter 12: Scaling e Chapter 15: Proficiency Scale Construction*. OECD Publishing, Paris, 2017. Disponível em http://www.oecd.org/pisa/data/2015-technical-report/. Acesso em: 23 set. 2018.

OLIVEIRA, M. A. M.; ROCHA, G. *Avaliação em larga escala no Brasil nos primeiros anos do Ensino Fundamental*. Porto Alegre: ANPAE, [sd]. UFRGS/FACED/PPGEDU, 2007.

PALLANT, J. F.; TENNANT, A. *An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS)*. British Journal of Clinical

Psychology, v. 46, n. 1, p. 1-18, 2007. Disponível em: https://onlinelibrary.wiley.com/doi/epdf/10.1348/014466506X96931. Acesso em: 15 out. 2018.

PASQUALI, L. *Princípios de elaboração de escalas psicológicas*. Rev Psiquiatria Clínica 1998; 25(5): 206-13.

_____. *Validade dos testes psicológicos: será possível reencontrar o caminho*. Psicologia: teoria e pesquisa, v. 23, p. 99-107, 2007.

_____. *Psychometrics*. Revista da Escola de Enfermagem da USP, v. 43, n. SPE, p. 992-999, 2009.

PASQUALI, L; PRIMI, R. *Fundamentos da teoria da resposta ao item: TRI*. Avaliação Psicologica: Interamerican Journal of Psychological Assessment, v. 2, n. 2, p. 99-110, 2003.

PARRA-LÓPEZ, E.; OREJA-RODRÍGUEZ, J. R. *Evaluation of the competiveness of tourist zones of an island destination: An application of a Many-Facet Rasch Model (MFRM)*. Journal of Destination Marketing & Management, v. 3, n. 2, p. 114-121, 2014.

REISE, S. P.; MORIZOT, J.; HAYS, R. D. *The role of the bifactor model in resolving dimensionality issues in health outcomes measures*. Quality of Life Research, v. 16, n. 1, p. 19-31, 2007.

STADLER, J. P.; HUSSEIN, F. R. G. S. *O perfil das questões de ciências naturais do novo Enem: interdisciplinaridade ou contextualização*, Ciência & Educação, v. 23, n. 2, p. 391-402, 2017.

SOARES, J. F. *Qualidade e equidade na educação básica brasileira: fatos e possibilidades*. In: BROCK, Colin; SCHWARTZMAN, Simon. (Ed.). Os desafios da educação no Brasil. Rio de Janeiro: Nova Fronteira, p. 87-114, 2005.

SOARES NETO, J. J.; JESUS, G. R.; KARINO, C. A.; ANDRADE, D. F. *Uma escala para medir a infraestrutura escolar*. Est. Aval. Educ., São Paulo, v. 24, n. 54, p. 78-99, jan./abr. 2013.

STEMLER, S. E. *A comparison of consensus, consistency and measurement approaches to estimating interrater reliability*. Practical Assessment, Research and Evaluation, v. 9, n. 4, p. 1-11, 2004.

TATTERSALL, K. *A brief history of policies, practices and issues relating to comparability*. In Techniques for monitoring the comparability of examination standards, ed. P. Newton, J. Baird, H. Goldstein, H. Patrick, P. Tymms, 43-96. London: QCA, 2007.

TOFFOLI, S. F. L. *Avaliações em larga escala com itens de respostas construídas no contexto do modelo Multifacetas de Rasch*. 2015. 313 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2015.

TOFFOLI, S. F. L.; SIMON, C. V. B. *A utilização do modelo multifacetas de Rasch na análise das influências dos avaliadores sobre as avaliações com itens abertos*. Ensaio: Avaliação e Políticas Públicas em Educação, n. 101, p. 1303-1323, 2018.

VIANNA, H. M. *Avaliações nacionais em larga escala: análises e propostas*. Estudos em avaliação educacional, n. 27, p. 41-76, 2003.

_____. *Avaliação educacional: uma perspectiva histórica*. Estudos em Avaliação Educacional, v. 25, n. 60, p. 14-35, 2014.

VIEIRA, N. N.; BARBETTA, P. A. *As provas das quatro áreas do ENEM vista como prova única na óptica de modelos da Teoria da Resposta ao Item uni e multidimensional*. In: CONBRATRI Congresso Brasileiro de Teoria da Resposta ao Item. p. 107-108. 2016.

VITÓRIA, F.; ALMEIDA, L. S.; PRIMI, R. *Unidimensionalidade em testes psicológicos: conceito, estratégias e dificuldades na sua avaliação*. Psic: revista da Vetor Editora, v. 7, n. 1, p. 01-07, 2006.

UNESCO. *Accountability in education: meeting our commitments – Global Education Monitoring Report 2017/18*. França, 2017.

WANG, N.; STAHL, J. *Obtaining content weights for test specifications from job analysis task surveys: An application of the many-facets Rasch model*. International Journal of Testing, v. 12, n. 4, p. 299-320, 2012.

WERLE, F. O. C. *Políticas de avaliação em larga escala na educação básica: do controle de resultados à intervenção nos processos de operacionalização do ensino*. Ensaio: Avaliação e Políticas Públicas em Educação, v. 19, n. 73, 2011.

WRIGHT, B. D.; STONE, M. H. *Measurement essentials*. Wilmington, DE: Wide Range. 1999.

7. ANEXOS

Anexo 1 – Alguns Resultados do *FACETS*

```
Data specification
Facets = 3
Delements = N
Dvalues = 3,2,$GROUP, ; fixed values for data facets
Non-centered = 1
Positive =  1
Labels =
 1,Children ; (elements = 100000)
 2,items ; (elements = 45, highest group = 3)
 3,subscales, D  ; (elements = 3)
Model = ?,?,?,D,1

; Output description
Arrange tables in order = mN,fN,N
Bias/Interaction direction = plus ; ability, easiness, leniency: higher score = positive logit
Fair score = Mean
Pt-biserial = Yes
Heading lines in output data files = Y
Scorefile = ENEMSC
Barchart = Yes
Total score for elements = Yes
T3onscreen show only one line on screen iteration report = Y
T4MAX maximum number of unexpected observations reported in Table 4 = 100
T8NBC show table 8 numbers-barcharts-curves = NBC
Unexpected observations reported if standardized residual >= 3
Usort unexpected observations sort order = u
Vertical ruler definitions = 1*,2*,2A,3A
WHexact - Wilson-Hilferty standardization = Y
Yardstick: dimensions in rulers = 112,4

; Convergence control
Convergence = .5, .01
Iterations (maximum) = 0 ; unlimited
Xtreme scores adjusted by = .3, .5  ;(estimation, bias)

ENEM CN 100000 ALUNOS 20/01/2019 12:11:25
Table 2. Data Summary Report.

Assigning models to Data= "DATA_ENEM_NS.sav"
Total lines in data file = 100005
Total data lines = 100000
Responses matched to model: ?,?,?,D,1 = 4500000
   Total non-blank responses found = 4500000
```

Number of blank lines = 4
Valid responses used for estimation = 4500000

ENEM CN 100000 ALUNOS
Table 3. Iteration Report.

Iteration		Max. Score Residual Elements	%	Categories	Max. Logit Change Elements	Steps
PROX	1	Recount required			-4.4998	
PROX	2				-.1264	
JMLE	3	-243.9238	-.4	.0000	-.2318	.0000
JMLE	4	-133.7520	-.2	.0000	-.1166	.0000
JMLE	5	-73.5313	-.1	.0000	-.0586	.0000
JMLE	6	-40.5098	.0	.0000	-.0295	.0000
JMLE	7	-22.3535	.0	.0000	-.0149	.0000
JMLE	8	-12.3516	.0	.0000	-.0075	.0000
JMLE	9	-6.8320	.0	.0000	-.0038	.0000
JMLE	10	-3.7852	.0	.0000	-.0019	.0000
JMLE	11	-2.0977	.0	.0000	-.0010	.0000
JMLE	12	-1.1641	.0	.0000	-.0005	.0000
JMLE	13	-.6465	.0	.0000	-.0002	.0000
JMLE	14	-.3594	.0	.0000	-.0001	.0000

Subset connection O.K.

ENEM CN 100000 ALUNOS
Table 4. Unexpected Responses - appears after Table 8.

ENEM CN 100000 ALUNOS
Table 5. Measurable Data Summary.

Cat	Score	Exp.	Resd	StRes	
.26	.26	.26	.00	.00	Mean (Cnt: 4497435)
.44	.44	.13	.42	1.00	S.D. (Population)
.44	.44	.13	.42	1.00	S.D. (Sample)

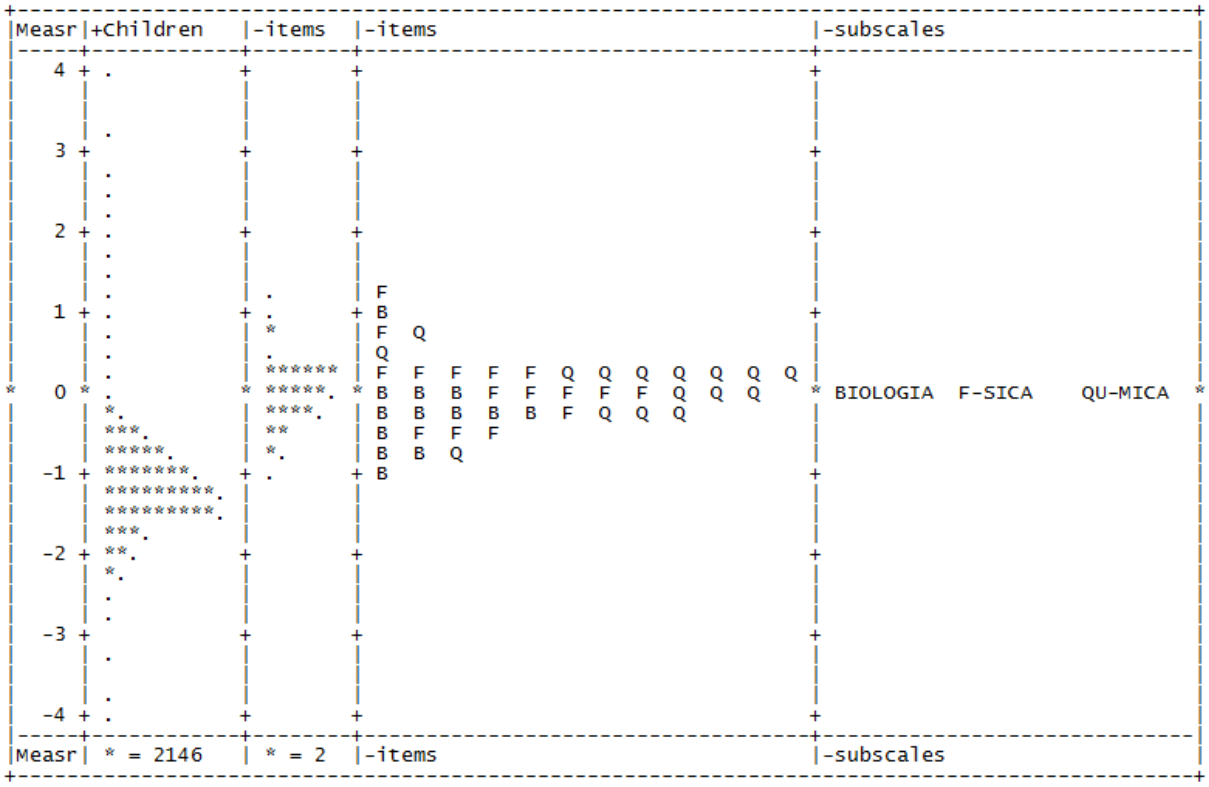
Data log-likelihood chi-square = 4739113.0000
Approximate degrees of freedom = 4397446
Chi-square significance prob. = .0000

	Count	Mean	S.D.	Params
Responses used for estimation	= 4497435	0,26	0,44	99989
Responses in one extreme score	= 2565	0,00	0,00	57
All Responses	= 4500000	0,26	0,44	100046

Count of measurable responses = 4500000
Raw-score variance of observations = 0,19 100.00%
Variance explained by Rasch measures = 0,02 8,92%
Variance of residuals = 0,17 91,08%

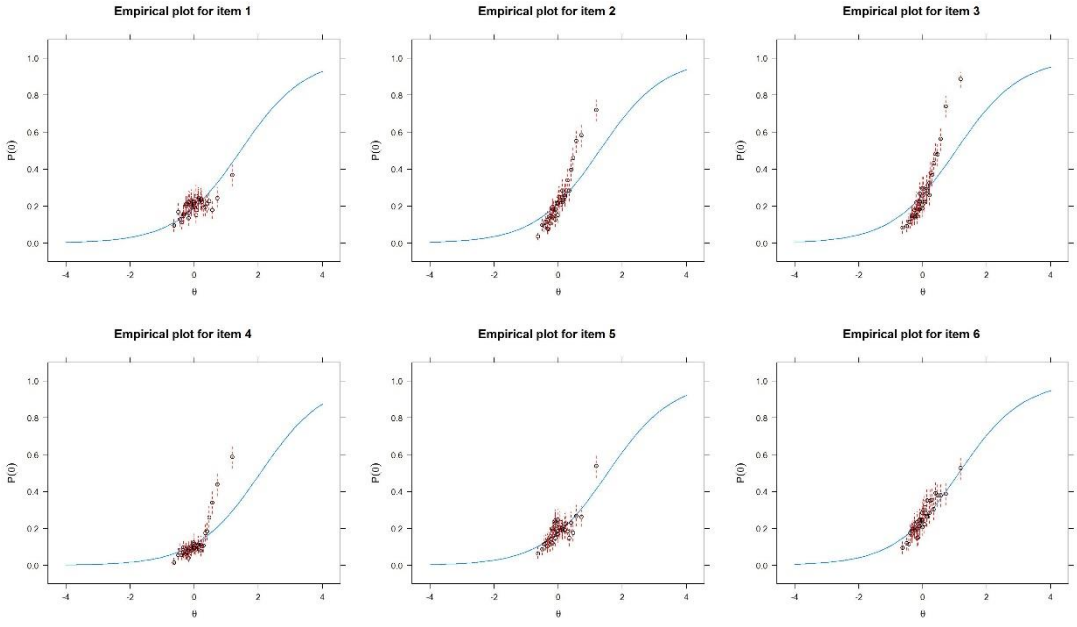
ENEM CN 100000 ALUNOS 20/01/2019 12:11:25
Table 6.0 All Facet Vertical "Rulers".

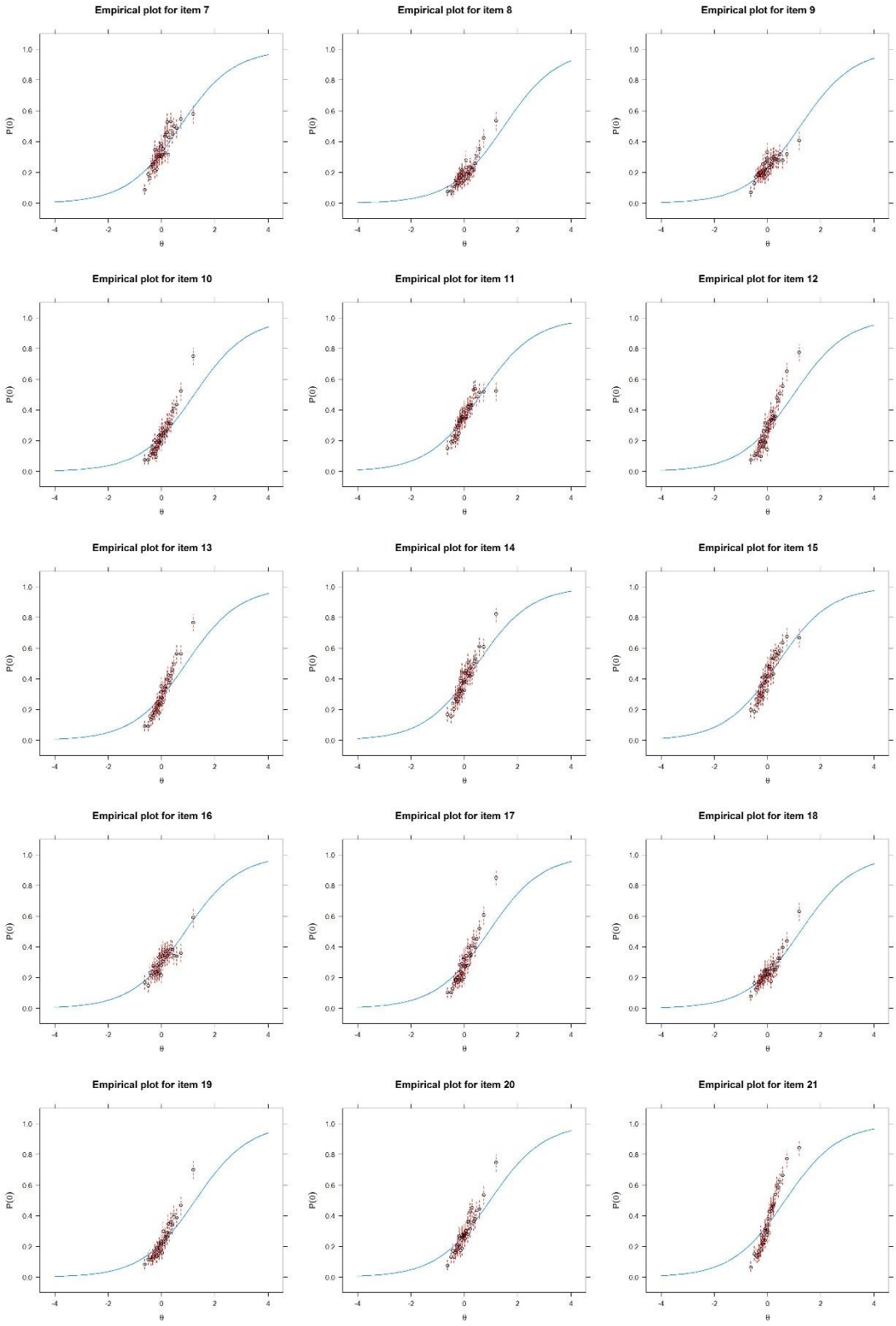
Vertical = (1*,2*,2A,3A,S) Yardstick (columns lines low high extreme)= 112,4,-4,4,End

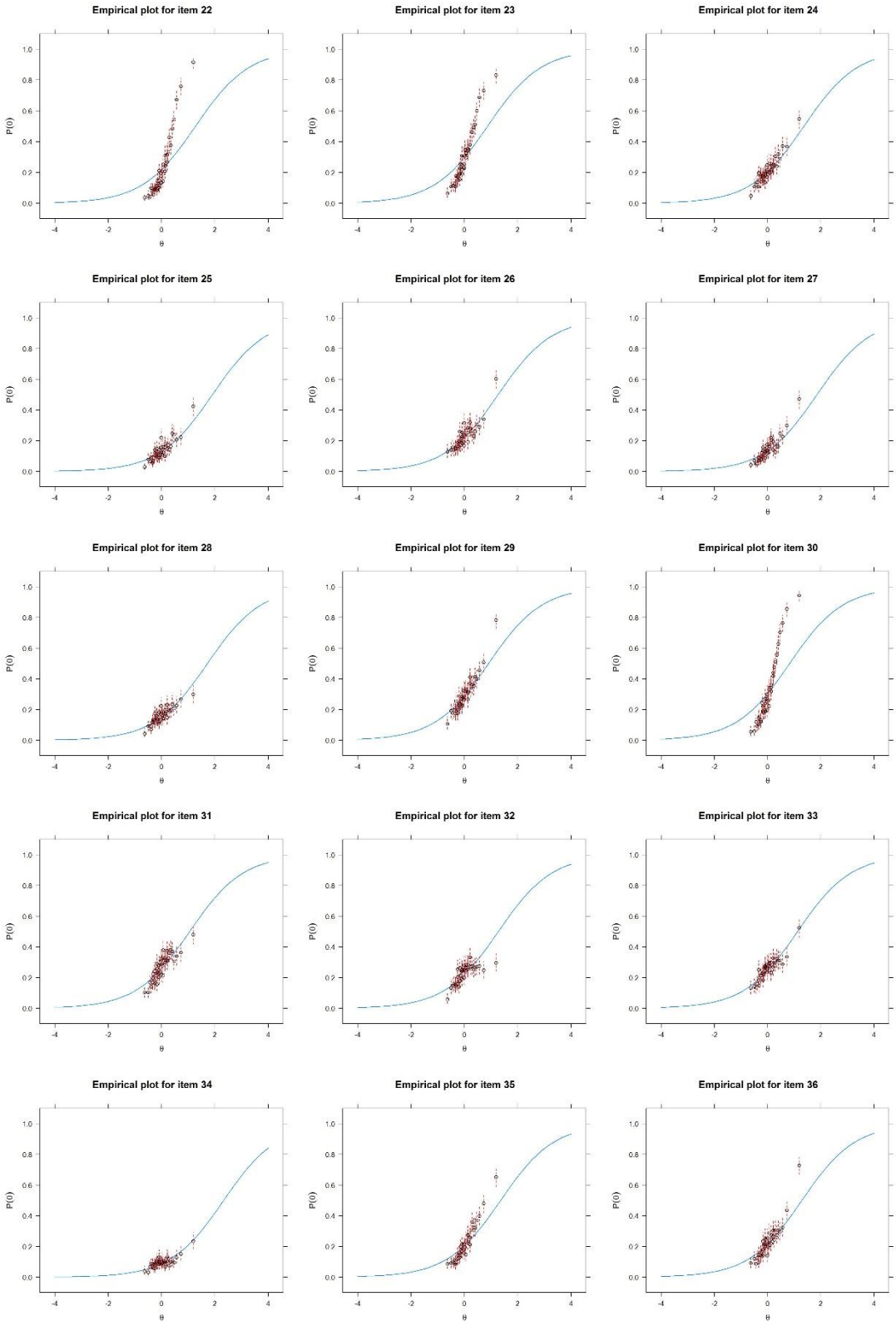


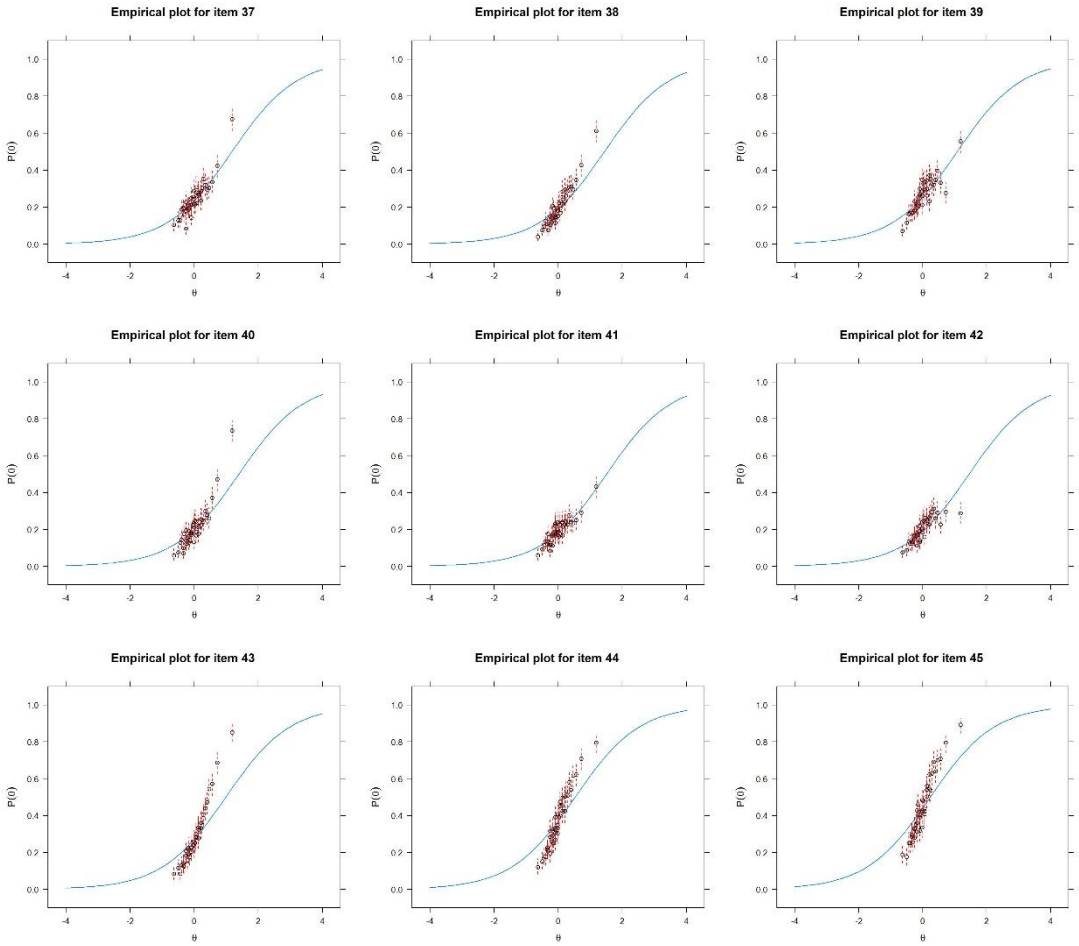
S.1: Model = ?,?,?,D

Gráficos da dificuldade das questões




Empirical plot for item 7
Empirical plot for item 8
Empirical plot for item 9
Empirical plot for item 10
Empirical plot for item 11
Empirical plot for item 12
Empirical plot for item 13
Empirical plot for item 14
Empirical plot for item 15
Empirical plot for item 16
Empirical plot for item 17
Empirical plot for item 18
Empirical plot for item 19
Empirical plot for item 20
Empirical plot for item 21
P(θ)
θ


Empirical plot for item 22
Empirical plot for item 23
Empirical plot for item 24
Empirical plot for item 25
Empirical plot for item 26
Empirical plot for item 27
Empirical plot for item 28
Empirical plot for item 29
Empirical plot for item 30
Empirical plot for item 31
Empirical plot for item 32
Empirical plot for item 33
Empirical plot for item 34
Empirical plot for item 35
Empirical plot for item 36
P(θ)
θ


Empirical plot for item 37
Empirical plot for item 38
Empirical plot for item 39
Empirical plot for item 40
Empirical plot for item 41
Empirical plot for item 42
Empirical plot for item 43
Empirical plot for item 44
Empirical plot for item 45
P(θ)
θ

Anexo 2 – Alguns Resultado da Análise Fatorial

```
R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

> library(mirt)
The following object is masked from ‘package:mirt’:

> library(psych)
Attaching package: ‘psych’
The following object is masked from ‘package:ltm’:
    factor.scores

> library(psychometric)
Carregando pacotes exigidos: multilevel
Carregando pacotes exigidos: nlme

Attaching package: ‘nlme’
The following object is masked from ‘package:mirt’:
    fixef
Attaching package: ‘psychometric’
The following object is masked from ‘package:psych’:
    alpha

> summary(dados)
      P1                P2                P3                P4
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.3632   Mean   :0.2015   Mean   :0.2696   Mean   :0.2155
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      P5               P6                P7                P8
 Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.226   Mean   :0.2475   Mean   :0.2256   Mean   :0.1353
 3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      P9            P10            P11           P12
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.00   Median :0.0000
 Mean   :0.2788   Mean   :0.2454   Mean   :0.26   Mean   :0.2473
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.00   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00   Max.   :1.0000
     P13            P14             P15            P16
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000
 Mean   :0.3399   Mean   :0.09968   Mean   :0.2225   Mean   :0.3218
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
     P17            P18            P19            P20
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
```

```
 Mean   :0.3022   Mean   :0.2812   Mean   :0.2243   Mean   :0.1993
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
     P21              P22              P23              P24
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.2478   Mean   :0.3919   Mean   :0.2638   Mean   :0.2044
 3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
     P25              P26              P27              P28
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.2387   Mean   :0.2158   Mean   :0.1597   Mean   :0.1525
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
     P29              P30              P31              P32
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.2002   Mean   :0.2004   Mean   :0.1916   Mean   :0.3004
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
     P33              P34              P35              P36
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.3075   Mean   :0.2388   Mean   :0.2874   Mean   :0.2758
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
     P37              P38              P39              P40
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000
 Mean   :0.3004   Mean   :0.3751   Mean   :0.4101   Mean   :0.232
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
     P41              P42              P43              P44
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.4527   Mean   :0.1187   Mean   :0.2837   Mean   :0.3521
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
     P45
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.2324
 3rd Qu.:0.0000
 Max.   :1.0000
```

```
> dados_cor <- tetrachoric(dados, y=NULL, correct=.5, smooth=TRUE, global=TRUE, weight=NULL, na.rm=TRUE, delete=TRUE)

> dados_cor
Call: tetrachoric(x = dados, y = NULL, correct = 0.5, smooth = TRUE,
   global = TRUE, weight = NULL, na.rm = TRUE, delete = TRUE)
tetrachoric correlation
```

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P1	1.00										
P2	0.01	1.00									
P3	0.05	0.01	1.00								
P4	0.04	0.01	-0.03	1.00							
P5	0.05	-0.04	0.04	0.10	1.00						
P6	0.04	0.01	0.00	0.05	0.05	1.00					
P7	0.01	0.03	0.01	0.00	-0.01	-0.01	1.00				
P8	0.01	0.02	0.02	0.07	0.08	0.07	0.01	1.00			
P9	0.05	0.00	0.05	0.09	0.19	0.04	-0.01	0.08	1.00		
P10	0.04	0.00	0.04	0.07	0.20	0.07	-0.01	0.08	0.16	1.00	
P11	0.03	0.02	0.07	0.07	0.04	0.00	0.00	0.04	0.06	0.04	1.00
P12	0.05	0.01	0.03	0.03	0.06	0.06	0.00	0.07	0.06	0.06	0.04
P13	0.03	0.00	0.03	0.02	0.10	0.02	0.00	0.02	0.07	0.08	0.02
P14	-0.01	0.02	-0.01	0.02	0.06	0.04	-0.03	0.05	0.01	0.06	0.02
P15	0.04	-0.01	0.05	0.07	0.17	0.03	-0.01	0.09	0.13	0.11	0.04
P16	0.06	-0.07	0.07	0.08	0.30	0.03	-0.01	0.08	0.27	0.23	0.05
P17	0.09	-0.03	0.06	0.07	0.23	0.02	-0.01	0.08	0.19	0.16	0.04
P18	0.04	-0.05	0.04	0.08	0.20	0.04	0.00	0.07	0.16	0.14	0.04
P19	0.08	0.06	0.04	0.07	0.11	0.06	0.01	0.09	0.10	0.09	0.05
P20	0.02	0.00	0.03	0.04	0.10	0.02	0.01	0.05	0.12	0.07	0.04
P21	0.01	0.01	0.03	0.03	0.08	0.01	0.01	0.02	0.09	0.08	0.03
P22	0.04	0.00	0.04	0.02	0.08	0.00	0.02	0.02	0.11	0.08	0.04
P23	0.05	0.04	0.07	0.03	0.08	0.00	0.03	0.02	0.07	0.04	0.03
P24	0.05	0.00	0.05	0.05	0.12	0.06	0.01	0.05	0.17	0.12	0.05
P25	0.00	0.01	0.01	0.05	0.06	0.04	0.00	0.02	0.05	0.07	0.03
P26	0.04	0.01	0.04	0.07	0.15	0.07	0.00	0.12	0.16	0.13	0.06
P27	-0.01	-0.01	0.00	0.03	0.08	0.04	-0.02	0.03	0.03	0.04	0.01
P28	0.03	0.00	0.05	0.08	0.13	0.02	-0.01	0.08	0.13	0.11	0.06
P29	0.02	0.02	0.01	0.02	0.06	0.02	0.00	0.05	0.03	0.05	0.03
P30	0.01	0.00	0.02	0.03	0.02	-0.01	0.02	0.03	0.04	0.00	0.04
P31	0.01	0.02	0.02	0.06	0.06	0.01	0.00	0.04	0.09	0.05	0.05
P32	0.06	0.01	0.05	0.08	0.07	0.03	-0.01	0.04	0.11	0.07	0.05
P33	0.06	-0.01	0.05	0.06	0.16	0.03	0.01	0.07	0.14	0.11	0.04
P34	0.05	0.02	0.04	0.08	0.10	0.02	0.01	0.05	0.09	0.07	0.05
P35	0.05	-0.03	0.03	0.05	0.16	0.03	0.00	0.06	0.15	0.12	0.03
P36	0.05	-0.01	0.05	0.07	0.20	0.02	0.01	0.06	0.20	0.15	0.05
P37	0.01	0.01	0.01	0.01	0.04	0.01	0.01	0.01	0.05	0.03	0.03
P38	0.05	-0.02	0.04	0.06	0.14	0.02	0.00	0.03	0.15	0.10	0.05
P39	0.05	-0.01	0.04	0.03	0.07	0.01	0.00	0.01	0.09	0.08	0.03
P40	0.04	0.01	0.04	0.05	0.12	0.02	0.01	0.05	0.15	0.11	0.04
P41	0.06	-0.03	0.06	0.05	0.15	0.02	0.01	0.04	0.16	0.12	0.05
P42	0.06	0.02	0.06	0.11	0.23	0.05	0.00	0.11	0.23	0.17	0.07
P43	0.05	0.00	0.05	0.06	0.11	0.03	0.00	0.06	0.12	0.11	0.04
P44	0.06	-0.03	0.05	0.05	0.18	0.02	0.01	0.04	0.19	0.14	0.04
P45	0.09	-0.05	0.09	0.11	0.32	0.04	0.01	0.11	0.31	0.23	0.07

	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22
P12	1.00										
P13	0.04	1.00									
P14	0.02	0.00	1.00								
P15	0.06	0.06	0.02	1.00							
P16	0.08	0.13	-0.03	0.20	1.00						
P17	0.08	0.09	-0.01	0.14	0.29	1.00					
P18	0.04	0.10	0.00	0.15	0.27	0.21	1.00				
P19	0.10	0.04	0.05	0.09	0.10	0.08	0.08	1.00			
P20	0.04	0.02	0.01	0.06	0.11	0.10	0.06	0.08	1.00		
P21	0.01	0.03	0.02	0.04	0.13	0.08	0.07	0.06	0.07	1.00	
P22	0.03	0.04	-0.01	0.04	0.14	0.11	0.07	0.06	0.09	0.08	1.00
P23	0.03	0.02	0.00	0.04	0.07	0.06	0.05	0.05	0.04	0.04	0.04
P24	0.06	0.05	0.01	0.08	0.19	0.16	0.12	0.08	0.09	0.08	0.11

P25 0.05 0.03 0.01 0.05 0.08 0.06 0.04 0.07 0.04 0.05 0.04
P26 0.09 0.05 0.06 0.11 0.13 0.12 0.11 0.14 0.08 0.05 0.06
P27 0.02 0.03 0.07 0.03 0.04 0.03 0.05 0.04 0.02 0.01 -0.01
P28 0.04 0.04 0.02 0.09 0.15 0.11 0.12 0.08 0.10 0.05 0.05
P29 0.03 0.03 0.03 0.05 0.05 0.05 0.06 0.07 0.02 0.03 0.03
P30 0.02 0.00 0.00 0.02 0.03 0.02 0.03 0.03 0.03 0.00 0.02
P31 0.07 0.00 0.03 0.05 0.07 0.05 0.06 0.05 0.05 0.05 0.06
P32 0.04 0.04 0.02 0.07 0.09 0.08 0.06 0.09 0.08 0.05 0.05
P33 0.06 0.06 0.01 0.12 0.18 0.15 0.13 0.11 0.07 0.06 0.09
P34 0.04 0.04 0.02 0.07 0.11 0.09 0.08 0.07 0.06 0.04 0.06
P35 0.04 0.09 0.02 0.09 0.22 0.17 0.18 0.08 0.07 0.08 0.08
P36 0.07 0.08 0.00 0.13 0.26 0.20 0.20 0.09 0.10 0.10 0.12
P37 0.00 0.01 0.00 0.04 0.06 0.04 0.02 0.04 0.03 0.05 0.03
P38 0.04 0.06 -0.02 0.09 0.20 0.17 0.19 0.06 0.07 0.07 0.13
P39 0.04 0.05 -0.01 0.06 0.12 0.10 0.08 0.04 0.05 0.04 0.06
P40 0.05 0.06 0.00 0.10 0.18 0.13 0.12 0.07 0.07 0.07 0.08
P41 0.04 0.08 -0.04 0.10 0.24 0.19 0.15 0.08 0.08 0.07 0.10
P42 0.07 0.09 0.07 0.17 0.26 0.20 0.17 0.13 0.14 0.10 0.11
P43 0.08 0.06 0.01 0.08 0.16 0.12 0.08 0.08 0.06 0.07 0.08
P44 0.06 0.07 -0.03 0.17 0.29 0.21 0.16 0.07 0.08 0.09 0.12
P45 0.10 0.12 0.02 0.21 0.40 0.33 0.30 0.14 0.14 0.11 0.16
  P23 P24 P25 P26 P27 P28 P29 P30 P31 P32 P33
P23 1.00
P24 0.06 1.00
P25 0.01 0.05 1.00
P26 0.06 0.09 0.08 1.00
P27 0.00 0.03 0.02 0.05 1.00
P28 0.04 0.10 0.05 0.09 0.03 1.00
P29 0.02 0.04 0.04 0.07 0.02 0.03 1.00
P30 0.05 0.06 -0.01 0.02 0.00 0.02 0.00 1.00
P31 0.07 0.08 0.05 0.07 0.01 0.06 0.03 -0.01 1.00
P32 0.03 0.08 0.06 0.09 0.03 0.07 0.03 0.02 0.08 1.00
P33 0.07 0.10 0.06 0.11 0.01 0.09 0.06 0.02 0.05 0.09 1.00
P34 0.04 0.05 0.04 0.08 0.03 0.07 0.02 0.01 0.05 0.07 0.10
P35 0.04 0.12 0.05 0.10 0.03 0.09 0.08 0.01 0.05 0.07 0.15
P36 0.05 0.14 0.07 0.14 0.03 0.11 0.05 0.03 0.07 0.10 0.17
P37 0.02 0.02 0.02 0.01 0.06 0.03 0.02 0.00 0.05 0.02 0.00
P38 0.05 0.11 0.03 0.08 -0.01 0.09 0.05 0.02 0.05 0.05 0.11
P39 0.02 0.05 0.03 0.04 -0.01 -0.04 0.02 0.01 0.01 0.05 0.08
P40 0.04 0.10 0.05 0.09 0.01 0.08 0.08 0.01 0.05 0.07 0.09
P41 0.06 0.11 0.04 0.08 -0.01 0.07 0.06 0.02 0.04 0.07 0.13
P42 0.07 0.15 0.08 0.19 0.07 0.15 0.06 0.04 0.10 0.11 0.17
P43 0.04 0.10 0.05 0.09 0.00 0.06 0.04 0.02 0.06 0.09 0.13
P44 0.06 0.15 0.04 0.09 -0.01 0.09 0.05 0.01 0.05 0.08 0.13
P45 0.08 0.22 0.07 0.19 0.05 0.17 0.06 0.05 0.09 0.13 0.22
  P34 P35 P36 P37 P38 P39 P40 P41 P42 P43 P44
P34 1.00
P35 0.10 1.00
P36 0.09 0.17 1.00
P37 0.02 0.03 0.03 1.00
P38 0.06 0.16 0.18 0.05 1.00
P39 0.04 0.07 0.08 0.03 0.09 1.00
P40 0.06 0.11 0.15 0.04 0.12 0.06 1.00
P41 0.06 0.14 0.17 0.03 0.15 0.09 0.11 1.00
P42 0.15 0.16 0.23 0.05 0.15 0.06 0.15 0.15 1.00
P43 0.08 0.10 0.12 0.02 0.08 0.07 0.08 0.12 0.12 1.00
P44 0.06 0.16 0.19 0.08 0.19 0.10 0.15 0.19 0.17 0.11 1.00
P45 0.13 0.26 0.34 0.05 0.26 0.14 0.19 0.26 0.32 0.18 0.29
[1] 1.00

```
 with tau of
  P1   P2   P3   P4   P5   P6   P7   P8   P9  P10  P11  P12  P13  P14  P15  P16
0.35 0.84 0.61 0.79 0.75 0.68 0.75 1.10 0.59 0.69 0.64 0.68 0.41 1.28 0.76 0.46
 P17  P18  P19  P20  P21  P22  P23  P24  P25  P26  P27  P28  P29  P30  P31  P32
0.52 0.58 0.76 0.84 0.68 0.27 0.63 0.83 0.71 0.79 1.00 1.03 0.84 0.84 0.87 0.52
 P33  P34  P35  P36  P37  P38  P39  P40  P41  P42  P43  P44  P45
0.50 0.71 0.56 0.60 0.52 0.32 0.23 0.73 0.12 1.18 0.57 0.38 0.73
> dados_cor$rho


> KMO(dados_cor$rho)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = dados_cor$rho)
Overall MSA =  0.92
MSA for each item =
  P1   P2   P3   P4   P5   P6   P7   P8   P9  P10  P11  P12  P13  P14  P15  P16
0.86 0.62 0.84 0.88 0.93 0.78 0.53 0.87 0.94 0.94 0.87 0.87 0.94 0.63 0.94 0.93
 P17  P18  P19  P20  P21  P22  P23  P24  P25  P26  P27  P28  P29  P30  P31  P32
0.94 0.93 0.89 0.93 0.92 0.92 0.86 0.94 0.89 0.92 0.71 0.92 0.87 0.72 0.86 0.92
 P33  P34  P35  P36  P37  P38  P39  P40  P41  P42  P43  P44  P45
0.94 0.93 0.94 0.95 0.78 0.93 0.88 0.95 0.94 0.94 0.94 0.93 0.93
> cortest.bartlett(dados_cor$rho, n = 10000, diag=TRUE)
$chisq
[1] 30403.76

$p.value
[1] 0

$df
[1] 990

> alpha(dados_cor$rho)
[1] 0.6278158

> eigen_dados <-fa.parallel( dados, n.iter=10,SMC=TRUE,  fm = "minres", sim=TRUE, fa="pc", cor="poly")
Parallel analysis suggests that the number of factors =  NA  and the number of components =  8
> names(eigen_dados)
 [1] "fa.values" "pc.values" "pc.sim"    "pc.simr"   "fa.sim"    "fa.simr"
 [7] "nfact"     "ncomp"     "Call"      "values"
> list(eigen_dados)
[[1]]
Call: fa.parallel(x = dados, fm = "minres", fa = "pc", n.iter = 10,
    SMC = TRUE, sim = TRUE, cor = "poly")
Parallel analysis suggests that the number of factors =  NA  and the number of components =  8

 Eigen Values of

 eigen values of factors
 [1]  4.00  0.53  0.31  0.19  0.15  0.11  0.11  0.09  0.09  0.07  0.06  0.05
[13]  0.05  0.03  0.03  0.02  0.01  0.01  0.00 -0.01 -0.01 -0.02 -0.02 -0.03
[25] -0.03 -0.04 -0.04 -0.04 -0.05 -0.06 -0.07 -0.07 -0.07 -0.09 -0.09 -0.09
[37] -0.10 -0.10 -0.11 -0.12 -0.12 -0.13 -0.14 -0.14 -0.15

 eigen values of components
 [1] 4.83 1.44 1.24 1.13 1.10 1.06 1.05 1.04 1.02 1.02 1.01 1.00 0.99 0.98 0.97
[16] 0.96 0.96 0.95 0.94 0.93 0.93 0.92 0.91 0.90 0.90 0.89 0.88 0.87 0.87 0.85
[31] 0.85 0.84 0.84 0.83 0.81 0.81 0.79 0.78 0.78 0.76 0.74 0.73 0.72 0.62 0.55
```

```
> dados_facpri <- principal(dados_cor$rho, nfactors = 3, residuals=FALSE, rotate="none", n.obs=NA, covar=FALSE, scores=FALSE, missing=FALSE, impute="median", oblique.scores=FALSE)
> dados_facpri
Principal Components Analysis
Call: principal(r = dados_cor$rho, nfactors = 3, residuals = FALSE,
   rotate = "none", n.obs = NA, covar = FALSE, scores = FALSE,
   missing = FALSE, impute = "median", oblique.scores = FALSE)
Standardized loadings (pattern matrix) based upon correlation matrix
     PC1   PC2   PC3    h2   u2 com
P1   0.16  0.03  0.24 0.082 0.92 1.8
P2  -0.03  0.29  0.30 0.173 0.83 2.0
P3   0.15  0.01  0.30 0.114 0.89 1.5
P4   0.21  0.23 -0.06 0.104 0.90 2.1
P5   0.50  0.02 -0.27 0.320 0.68 1.5
P6   0.11  0.27 -0.15 0.105 0.89 2.0
P7   0.00  0.00  0.28 0.077 0.92 1.0
P8   0.21  0.34 -0.10 0.166 0.83 1.9
P9   0.49 -0.01  0.02 0.239 0.76 1.0
P10  0.40  0.07 -0.17 0.191 0.81 1.4
P11  0.15  0.17  0.22 0.102 0.90 2.7
P12  0.19  0.23  0.10 0.098 0.90 2.4
P13  0.21 -0.06 -0.12 0.064 0.94 1.8
P14  0.04  0.38 -0.26 0.212 0.79 1.8
P15  0.36  0.04 -0.14 0.150 0.85 1.3
P16  0.62 -0.22 -0.11 0.440 0.56 1.3
P17  0.50 -0.17 -0.06 0.284 0.72 1.3
P18  0.45 -0.16 -0.21 0.267 0.73 1.7
P19  0.28  0.36  0.12 0.223 0.78 2.1
P20  0.26  0.09  0.13 0.092 0.91 1.7
P21  0.23  0.01  0.13 0.069 0.93 1.6
P22  0.27 -0.11  0.29 0.170 0.83 2.3
P23  0.17  0.07  0.32 0.134 0.87 1.6
P24  0.37  0.00  0.12 0.148 0.85 1.2
P25  0.17  0.18  0.00 0.062 0.94 2.0
P26  0.34  0.32 -0.04 0.224 0.78 2.0
P27  0.09  0.23 -0.36 0.193 0.81 1.8
P28  0.30  0.15 -0.06 0.118 0.88 1.6
P29  0.16  0.13 -0.02 0.041 0.96 1.9
P30  0.07  0.04  0.17 0.036 0.96 1.5
P31  0.19  0.22  0.20 0.123 0.88 3.0
P32  0.25  0.20  0.19 0.136 0.86 2.8
P33  0.38  0.02  0.07 0.150 0.85 1.1
P34  0.25  0.16  0.06 0.093 0.91 1.9
P35  0.40 -0.10 -0.09 0.183 0.82 1.2
P36  0.50 -0.10 -0.02 0.255 0.74 1.1
P37  0.11  0.02  0.05 0.015 0.99 1.4
P38  0.39 -0.24  0.06 0.214 0.79 1.7
P39  0.22 -0.18  0.12 0.094 0.91 2.6
P40  0.34 -0.05  0.06 0.124 0.88 1.1
P41  0.41 -0.23  0.09 0.225 0.77 1.7
P42  0.52  0.16 -0.06 0.298 0.70 1.2
P43  0.31  0.04  0.13 0.117 0.88 1.4
P44  0.46 -0.27  0.05 0.287 0.71 1.6
P45  0.69 -0.13 -0.08 0.502 0.50 1.1

                       PC1  PC2  PC3
SS loadings           4.83 1.44 1.24
Proportion Var        0.11 0.03 0.03
Cumulative Var        0.11 0.14 0.17
Proportion Explained  0.64 0.19 0.16
```

Cumulative Proportion 0.64 0.84 1.00

Mean item complexity = 1.7
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is 0.03

Fit based upon off diagonal values = 0.85> dados_facpa <- fa(dados_cor$rho, nfactors=3, rotate="none", n.obs=100000, np.obs=100000, cor="tet", n.iter=1, scores= "none", fm="pa")
> dados_facpa
Factor Analysis using method = pa
Call: fa(r = dados_cor$rho, nfactors = 3, n.obs = 1e+05, n.iter = 1,
rotate = "none", scores = "none", fm = "pa", np.obs = 1e+05,
cor = "tet")
Standardized loadings (pattern matrix) based upon correlation matrix

	PA1	PA2	PA3	h2	u2	com
P1	0.14	0.02	0.10	0.0306	0.97	1.9
P2	-0.03	0.16	0.13	0.0431	0.96	2.0
P3	0.13	0.01	0.12	0.0315	0.97	2.0
P4	0.19	0.13	-0.03	0.0535	0.95	1.9
P5	0.46	0.02	-0.21	0.2574	0.74	1.4
P6	0.09	0.14	-0.06	0.0319	0.97	2.1
P7	0.00	0.00	0.10	0.0102	0.99	1.0
P8	0.18	0.20	-0.05	0.0742	0.93	2.1
P9	0.45	0.01	0.03	0.2006	0.80	1.0
P10	0.36	0.05	-0.09	0.1382	0.86	1.2
P11	0.13	0.10	0.09	0.0355	0.96	2.7
P12	0.17	0.13	0.05	0.0466	0.95	2.1
P13	0.19	-0.03	-0.05	0.0386	0.96	1.2
P14	0.04	0.21	-0.13	0.0648	0.94	1.7
P15	0.32	0.03	-0.08	0.1091	0.89	1.1
P16	0.59	-0.18	-0.09	0.3894	0.61	1.2
P17	0.46	-0.11	-0.03	0.2257	0.77	1.1
P18	0.41	-0.10	-0.13	0.1942	0.81	1.3
P19	0.25	0.24	0.07	0.1219	0.88	2.2
P20	0.23	0.07	0.07	0.0614	0.94	1.4
P21	0.20	0.01	0.07	0.0456	0.95	1.3
P22	0.24	-0.05	0.17	0.0913	0.91	1.9
P23	0.15	0.04	0.12	0.0381	0.96	2.1
P24	0.33	0.01	0.08	0.1135	0.89	1.1
P25	0.15	0.10	0.02	0.0323	0.97	1.8
P26	0.31	0.23	-0.02	0.1459	0.85	1.8
P27	0.08	0.13	-0.18	0.0532	0.95	2.2
P28	0.27	0.10	-0.03	0.0816	0.92	1.3
P29	0.14	0.07	0.00	0.0237	0.98	1.5
P30	0.06	0.03	0.05	0.0075	0.99	2.3
P31	0.17	0.13	0.10	0.0527	0.95	2.5
P32	0.22	0.13	0.10	0.0748	0.93	2.1
P33	0.34	0.03	0.05	0.1194	0.88	1.1
P34	0.22	0.10	0.03	0.0595	0.94	1.5
P35	0.36	-0.05	-0.03	0.1359	0.86	1.1
P36	0.45	-0.05	0.01	0.2088	0.79	1.0
P37	0.10	0.01	0.02	0.0101	0.99	1.2
P38	0.35	-0.14	0.06	0.1496	0.85	1.4
P39	0.19	-0.09	0.07	0.0501	0.95	1.8
P40	0.30	-0.02	0.05	0.0954	0.90	1.1
P41	0.37	-0.14	0.08	0.1607	0.84	1.4
P42	0.48	0.14	-0.04	0.2514	0.75	1.2
P43	0.28	0.03	0.09	0.0850	0.92	1.2
P44	0.42	-0.18	0.06	0.2164	0.78	1.4

P45 0.68 -0.11 -0.06 0.4716 0.53 1.1

	PA1	PA2	PA3
SS loadings	4.04	0.56	0.33
Proportion Var	0.09	0.01	0.01
Cumulative Var	0.09	0.10	0.11
Proportion Explained	0.82	0.11	0.07
Cumulative Proportion	0.82	0.93	1.00

Mean item complexity = 1.6
Test of the hypothesis that 3 factors are sufficient.

The degrees of freedom for the null model are 990 and the objective function was 3.05 with Chi Square of 304499
The degrees of freedom for the model are 858 and the objective function was 0.17

The root mean square of the residuals (RMSR) is 0.01
The df corrected root mean square of the residuals is 0.01

The harmonic number of observations is 1e+05 with the empirical chi square 27196.92 with prob < 0
The total number of observations was 1e+05 with Likelihood Chi Square = 16798.44 with prob < 0

Tucker Lewis Index of factoring reliability = 0.939
RMSEA index = 0.014 and the 90 % confidence intervals are 0.013 0.014
BIC = 6920.35
Fit based upon off diagonal values = 0.98
Measures of factor score adequacy

	PA1	PA2	PA3
Correlation of (regression) scores with factors	0.92	0.63	0.52
Multiple R square of scores with factors	0.84	0.40	0.27
Minimum correlation of possible factor scores	0.68	-0.21	-0.45

	PC1	PC2	PC3	PA1	PA2	PA3
PC1	1.00	0.00	0.00	1.00	0.02	0.01
PC2	0.00	1.00	0.00	-0.01	0.99	-0.04
PC3	0.00	0.00	1.00	-0.01	0.03	0.97
PA1	1.00	-0.01	-0.01	1.00	0.00	0.00
PA2	0.02	0.99	0.03	0.00	1.00	0.00
PA3	0.01	-0.04	0.97	0.00	0.00	1.00

Análise do Bifatorial - PROC CALIS – SAS

The CALIS Procedure
Correlation Structure Analysis: Maximum Likelihood Estimation

Fit Summary	
Chi-Square	16.9154
Chi-Square DF	900
Pr > Chi-Square	1.0000