MENSURAÇÃO EM SISTEMAS ADMISSIONAIS NO ENSINO SUPERIOR E AVALIAÇÕES DE IMPACTO

Tufi Machado Soares[1]

Mariana Calife Nóbrega Soares[2]

[1] Professor Titular do Departamento de Estatística da UFJF e pesquisador da Fundação CAEd/UFJF, Juiz de Fora, Minas Gerais, Brasil. E-mail: tufi@caed.ufjf.br.

[2] Doutora em Educação pela PUC-Rio e analista de avaliação da Fundação CAEd/UFJF. Juiz de Fora, Minas Gerais, Brasil. E-mail: mariana.soares@caed.ufjf.br .

Resumo

Este trabalho apresenta um estudo que revisa os principais sistemas de admissão ao ensino superior no mundo. Em particular, revisa os diferentes processos de admissão e os exames e provas usados além dos métodos de mensuração empregados. Além disso, aborda a necessidade de se avaliar a qualidade desses testes e suas medidas e revisa alguns desses métodos. Finalmente, uma análise contextualizada da prova do ENEM, assim como seu uso no SISU, é realizada utilizando os resultados encontrados no estudo.

Palavras-chave: Avaliação de Impacto. Medidas. ENEM. Qualidade. SISU.

Abstract

This work presents a study that reviews the main admission systems to higher education Institutions in the world. In particular, it reviews the different admission processes and the exams and tests used in addition to the measurement methods employed. Moreover, it addresses the need to evaluate the quality of these tests and their measures and reviews some of these methods. Finally, it performs a contextualized analysis of the ENEM test, as well as its use in SISU using the results found in the study.

Keywords: Impact Evaluation. Measures. ENEM. Quality. SISU.

INTRODUÇÃO

A admissão no ensino superior no mundo tem características bastante variadas, não havendo padrões específicos que são seguidos exatamente, mesmo por grupos de países política e culturalmente parecidos. Além disso, mesmo dentro de um país, variações no processo de admissão dos candidatos são observadas. Assim, cada país cria ou adapta os critérios de admissão conforme suas necessidades e tradições.

Alguns padrões gerais serão discutidos com vistas a embasar a discussão da prova do ENEM. No contexto dos países da OCDE (SARGEANT et. al., 2012), uma primeira classificação dos sistemas de admissão é a separação entre **sistemas abertos** e **sistemas seletivos de admissão**: um sistema aberto existe quando resultados ou certificados obtidos no ensino médio garantem o acesso automático à educação superior desde que certas credenciais sejam alcançadas pelos candidatos; já em um sistema seletivo outros critérios mais rígidos são utilizados para a admissão.

Não significa dizer, portanto, que um sistema aberto garante a Educação Superior a todos que desejam fazê-la, mas que a seleção/credenciamento pode ocorrer em diferentes momentos da educação básica. A Alemanha, por exemplo, tem um sistema considerado aberto para a admissão no ensino superior, no entanto, o sistema pode ser considerado altamente seletivo ao longo da educação básica. De fato, a taxa de graduação é de 30% na Alemanha, contra 38% nos Estados Unidos, que apresenta um sistema mais seletivo na admissão (MCGRATH et. al., 2014).

Por outro lado, mesmo em um sistema aberto, para determinadas universidades de grande prestígio e para certos cursos mais disputados, ou que requeiram habilidades especiais, como é comum no caso dos cursos de Medicina, por exemplo, há processos seletivos complementares aos resultados obtidos no ensino médio. Além disso, muitos sistemas abertos utilizam os resultados de uma prova de credenciamento ao final do ensino médio, como é o caso do exame Baccalaureate, na França.

Segundo uma tipologia adaptada de Palmer, Bexley e Jame (2011) os diferentes sistemas de admissão ou não utilizam exames ou testes, como os que utilizam os resultados do ensino eédio, como é o caso da Noruega e do Canadá, ou usam testes ou exames: exames de saída do ensino médio[3]; exames para en-

[3] *Secondary leaving examinations.*

trada no ensino superior[4], que em geral medem o conhecimento adquirido em certas disciplinas ou áreas de conhecimento; testes padronizados de aptidões ou habilidades[5], que incluem medidas de habilidades ou competências cognitivas, compreensão e outras habilidades gerais; múltiplos exames realizados ao longo do ensino médio e/ou diferentes exames realizados em diferentes momentos no processo de entrada no ensino superior[6]. De qualquer forma, mesmo dentro de um país, a existência de um sistema unificado é exceção.

Edwards, Coates e Friedman (2012) classificam os sistemas de admissão, segundo o uso dos testes da seguinte forma: sistemas que utilizam testes como o único determinante para a entrada e citam, especificamente, China, Grécia, Portugal e Coréia do Sul; sistemas que utilizam os testes como um critério adicional para a admissão e citam Japão, África do Sul, Suécia, Turquia e Estados Unidos; e sistemas que utilizam menos frequentemente os testes no processo admissional citam, como exemplos, a Austrália, Canadá, México e Reino Unido. Nestes últimos, os testes são mais comuns para determinadas universidades e cursos específicos.

Em um estudo[7] sistemático realizado por McGrath et. al. (2014), os sistemas de ingresso (ou admissão) à educação superior de 10 países foram analisados e comparados detalhadamente, entre eles: França, Itália, Alemanha, Eslovênia, Suécia e Reino Unido, Turquia, Austrália, Japão e Estados Unidos. Três aspectos principais usados para a comparação dos sistemas de ingresso foram a capacidade de produzir equidade, a qualidade dos candidatos selecionados e a capacidade de gerar mobilidade dentro da União Europeia. Alguma evidência foi achada de que a forma de ingresso está associada a essas três características do sistema de educação superior, mas o mais interessante, nesse caso, é que o parlamento europeu entende que essas são características importantes para a composição de um sistema de educação superior. Normalmente, nos EUA, como se verá na seção 3, um aspecto importante dos testes aplicados aos alunos é que seus resultados estejam altamente correlacionados ao sucesso acadêmico futuro, como o êxito na conclusão total ou parcial de determinadas etapas do curso selecionado. Mais importante do que um ranqueamento inicial dos candidatos, portanto, nesses países, é a previsibilidade do êxito na conclusão do ensino superior.

[4] *Entrance examinations.*

[5] *Standardized aptitude tests.*

[6] *Multiple examinations.*

[7] Estudo patrocinado e apoiado pelo *Policy Departament B: Structural cohesion policies – Culture and Education* do Parlamento Europeu.

Outra caracterização importante dos sistemas de admissão é o grau de centralização do processo. Centralização quer dizer que existe um ou mais processos unificados para organizar o ingresso em um grupo de universidades. De fato, portanto, podemos dizer que um sistema é mais centralizado ou mais descentralizado que outro. Dificilmente existem sistemas totalmente centralizados em um único sistema admissional regidos pelos mesmos critérios. Sistemas mais centralizados podem ainda dar pouca autonomia às universidades ou mais autonomia no processo de escolha, mas, em geral, sistemas mais descentralizados dão autonomia para a escolha dos candidatos, a escolha dos critérios de admissão e a operacionalização da aplicação.

OS PAÍSES E SEUS SISTEMAS DE ADMISSÃO

São vários os tipos de instrumentos de admissão existentes para auxiliar as universidades a escolherem seus estudantes, incluindo exames de saída do ensino médio ou exames de certificação ou exames de entrada e testes patronizados de aptidões. Em alguns casos, os alunos podem ser admitidos sem qualquer exame, mas, ainda assim, selecionados por meio de demonstração de requisitos obtidos ao longo de sua educação básica formal. Por outro lado, mesmo quando se considera um mesmo país, há diferentes formas de requisitos e seleção, dependendo se o curso ou a universidade possui uma grande procura, prestígio ou requisitos específicos. Apresentam-se, sucintamente, as principais características dos sistemas de admissão de alguns países:

V. Na **Austrália**, existem dois tipos de admissão, a mais comum ocorre por meio do "*Year12 Apllicants*" (um sistema centralizado de admissão, destinado aos alunos do último ano do ensino médio) e o "*Mature Applicants*" destinado a alunos mais velhos. Normalmente, quando a entrada acontece por meio do "*Year12 Apllicants*", os candidatos apresentam um certificado que os habilita para a educação superior denominado "*Australian Tertiary Admissions Rank*" (ATAR). Os escores do **ATAR** são calculados comparativamente aos estudantes do mesmo ano em dado estado ou território e são calculados em termos de níveis de percentis. Algumas universidades ou centros utilizam ainda testes padronizados denominados "*Special Tertiary Admissions Test*" (STAT), e outras exigem exames da própria instituição ou outros exames específicos, geralmente para medição de habilidades ou conhecimentos específicos em determinadas áreas.

VI. Na **Inglaterra,** a forma mais comum de admissão é a submissão dos resultados dos exames do GCE A-Levels, ou equivalentes, realizados ao final do ensino médio para as universidades que, em sua maioria, estabelecem os níveis mínimos de aceitação. Além disso, um grande número de universidades realiza seus próprios exames.

VII. Na **França**, os candidatos devem ter o certificado de nível médio obtido por meio do exame Baccalaureate, ou o diploma de nacional (*Diplôme d'accès aaux études universitaries – DAEU*) ou equivalentes internacionais. Note-se que nem todos os que concluíram o ensino médio francês se habilita, assim, para o sistema universitário. O sistema francês é em sua maioria um sistema aberto para acesso ao ensino superior, desde que os candidatos tenham o certificado. No entanto, outros exames de entrada e/ou entrevistas são exigidos para cursos e instituições de grande prestígio, como as chamadas *grandes écoles*.

VIII. Na **Alemanha**, encontra-se um sistema aberto de admissão, os candidatos devem possuir um certificado de qualificação para a entrada na educação superior chamado Abitur. Alternativas existem para candidatos mais velhos e experientes, que podem entrar por meio de outros tipos de testes ou qualificação.

IX. Na **Itália**, geralmente, a admissão transcorre a partir da qualificação obtida com a conclusão do ensino médio e algum certificado emitido pelas escolas. Alguns cursos possuem a admissão regulada nacionalmente por meio de testes padronizados, como medicina e cirurgia, odontologia, medicina veterinária, enfermagem e arquitetura.

X. Todas as universidades no **Japão**, públicas ou privadas, devem seguir o guia para a implementação da seleção dos alunos determinado pelo Ministério da Educação, Cultura, Esportes e Ciência e Tecnologia (MEXT) no final do mês de maio. Apesar disso, todas as universidades têm liberdade para se decidirem pelos critérios de seleção, que geralmente são baseados em exames de entrada geral das universidades, recomendação dos diretores das escolas de ensino médio, testes padronizados produzidos pelo Centro Nacional de Testagem para Admissão nas Universidades ou combinação desses requisitos. O sistema de seleção é quase todo descentralizado.

XI. Os requisitos de entrada na **Suécia** são baseados nos *school-leaving grades*, resultados escolares do final do ensino médio, para a maioria dos cursos e por meio dos resultados do *Swedish Scholastic Aptitude Test* (SweSAT). Requisitos adicionais são exigidos para algumas áreas em particular como saúde, direito e áreas específicas que exigem habilidades artísticas.

XII. O sistema educacional Norte Americano (**EUA**) é provavelmente o mais descentralizado de todos. As instituições definem os critérios, selecionam e controlam a aplicação dos alunos. No entanto, para alguns cursos de algumas universidades, existem alguns sistemas unificados. Entre os requisitos de entrada nas universidades norte-americanas são muito comuns os resultados dos testes de aptidões, sendo os mais comuns o *Scholastic Aptitude Tests* (SATs) e os testes de *Subject American College Testing* (ACTs), realizados independentemente das universidades por organizações ou empresas especializadas em testagem. Outros testes de aptidões e/ou conhecimentos são requisitados para determinados cursos, como direito e medicina. Os resultados dos testes, geralmente, são parte dos requisitos, mas não constituem o único requisito de entrada.

XIII. Na **Turquia**, o sistema é totalmente centralizado. O sistema nacional seleciona os critérios, os alunos e controla a aplicação. A entrada na universidade é baseada no ÖSS, um exame compulsório baseado em testes para habilidades verbais e quantitativas. O teste abrange uma variedade de disciplinas, incluindo ciências, matemática, língua turca, língua estrangeira e ciências sociais. Os resultados do ÖSS são combinados com as médias do certificado *Lise Diplomasi*, de saída do Eesino básico. O *Yükseköğretim Kurulu* (YÖK), conselho de educação superior, é responsável por coordenar e determinar os critérios e realizar a seleção dos candidatos em si.

XIV. No **Chile**, existe a prova de seleção Universitária (*La Prueba de Selección Universitaria* – PSU), cujos resultados são usados para a admissão por muitas das Universidades do país. Os testes de Matemática e Linguagem e comunicação são obrigatórios, enquanto os testes de ciências (biologia, física, química e técnico-profissional) e história, geografia e ciências sociais são eletivos. Os alunos escolhem as eletivas de acordo com os requisitos das universidades. Existe um sistema único de admissão, simultâneo e integrado, para as Universidades Chilenas Públicas do *Consejo de Rectores* (CRUCH) e Universidades privadas associadas. O sistema é

coordenado pelo *Departamento de Evaluación, Medición Y Registro Educacional* (DEMRE) da *Univesidad de Chile* que também coordena a prova PSU (cf. UNIVERSIDAD DE CHILE, [c2017]b). As Universidades consideram para a admissão, além das pontuações na PSU, o ranking de pontuações obtidas pelos colegas na escola de ensino médio (medida do contexto) e as notas obtidas nas escolas de nível médio.

Variações no processo seletivo dentro de um mesmo país ocorrem, frequentemente, entre os cursos. A disciplina de ciências médicas representa o principal exemplo, em muitos países, em que testes de aptidão complementam os certificados do ensino médio. Na Inglaterra, os estudantes têm que realizar testes como o *Bio-Medical Admissions Test* (BMAT), o *The Helth Professions Admission Test* (HPAt) ou o *UK Clinical Aptitude Test* (UKCAT). Na Itália, os cursos de medicina e cirurgia, veterinária e odontologia estão entre os que têm sua entrada regulada a nível nacional por meio de testes padronizados que focam conhecimentos gerais (história, raciocínio lógico etc.) e conhecimentos específicos relevantes para a área em questão. Na Austrália, o teste *Undergraduate Medical Admissions Test* (UMAT) é requisito para muitos cursos de odontologia e médicos. Nos EUA, o *Law School Admission Test* (LSAT) é muito usado na admissão dos candidatos nas escolas de Direito e o *Medical College Admission Test* (MCAT) para as escolas de medicina.

OS TESTES E SEUS SISTEMAS DE MENSURAÇÃO

Os testes, dentro das Teorias dos Testes de Avaliação, incluindo os de seleção, são construídos para medir (avaliar) algum constructo: o aprendizado em uma disciplina, a proficiência em física, o conhecimento em matemática, a aptidão em relações humanas, a habilidade de raciocínio quantitativo, a habilidade de interpretação de textos em língua portuguesa etc. Genericamente, utilizaremos o termo **proficiência cognitiva** para o constructo a ser medido, tendo em vista que nos exames admissionais para o ensino superior esses constructos geralmente representam alguma dimensão da cognição humana. Portanto, os testes são, essencialmente, instrumentos de medição de uma característica do ser humano que não pode ser observada diretamente por outro meio (por isso ela é denominada, então, de latente).

Os testes têm usos diversos. Testes de avaliação educacional em larga escala, como o PISA[8], o NAEP[9], o SAEB[10] e a Prova Brasil, são normalmente desenhados para avaliar sistemas educacionais. Por outro lado, há os testes que necessitam ser desenhados para avaliar uma ou mais características de um indivíduo especificamente. É o caso de testes de diagnósticos psicológicos, psiquiátricos e, naturalmente, aqueles que orientam a seleção para ingresso no ensino superior. Essa diferença é substancial, pois a composição do teste, o conjunto de habilidades a serem medidas, as dificuldades dos itens entre outras características dos itens dos testes devem ser calibradas para atender às suas finalidades específicas.

Os testes de admissão são, geralmente, divididos entre os testes de realização ou êxito (*achievement*), mais comuns no Brasil, também denominados de testes de conhecimento (*subject*), como os que medem o conhecimento adquirido em uma disciplina em uma etapa escolar, por exemplo, e que naturalmente se referem a uma realização passada do estudante; os testes de habilidades (*abilities*) e competências cognitivas que medem a capacidade de realizar determinadas tarefas como, por exemplo, o raciocínio quantitativo e que, em geral, representam uma habilidade atual do indivíduo; e os testes de aptidão (*aptitude*) que procuram revelar uma capacidade do indivíduo para uma realização exitosa no futuro. Evidentemente, nem sempre os testes são exclusivamente de um tipo somente, mas costumam transitar entre essas três características, mesmo que na maioria dos casos uma delas predomine sobre as demais.

Definido o constructo ou constructos que são relevantes, os testes de seleção, por exemplo, são construídos para escolher, entre os candidatos, aqueles que apresentam os maiores valores de proficiência para o constructo ou, ainda, para distinguir aqueles que alcançaram determinados níveis de habilidades mínimas requisitadas para o curso daqueles que não alcançaram essas habilidades.

Como não é possível observar diretamente o constructo de interesse, usa-se o teste como instrumento para se obter uma avaliação do constructo. Chama-se a esse processo de medição, sendo o teste o instrumento de medição, e o resultado que se extrai dele, no caso uma pontuação calculada com base nos resultados das questões, uma medida do constructo. Assim, a pontuação do

[8] *Program for International Student Assessment* (Pisa) – OECD.

[9] *National Assessment of Educational Progress* (NAEP) – USA.

[10] **Sistema de Avaliação da Educação Básica.**

resultado de um teste respondido por um candidato é, então, uma avaliação (medição) do constructo que se quer conhecer. Na próxima subseção, as técnicas de pontuação serão discutidas e suas implicações analisadas.

Formas de pontuação e medição dos testes

Infelizmente, toda medida é uma representação parcial do valor que se quer de fato conhecer, denominado por alguns autores de **valor verdadeiro** do constructo, embora não haja consenso sobre essa interpretação. Assim, a toda medida se associa uma incerteza, que é inerente a qualquer processo de medição. Esse argumento vale para medidas em todas as áreas do conhecimento e suas aplicações: desde a física até as ciências humanas e sociais. As diferentes teorias das medidas, nas diferentes áreas e aplicações procuram, então, quantificar essas incertezas para que novos métodos de medição possam ser propostos com o objetivo de produzir resultados mais acurados e as medidas possam ser usadas de forma mais adequada.

A incerteza da medida é usualmente denominada no campo da Psicometria e da Estatística como o **erro da medida**. O termo erro é, infelizmente, mal entendido muitas vezes. Ele não quer dizer que a medida esteja errada em si mesma, mas que toda medida tem associada uma incerteza em relação ao valor verdadeiro do que se quer de fato conhecer e que não pode ser observado diretamente.

Na Psicometria, que é a ciência que estuda os processos de medição de constructos latentes por meio de testes psicométricos, dos quais os testes de avaliação e seleção são exemplos, duas teorias principais norteiam a construção e a verificação da qualidade das medidas. Com a primeira, denominada de Teoria Clássica dos Testes (TCT), procura-se analisar os erros das medidas calculadas pelos métodos clássicos de pontuação dos testes e, assim, sugerir formas de pontuação mais acuradas e fidedignas dos valores verdadeiros, além de preconizar métodos de pontuação que permitam a comparabilidade de resultados quando os instrumentos (testes) são diferentes, por exemplo.

A segunda teoria mais comum no campo da Psicometria é a Teoria da Resposta ao Item (TRI) que, apontando determinadas limitações nas formas clássicas de pontuação, preconiza formas diferentes de se pontuar os testes e construir medidas para os constructos latentes. Particularmente, ela é muito útil na comparabilidade de resultados de testes diferentes aplicados a diferentes alunos.

A seguir, analisadas no campo das duas teorias, serão discutidas as principais formas de pontuação, suas limitações e vantagens.

Forma Clássica de Pontuação e Teoria Clássica do Teste (TCT)

Os escores reportados como resultados dos testes são derivados do desempenho obtido pelo candidato no teste por meio de um processo estatístico chamado pontuação (*scaling*). O exemplo mais simples é aquele em que um candidato que responde corretamente a 55 itens, em um teste de 60, recebe uma pontuação (*score*) de 55. Nesse caso, a pontuação é chamada de pontuação bruta (*raw score*). Outros tipos de *raw scores* são o percentual de acerto no teste ou uma pontuação ponderada dos itens, por exemplo.

Para testes padronizados, e testes admissionais no mundo, a pontuação bruta quase nunca é utilizada diretamente. Isso ocorre porque os pontos brutos são diretamente dependentes dos itens presentes em uma forma[11] particular de um teste. Por exemplo, uma determinada forma pode conter itens mais fáceis que outra forma e, portanto, a mesma pontuação, em ambas as formas do mesmo teste, não expressa o mesmo nível de habilidade dos candidatos. Principalmente por isso, quase sempre, os pontos brutos são transformados para uma escala específica. Este processo é denominado **escalonamento** (*scaling*).

Grosso modo, a TCT lida com a incerteza da medida da seguinte forma: admita que seja possível construir formas paralelas de um teste, isto é, formas de um teste cujos resultados sejam indistinguíveis. Ainda assim, estudos empíricos mostram que os pontos obtidos não são exatamente os mesmos quando o mesmo aluno responde a essas formas. Portanto, o mesmo aluno, respondendo a formas absolutamente equivalentes, ainda assim apresenta resultados diferentes.

A teoria clássica, então, admite que o **escore verdadeiro** (**T**) do candidato é a pontuação média obtida em todas as possíveis formas paralelas do teste; na prática, é a média dos pontos brutos que o aluno teria ao responder a uma quantidade grande de formas paralelas. Assim, matematicamente, pode-se relacionar os pontos brutos obtidos em um teste (**O**), chamado de escore observado, como sendo **O = T + e**, onde **e** é denominado de erro de medida observado na pontuação de um teste.

[11] Formas de um mesmo teste são conjuntos de itens (geralmente organizados em um caderno de teste), pontuados segundo uma escala específica, medindo a mesma habilidade cognitiva, portanto, referenciados aos mesmos conteúdos. As formas são denominadas **paralelas**, quando apresentam as mesmas características psicométricas, como a dificuldade dos itens, de tal forma que sejam indistinguíveis do ponto de vista da medida produzida. Na prática, formas paralelas são muito difíceis de serem construídas e, por isso, faz-se necessário algum processo de equalização que estabeleça a comparabilidade das pontuações.

A pontuação bruta obtida no teste pelo candidato é, nesse sentido, uma estimativa para o escore verdadeiro desconhecido. Sob determinadas hipóteses, admitidas sobre a natureza probabilística do erro e sobre o escore verdadeiro, é possível se estimar, por meio de técnicas estatísticas, as propriedades estatísticas do erro **e,** como seu desvio-padrão, denominado de **erro padrão** da medida. Para uma explicação mais pormenorizada desses procedimentos, sugerem-se as seguintes referências: Lord e Novick (1968), Thissen e Wainer (2001).

Para sistemas de testagem que necessitam ter múltiplas aplicações, com diferentes formas do teste, é necessário manter a comparabilidade dos resultados para que possam ser usados adequadamente nos processos de admissão, por exemplo. Quando se trata do mesmo teste aplicado, a comparabilidade dos resultados obtidos pelas diferentes formas é produzida por meio de um processo estatístico chamado de **equalização** (*equating*). O conceito exato de equalização pode variar um pouco, conforme o autor. Mas, na tradição Norte Americana, conforme Kolen e Brennan (2004), a equalização é um procedimento de comparabilidade que ajusta os escores nas diferentes formas de um mesmo teste, que são construídas para serem similares em dificuldade e conteúdo. O escalonamento, portanto, geralmente está associado a um processo de equalização utilizado para manter comparáveis os resultados de diferentes formas de testes aplicadas a diferentes grupos de candidatos. Quando as formas não são similares, ou os testes não são exatamente os mesmos em conteúdo, os métodos de comparabilidade ganham outras denominações como *linking*, moderação, projeção etc. Também é comum o termo “equalização horizontal”, quando os grupos de alunos que realizam as diferentes formas do teste são similares em níveis de proficiências, e “equalização vertical”, quando os grupos de alunos não são similares.

Naturalmente, ao se fazer uma equalização, assim como ao se realizar qualquer outro método de comparabilidade, novas incertezas são introduzidas nas medidas construídas. Por esse motivo, busca-se construir formas de testes que sejam as mais parecidas, do ponto de vista psicométrico e do conteúdo, manter a padronização na aplicação dos testes e empregar os métodos de equalização mais precisos que se disponham. Infelizmente, muitas vezes não se sabe qual é exatamente o melhor método que possa ser empregado em uma situação ou outra. Na teoria, alguns métodos foram comparados em situações ideais e, geralmente, os psicometristas procuram seguir uma fonte ou outra, além da sua própria *expertise*.

Existem muitas maneiras de se produzir a equalização dos resultados. Não se vai aqui discuti-las, tendo em vista a necessidade de conhecimento técnico

específico para isso. Uma boa referência para esses métodos é Kolen e Brennan (2004), que apresentam os métodos de equalização tanto para a forma clássica de pontuação quanto para a Teoria da Resposta ao Item.

Modelos de Resposta ao Item

A Teoria da Resposta ao Item (ver Lord, 1980) surge da necessidade de superar as limitações de se pontuar os resultados dos alunos em testes educacionais na forma clássica, como os percentuais de acertos ou escores brutos dos testes, e da dificuldade de se comparar esses resultados para as diferentes formas de um teste aplicadas em diferentes situações, como em diferentes datas, por exemplo. Na teoria clássica, para se contornar esses problemas, hipóteses muito restritivas são impostas, muitas das quais difíceis de serem comprovadas empiricamente

A TRI muda o foco de análise do teste como um todo para a análise de cada item. A ideia básica consiste no emprego de modelos para o funcionamento dos itens com relação aos valores do constructo que se deseja medir. Geralmente, esses modelos são paramétricos, de tal forma que os parâmetros podem representar características importantes dos itens e, nesse sentido, podem até ser interpretáveis.

A suposição de que, fixada a escala de proficiências, os parâmetros dos itens não variam mesmo quando aplicados a diferentes grupos de alunos, **se observada empiricamente**, garante que as medidas de proficiências construídas por meio desses modelos, para os alunos nos diferentes grupos, estejam na mesma escala e, portanto, sejam comparáveis. Essa propriedade é conhecida como **invariância dos parâmetros dos itens**.

Assim, parte-se do pressuposto de que o modelo do item seja invariante, independente do grupo de alunos a que esteja sendo submetido o teste – suposição esta que pode ou não ser verificada empiricamente –, a TRI permitirá a comparabilidade dos resultados produzidos para grupos de indivíduos diferentes, mesmo quando os testes aplicados são diferentes. É frequente a afirmativa dos autores de que a medida de habilidade produzida por meio da TRI é independente do instrumento aplicado e independente do grupo de indivíduos ao qual é aplicado. Claro que certos cuidados técnicos devem ser tomados para garantir essa propriedade, além da verificação dos pressupostos sob os quais os modelos dos itens são propostos.

Existem muitos tipos de modelos, dependendo de como os itens dos testes são pontuados, do número de constructos a serem medidos, entre outras características. Consideram-se aqui apenas itens pontuados segundo uma regra

dicotômica, certo ou errado, por exemplo, e modelos para medir um único constructo latente, chamados de unidimensionais. Nesse caso, o modelo associa à probabilidade de uma resposta correta no item a proficiência cognitiva e determinados parâmetros dos itens, que representam certas propriedades do item, por meio de uma função matemática específica. Pela importância que possuem nos testes admissionais em questão, vai-se apresentar aqui muito superficialmente o modelo de 3 parâmetros.

Modelo de três parâmetros

Em geral, a maioria dos itens de múltipla escolha em avaliações educacionais em larga escala, pelo menos no Brasil, apresenta um comportamento tal que mesmo alunos com baixa proficiência apresentam uma pequena probabilidade de acertá-los. Claro que, em parte, esse comportamento do item depende da forma como os especialistas elaboram os itens do teste e constroem as alternativas de respostas não corretas (chamadas de distratores[12]). Alguns testes recomendam que o aluno, caso não saiba a resposta correta, não responda ao item, numa tentativa de evitar esse acerto casual, conhecido popularmente como "chute". No entanto, infelizmente, na prática, essa recomendação dificilmente é observada e, mesmo que o seja, muitos itens provavelmente manterão o mesmo comportamento. Esse fenômeno foi interpretado sob diversas formas: Birnbaum (1968), exatamente nas linhas dos argumentos apresentados acima, sugeriu que os alunos de baixa habilidade estariam acertando este item devido a escolhas casuais, por isso sugeriu introduzir o parâmetro c_i no modelo que será apresentado a seguir; Lord (1980) notou que, em geral, o percentual de acerto nesses níveis de habilidades muito baixas era menor do que o inverso do número de alternativas, provavelmente devido à forma como os itens eram elaborados, fazendo com que um dos distratores se tornasse mais atraente para esses alunos. Por esses e outros motivos, o modelo da TRI para os itens utilizados nas avaliações em larga escala de sistemas, no Brasil e em boa parte do mundo, é o chamado modelo logístico de 3 parâmetros, cuja expressão matemática na representação logística é dada por:

$$\boldsymbol{P}\left(y_{ij}=1|\boldsymbol{\theta_j},a_i,b_i,c_i\right)=c_i+\left(1-c_i\right)\frac{1}{1+e^{-D\,a_i\left(\theta_j-b_i\right)}}$$

O modelo acima, $\boldsymbol{P}\left(y_{ij}=1|\boldsymbol{\theta_j},a_i,b_i,c_i\right)$, representa a probabilidade do aluno *j* acertar ao item *i* tendo uma proficiência $\boldsymbol{\theta_j}$, sendo que o item apresenta um índice de discriminação $\boldsymbol{a_i}$, uma dificuldade $\boldsymbol{b_i}$ e um parâmetro de *guessing* $\boldsymbol{c_i}$, que representa a probabilidade de um aluno de proficiência mui-

[12] Tradução literal do inglês *distractors*.

to baixa acertar o item. Para ver mais detalhes dessa representação, incluindo a justificativa para a interpretação dos parâmetros, sugere-se, por exemplo, Baker e Kim (2004).

Pontuação por Meio da Teoria da Resposta ao Item

A forma de se pontuar os resultados dos testes usando a TRI é chamada de métodos de estimação das proficiências. Se conhecidos os parâmetros dos itens, os principais métodos de estimação das proficiências são os métodos de verossimilhança máxima, de máximo da distribuição a posteriori e a média da distribuição a posteriori. O mais usado, na prática, é o método da média da distribuição a posteriori (EAP[13]), que é o método usado no ENEM.

Uma estatística importante que caracteriza a incerteza da medida é chamada de **erro padrão da medida de proficiência**. Os principais *softwares* usados na produção das medidas de proficiências na TRI fornecem estimativas do erro padrão que indicam a acurácia da medida produzida, e inclusive por meio deles pode-se construir intervalos de confiança para o valor verdadeiro da proficiência. Esses intervalos são construídos como forma de indicar a incerteza da medida produzida, fornecendo não uma estimativa pontual, mas um intervalo de valor onde provavelmente o valor verdadeiro está.

As proficiências podem, ainda, ser estimadas conjuntamente com os parâmetros dos itens. Os métodos mais comuns são os de máxima verossimilhança conjunta, máxima verossimilhança conjunta marginalizada, máxima distribuição a posteriori conjunta marginalizada (MMAP), e variações. Em todos os casos, os erros de medida são fornecidos tanto para os parâmetros estimados quanto para as proficiências estimadas, de tal forma que a acurácia das medidas pode ser verificada.

Infelizmente, os métodos de pontuação nesses casos não são baseados em formas explícitas e fáceis de serem apresentadas. Quase sempre, constituem-se de cálculos realizados por métodos iterativos e computacionalmente intensivos. No entanto, os cálculos são totalmente objetivos, uma vez definidos os métodos e os pressupostos básicos. Para uma revisão e explicação pormenorizada desses métodos sugere-se, por exemplo, Baker e Kim (2004). Assim, conhecidos os métodos que são empregados, pode-se reproduzir muito aproximadamente o cálculo de forma independente, desde que se tenham todas as informações e bases necessárias.

Comparação Entre as Formas Clássicas e da TRI de Pontuar

[13] *Expected A Posterori*.

Frequentemente, a literatura aponta uma série de vantagens teóricas da pontuação obtida pelos métodos da TRI sobre a pontuação tradicional. Ver, por exemplo, Hambleton e Jones (1993) para uma comparação das duas teorias. No entanto, como afirmam Thissen e Orlando (2001), ambas as teorias (TCT e TRI) apresentam formas de avaliar a precisão das pontuações e podem ser usadas de forma complementar.

A vantagem mais apontada das pontuações obtidas por meio da TRI é que elas permitem comparabilidade de resultados de testes diferentes, respondidos por candidatos diferentes, por procedimentos mais diretos e, na maioria das vezes, com maior precisão do que os artifícios que necessitam ser usados quando se está pontuando os testes na forma tradicional. Isto é, em geral, afirma-se que é mais simples e preciso produzir a comparabilidade de resultados quando se usa a TRI.

O desempenho dos testes computadorizados e adaptativos, por exemplo, que produzem medidas muito mais precisas do que as produzidas pelos métodos tradicionais de papel e lápis é totalmente potencializado pelo emprego da TRI (Thissen; Orlando, 2001). Assim, devido à sua versatilidade, justifica-se o emprego da TRI, quando é necessário aplicar sistematicamente diferentes formas de teste.

Estudos, no entanto, apontam que aparentemente alguns métodos de comparabilidade desenhados para as técnicas clássicas de pontuação, **com os correspondentes processos de escalonamento,** podem alcançar desempenho similar em algumas situações. Porém, ainda há muita controvérsia na literatura. Alguns estudos demonstram pequena superioridade da pontuação obtida pela TRI em algumas aplicações, mas não em todas as situações.

Em um trabalho recente, Jabrayilov, Emons e Sijtsma (2016) mostraram, em um estudo empírico, a superioridade da pontuação obtida pela TRI na avaliação do ponto de mudança em tratamentos psicológicos clínicos[14], quando os testes contêm ao menos 20 itens, mas não quando os testes são inferiores.

Especificamente em testes admissionais, Fan (1998) utilizou os dados do *Texas Assessment of Academic Skills* (TAAS) *tests*, aplicado em outubro de 1992 aos estudantes do 11° ano escolar. Essa avaliação é obrigatória no estado do Texas, USA, e se constitui de testes de leitura, matemática e escrita. Os testes de leitura (com 48 itens) e matemática (com 60 itens) são todos de múltipla escolha e dicotomicamente pontuados como certo ou errado. A base de dados se constitui das repostas de 193.000 respondentes aos testes de leitura e

[14] *Change assessment in clinical settings.*

matemática. Os achados indicam que tanto as estatísticas dos itens quanto as proficiências derivadas da TCT e da TRI são bastante comparáveis. Os resultados indicaram também que a invariância das estatísticas dos itens ao longo das formas é tão boa para os métodos tradicionais de escalonamento quanto para os métodos da TRI. Por outro lado, Salle (2009) compara, em um teste de química aplicado aos alunos do ensino médio, a pontuação clássica e a obtida pela TRI por meio de duas amostras equivalentes de alunos e duas formas do teste. Os resultados foram comparados considerando as dificuldades dos itens, consistência interna (fidedignidade) dos testes e erros das medidas de proficiências. O modelo da TRI utilizado foi o modelo de Rasch. Apontando menor erro de medida, portanto, nas pontuações da TRI.

Fan (1998), em seu estudo baseado em dados reais, achou alta correlação (maior que 0,96) entre as pontuações dos respondentes obtidas por métodos tradicionais e as pontuações obtidas pela TRI e, da mesma forma, alta correlação entre as medidas de dificuldade dos itens obtida pelos métodos tradicionais e pela TRI (maior que 0,90). Ele também não encontrou grande evidência de uma grande invariância dos parâmetros dos itens calculados por meio da TRI com relação ao da teoria clássica. Resultado similar foi encontrado por Courville (2005) em outro estudo empírico.

Uma limitação dos estudos empíricos é que os valores dos parâmetros não podem ser manipulados e os valores verdadeiros não são conhecidos. MacDonald e Paunonen (2002) realizaram um estudo de Monte Carlo (simulação intensiva) no qual eles controlaram a variação das dificuldades e as discriminações dos itens. De forma similar a Fan (1998) e Courville (2005), eles encontraram alta correlação entre as dificuldades dos itens estimadas pelos métodos da TRI. No entanto, no caso da discriminação, uma superioridade na TRI foi observada.

Grosso modo, então, os estudos indicam que, dependendo do contexto e das características dos testes, os métodos tradicionais podem até ter desempenho similar aos métodos da TRI. Estes autores desconhecem resultados em contrário, isto é, em que a forma de pontuação clássica possa ser superior à obtida pela TRI, a não ser aqueles em que se ficou demonstrado que os modelos e os métodos da TRI empregados foram inadequadamente utilizados. Por isso, tem-se aqui a convicção de que utilizar os métodos preconizados pela TRI num teste com a dimensão e a importância do ENEM é mais apropriado porque conduz a resultados mais justos no geral, embora menos transparentes para a sociedade.

Qualidade das medidas produzidas

Uma preocupação importante na Psicometria, tanto no que se refere à TCT quanto à TRI, é garantir a qualidade das medidas produzidas. No Brasil, não existe previsão legal que regule amplamente um sistema de testagem de tal forma que abarque o uso da TRI.

Tampouco existe no Brasil uma regulação ou um conjunto de sugestões propostas por alguma entidade científica ou profissional Nacional. Nos EUA, existem guias ou *standards* que constituem recomendações de algumas entidades científicas para garantir a qualidade das medidas e dos processos de testagem. O mais importante são os *standards* organizados pela *American Psychological Association* (APA) e pela *American Educational Research Association* (AERA), sendo o mais recente de 2014.

Aqui reportamos algumas características importantes do *Educational Testing Service (ETS) Standards for Quality and Fairness*, por estar disponível para consulta livre e por ser muito semelhante ao da APA e AERA. Segundo o documento (EDUCATIONAL TESTING SERVICE, 2015), os *standards* proveem padrões de excelência que são usados pelo *staff* da ETS ao longo do processo de planejamento, desenvolvimento e entrega das medidas, de modo a prover resultados de testes "justos", "válidos" e "fidedignos". O último documento (de 2014) é relativamente extenso, composto de 13 capítulos, 76 *standards*, cada um correspondente a um aspecto relevante, desde a construção, aplicação e validação de um teste.

Foram selecionados pelo autor deste texto alguns *standards* que ele considera como muito relevantes para a presente discussão, por serem mais críticos na questão da pontuação do ENEM e porque servirão para orientar documentos futuros.

Fidedignidade dos Resultados dos Testes, Validade e Equalização/*Linking*

A fidedignidade é a propriedade de um teste produzir medidas consistentes, com níveis aceitáveis de erros de medida. Tanto a TRI quanto a Teoria Clássica proveem medidas da fidedignidade dos testes para atestar a acurácia dos resultados.

A título de exemplificação, os *standards* preconizados pela ETS preconizam que se usem métodos adequados, que sejam analisadas as suas implicações, **mas, sobretudo, que se informe ao usuário do teste sobre a imprecisão das medidas calculadas e como ele pode lidar com ela**.

Uma maneira de quantificar a quantidade do erro de medida de uma estimativa de proficiência é usar o chamado erro padrão da medida. O erro padrão provê uma estimativa do erro que está presente nos escores obtidos pelo método de pontuação do teste devido à natureza da imperfeição de qualquer medida na natureza e, em particular, na medida da proficiência cognitiva obtida por meio de qualquer tipo de teste.

No caso do LSAT, por exemplo, o erro padrão é usado para construir intervalos (*score bands*) que podem ser usados para reportar os escores, de tal forma que a incerteza associada à medida é reportada juntamente com os resultados. Isto será visto em uma exemplificação dos resultados de um teste na seção 4.3 (Law School Admission Council, 2017b).

Muitos outros fatores podem afetar os resultados nos testes em um dia em particular, como a motivação, a saúde física e mental, a ansiedade produzida por diferentes fatores, algum transtorno produzido por uma obra ou pelo trânsito etc. Esses outros fatores, normalmente, não têm como serem levados em consideração no cálculo do erro padrão.

A fidedignidade de todo um teste pode ser estimada por diferentes medidas como, no caso de pontuações clássicas, pelo alfa de Cronbach e, no caso de pontuações obtidas pela TRI, a função de informação do teste.

Grosso modo, a validade de um teste consiste em estabelecer que aquilo que o teste mede é, de fato, o que se deseja que ele meça. Especificamente, existem muitos tipos de validades, citam-se aqui alguns tipos mais importantes: Validade de Constructo (o teste mede as habilidades e competências que deveriam medir); Validade de Conteúdo (o teste mede os conteúdos apropriadamente); Validade Preditiva (o teste prevê o sucesso do candidato razoavelmente, no caso de testes de admissão, por exemplo, o sucesso na conclusão do curso); Validade de consequências (o teste apresenta consequências adversas mínimas); Validade externa (o teste tem associação elevada com outras medidas do mesmo constructo).

Mesmo que seja impossível realizar estudos de validade amplos para todos esses tipos, em pouco tempo, deve ser uma preocupação constante a realização desses estudos ao longo do tempo, informando aos usuários das medidas sobre eles.

Outro *standard* da ETS (4.3: *Obtaining and Documenting the Evidence*) enfatiza a importância de se investigar e documentar a validade do teste.

Estudos de validade podem ser conduzidos tanto pelos responsáveis pelos testes, até para justificar a qualidade do seu produto, quanto pelos usuários (no caso as Universidades), para determinar se a escolha é apropriada para o sucesso acadêmico do indivíduo.

Já se apresentaram os conceitos de equalização e de *linking*. Vai-se, apenas, fazer a apresentação dos *standards* 8.1 a 8.5[15], sobre a importância de se documentar os métodos de equalização utilizados.

Embora não preconize especificamente os métodos de comparabilidade que devem ser usados, os *standards* deixam clara a necessidade de explicitá-los para que possam, inclusive, ser criticados. Além disso, enfatizam a necessidade de explicitar o erro de medida para que o usuário faça uso apropriado da medida.

Finalmente, os *standards* abordam outras questões importantes como sigilo, segurança na aplicação e nos resultados, documentação, conferência da pontuação obtida, entre outras.

Exemplos de testes de admissão ao ensino superior

Alguns testes de aptidão e conhecimentos gerais

Nesta seção selecionaram-se alguns exames e testes para serem analisados mais detalhadamente. Nos EUA, o uso de testes padronizados de aptidão (*aptitude*) e/ou de conhecimentos gerais são os mais comuns usados nos processos de admissão das universidades. Há, ainda, testes de aptidão ou conhecimentos específicos (*subject*) muito usados em disciplinas ou áreas específicas adicionalmente aos testes mais gerais. Existem muitos, mas destacamos como os testes de aptidão e/ou conhecimentos gerais nos EUA mais tradicionais os seguintes:

I. O *Scholastic Aptitude Test* (**SAT**), administrado pelo *College Entrance Examination Board* desde 1926, é um teste que tem o seu foco na habilidade de raciocínio (Liu; Harris; Schimidt, 2007). Segundo o *College Board*, o SAT mede as habilidades e os conhecimentos que os pesquisadores mostram que são os mais importantes para o sucesso no curso superior

[15] *Standard 8.1: Using Appropriate Equating or Linking Methodologies; Standard 8.2: Documenting Equating or Linking — Population and Comparability; Standard 8.3: Documenting Equating or Linking; Linking — Data Collection Design Describe the data collection design for the equating or linking study and state explicitly the assumptions implied by the use of that design.; Standard 8.4: Documenting Equating or Linking — Statistical Procedures; Standard 8.5: Documenting Equating or Linking — Results* (EDUCATIONAL TESTING SERVICE, 2015).

e na carreira[16]. Ele inclui as seguintes seções: Leitura e escrita, baseadas em evidências, e Matemática. Além disso, possui seções de ensaios (ou redações) opcionais, que medem habilidades de leitura, análise e escrita.

Existem, ainda, os chamados "SAT *Subjects*", que medem conhecimentos e habilidade em disciplinas e áreas específicas. Há, no momento, 20 testes destinados a diferentes disciplinas ou áreas do conhecimento divididas em cinco grandes áreas: Inglês, História, Linguagens e Matemáticas. Cada teste tem uma hora de duração e são todos de múltipla escolha e pontuados numa escala de 200-800 pontos (The College Board, c2017a). Os testes são realizados em diferentes datas e, assim como no caso do SAT, o candidato pode realizar o teste sempre que quiser e pode escolher a sua pontuação que julgar mais conveniente para ser enviada às Universidades[17]. Também existem testes destinados a alunos de séries inferiores da educação básica, conhecidos como *Preliminar SAT tests* (PSATs).

O SAT é pontuado utilizando-se **a teoria clássica do teste**, no entanto, os resultados das diferentes seções e testes são escalonados (*scaled*) para escalas comparáveis (isto é, são equalizados) de tal forma que, segundo seus organizadores, os resultados das diferentes formas (isto é, dos testes realizadas em diferentes datas) possam ser utilizados pelas universidades indistintamente. Os métodos de equalização empregados pelo SAT são genericamente descritos por Liu, Harris e Schimidt (2007). Portanto, não são os pontos brutos que são apresentados, **mas os escores escalonados para a escala do SAT.**

Os alunos não respondem a uma única forma de teste. Isto é, diferentes grupos de itens são apresentados a diferentes grupos de alunos. No entanto, um procedimento espiralar de aplicação do teste (*spiraling procedure*) garante que os grupos sejam equivalentes em proficiências. Assim, os escores de uma forma escolhida são equalizados para a escala do SAT por meio de um método conhecido como *Nonequivalent Anchor Test Design* (NEAT), sendo os escores das demais formas equalizadas com esta por meio de um *Random/ Equivalent Goup Design* (EG).

Tabelas de conversão entre os pontos brutos alcançados em um teste específico e os pontos equivalentes das escalas são fornecidas de tal forma que o candidato pode conferir a sua pontuação. No Quadro 1, apresenta-se a relação dos resultados que são reportados pelo SAT.

[16] *The SAT measures the skills and knowledge that research shows are the most important for success in college and career.*

[17] Ver: The College Board, 2017.

Quadro 1 – Relação dos resultados reportados pelo SAT

SAT Score Reported	Details	Score Range
Total score	Sum of the two section scores.	400-1600
Section scores	Evidence-Based Reading and Writing, and Math.	200-800
Test scores	Reading, Writing and Language, and Math.	10-40
SAT Essay scores*	Reading, Analysis, and Writing.	2-8
Cross-test scores	Analysis in History/Social Studies and Analysis in Science. Based on selected questions in the Reading, Writing and Language, and Math Tests. These scores show how well you use your skills to analyze texts and solve problems in these subject areas.	10-40
Subscores	Reading and Writing and Language: Command of Evidence and Words in Context. Writing and Language: Expression of Ideas and Standard English Conventions. Math: Heart of Algebra, Problem Solving and Data Analysis, and Passport to Advanced Math.	1-15

* The SAT Essay is optional.

Fonte: The College Board (c2017c).

Os escores são apresentados na escala do SAT e, em alguns casos, comparativamente em percentis (*percentile ranks*), para as distribuições dos escores em relação aos alunos que fizeram os testes nos últimos anos.

Quando não concorda com a pontuação recebida, o candidato pode recorrer, solicitando uma verificação, conforme procedimento específico.

Para se ter uma ideia, cerca de 2,4 milhões de indivíduos realiza o SAT e 1,2 milhões o ACT todos os anos.

II. O *Achievement College Test* (**ACT**) é um teste de avaliação com características similares ao SAT, realizado pela ACT Inc. Porém, segundo Liu, Harris e Schimidt (2007), ele tem foco na realização (*achievement*)[18] e inclui quatro seções obrigatórias de testes de múltipla escolha de Inglês, Matemática, Leitura e Ciências. Assim como o SAT, o ACT é pontuado por meio de estatísticas clássicas e suas pontuações brutas são escalonadas (por meio de um procedimento de equalização) para permitir a comparabilidade de resultados; o método é genericamente descrito em Liu, Harris e Schimidt (2007). Devido ao método de equalização utilizado, é possível produzir uma tabela que mostra a equivalência entre os pontos brutos alcançados pelos candidatos no teste e os pontos equivalentes na escala do ACT. Existem também tabelas de comparação entre os escores do ACT e SAT, possivelmente construídas por meio de uma projeção entre as duas escalas.

[18] They describe essential skills and knowledge students need to become ready for college and career.

III. O *General Test* (**GRE**) é realizado pela ETS e é um teste normalmente exigido para a admissão em programas de pós-graduação nos EUA ou em cursos de graduação considerados de 2º ciclo. Possui duas versões: impressa e computadorizada. O GRE mede raciocínio verbal, raciocínio quantitativo, pensamento crítico e competências em escrita. É organizado em três baterias: *Verbal Reasoning, Quantitative Reasoning* e *Analytical Writing*. Em geral, o GRE é utilizado pelas universidades de forma complementar às informações da graduação. O teste é pontuado da seguinte forma: as duas primeiras baterias são divididas em duas seções, com 20 itens cada, e os pontos brutos convertidos numa escala de 130-170, com incrementos de 1 ponto, e a última seção em uma escala de 0-6 pontos, com incremento de 0,5 pontos. Para as duas primeiras, os escores brutos (*raw scores*) são computados diretamente das respostas corretas e convertidos para a escala específica do GRE por meio de um processo de equalização (*equating*).

Publicações de autores da ETS afirmam que até a década de 1980 se utilizava o método de Tucker (*Tucker Equating Procedure*), ver, por exemplo, Mckinley e Kingston (1987). A partir da década de 1990, provavelmente com a introdução da versão computadorizada tradicional em *Computer Adaptive Test* (CAT), passaram-se a utilizar os métodos da TRI para a pontuação e a equalização dos escores (MISLEV et al, 2006). Com a mudança introduzida em 2011, com a substituição do CAT tradicional por uma versão *Multistage Adaptive Test* (MST), não só a escala como o método de equalização mudou. Pelo que é apresentado no site da ETS, e como primeiro são calculados os pontos brutos e depois convertidos para escala do GRE, é muito provável que, a exemplo dos demais testes de admissão realizados pela ETS e outros nos EUA, se utilize um método de equalização baseado num procedimento do tipo *Item Response Theory* (*IRT) True Score Equating* (KOLEN; BRENAN, 2004*).*

Os procedimentos de equalização acima utilizam a Teoria da Resposta ao Item para calcular os parâmetros dos itens e as proficiências dos examinandos, conforme uma determinada escala de referência. No entanto, os escores calculados pela TRI não são utilizados diretamente na pontuação dos alunos. Dada uma forma do teste, eles são usados, juntamente com os parâmetros dos itens presentes na forma, para estabelecer uma relação funcional entre as proficiências calculadas pela TRI e o chamado *true score* correspondente àquela forma. Como essa mesma relação já é conhecida para o caso de uma forma de referência ou de origem da escala, constrói-se uma tabela associando os *true scores* das duas formas (a forma que se quer equalizar e a forma de referência). Por meio dessas tabelas, os escores brutos dos examinandos

que responderam à forma presente do teste são transformados para a escala da forma de referência.

Assim, embora se utilize a TRI, isto é feito de forma indireta no processo de escalonamento. Por outro lado, esse procedimento garante que o ordenamento dos alunos em uma determinada forma, segundo os pontos brutos, não se altere em relação à pontuação convertida. A razão pela qual a maior parte dos testes de admissões nos EUA utiliza esse método de equalização não está absolutamente clara. Há estudos que procuram mostrar que a precisão dessas medidas é muito próxima da precisão de medidas obtidas diretamente pela TRI para diferentes tipos de modelos. Além disso, dada a tabela de conversão, é possível ao examinando conferir sua pontuação em função do número de acertos no teste e, assim, recorrer, se for o caso de eventuais discrepâncias. Pode-se também supor que esse tipo de procedimento atenda à legislação sobre testagem para admissão no ensino superior presente em vários estados daquele país.

Para a seção de escrita analítica, cada ensaio recebe um escore 0 a 6 com base na tarefa especificada no teste por um corretor humano. O ensaio é também submetido a um programa corretor. Se os escores coincidem aproximadamente, então a média de ambos é tomada como a média final, caso contrário, um segundo corretor humano avalia o ensaio e o escore final é o escore médio dos dois corretores humanos (EDUCATIONAL TESTING SERVICE, c2017).

Existem ainda os GRE *subject tests* que são usados para medir o conhecimento em áreas específicas: Biologia, Química, Literatura em Inglês, Matemática, Física e Psicologia.

Alguns testes de aptidão e conhecimentos específicos

Entre os testes de aptidões e conhecimentos específicos, destacam-se:

I. O *Test of English as a Foreign Language* (**TOEFL**) é a avaliação em língua inglesa mais usada em todo o mundo com finalidade admissional de alunos estrangeiros e é aceito por Universidades e organizações em 130 países diferentes, incluindo EUA, Reino Unido, Austrália, Nova Zelândia (Educational Testing Service, 2011). A versão mais recente do TOEFL iBT foi lançada em 2005. O TOEFL é de responsabilidade do importante centro de testagem ETS e é realizado desde 1965. O teste é organizado em quatro seções que são pontuadas em uma escala, conforme descrição na tabela a seguir:

Tabela 1 – Escala TOEFL

Score	Scale	Reliability Estimate	SEM
Reading	0-30	0.85	3.35
Listening	0-30	0.85	3.20
Speaking	0-30	0.88	1.62
Writing	0-30	0.74	2.76
Total	0-120	0.94	5.64

Fonte: *Educational Testing Service*, 2011.

As escalas para as medidas dos testes do TOEFL iBT foram estabelecidas para um intervalo de 0 a 30 para as 4 seções, conforme descrito na Tabela 1, e todas têm o mesmo peso na medida total da habilidade em Inglês. O escore total é simplesmente a soma de todos os escores obtidos nas diferentes seções.

Assim, os escores brutos das seções são computados e, posteriormente, são convertidos também por meio do método do escore verdadeiro, *IRT True Score Equating* (Kolen; Brennan, 2004), para estabelecer a relação entre os escores brutos do último teste com os escores brutos de uma forma de referência e, como existe uma equivalência dos escores brutos da forma de referência com os níveis da escala, é possível fornecer uma equivalência entre os escores brutos da forma atual com os pontos de 0 a 30 da escala original, aproximadamente.

Especificamente, Way e Reese (1990) afirmam que até então o modelo da TRI utilizado seria o Modelo Logístico de 3 Parâmetros.

II. O **LSAT** é, segundo o site da organização *Law School Admission Council* (LSAC), responsável pelo teste requerido na maioria das escolas de direito dos EUA (Law School Admission Council, c2017a). Ainda segundo a organização, atualmente, 221 escolas de direito (*law schools*) nos EUA, Canadá e Austrália são membros do conselho e utilizam o LSAT. O teste consiste de cinco seções de 35 minutos com questões de múltipla escolha. Quatro dessas seções contribuem para a pontuação total no teste: uma que avalia a compreensão em leitura, uma que avalia o raciocínio analítico e duas que avaliam o raciocínio lógico. A seção, cujos resultados não são usados na pontuação, é normalmente usada para testar novos itens ou pré-equalizar novas formas do teste. Uma seção de 35 minutos que avalia a escrita é aplicada ao final do teste, e uma cópia do texto é enviada diretamente à Universidade. Segundo os organizadores, o LSAT é projetado para medir as competências que são consideradas essenciais na escola de direito: a leitura e compreensão de textos complexos com acurácia e *insights*; a

organização e administração das informações e a habilidade de construir inferências razoáveis a partir delas; a habilidade de pensar criticamente e a análise e avaliação do raciocínio e dos argumentos dos outros. Os escores de cada seção são pontuados em uma escala de 120-180 pontos, no entanto, além do último escore, são apresentados também os últimos resultados de todos os testes realizados pelo examinando em diferentes momentos (até o limite de 12), a partir de uma determinada data (atualmente, junho de 2012), incluindo cancelamentos e não comparecimentos. Além disso, é apresentado um intervalo de escores denominado *score band*. Segundo o LSAT, o *score band* reflete a precisão da medida do LSAT e é apresentado para enfatizar que o escore LSAT é uma estimativa para o verdadeiro e desconhecido nível de proficiência. Trata-se de um intervalo de confiança construído dentro de um nível de confiança de 68%, admitindo-se normalidade para os escores. Como o erro padrão do *score* também é fornecido, intervalos com níveis de confiança maiores podem ser construídos com base na hipótese de normalidade do estimador de proficiências. Também é apresentado o percentil do escore alcançado, calculado a partir dos escores dos candidatos que fizeram o teste nos três anos anteriores.

Na sua versão em papel e lápis, o LSAT equaliza os escores da seguinte forma: os pontos brutos são transformados na escala LSAT usando IRT e *true-scoring equating* (cf. Van der Linden; Pashley, 2013). Segundo Weissman (2011), em artigo patrocinado pelo LSAC, o modelo usado na equalização é o modelo logístico de 3 parâmetros.

III. O *Medical College Admission Test* (**MCAT**) é administrado e desenvolvido pela *Association of American Medical Colleges* (AAMC) e utilizado para ingresso nos cursos de medicina. Seu processo de pontuação também produz uma transformação dos escores brutos para uma escala por meio de um processo de equalização. Os métodos de pontuação e equalização são muito parecidos com os empregados no LSAT. O teste é organizado em quatro seções: *Biological and Biochemical Foundations of Living Systems; Chemical and Physical Foundations of Biological Systems; Psychological, Social, and Biological Foundations of Behavior;* e *Critical Analysis and Reasoning Skills*. Assim, o número de respostas corretas em cada seção é transformado para uma escala entre 118 e 132 pontos. Assim como no LSAT, os percentis dos escores são apresentados sendo atualizados anualmente em função do desempenho dos candidatos nas últimas avaliações. O Anexo II contém um exemplo da apresentação do resultado de

um indivíduo. É apresentado o escore estimado, o intervalo de confiança de 68% do escore verdadeiro e o percentil correspondente na distribuição de proficiências nas últimas avaliações. Um detalhe importante é que os escores são apresentados sem casas decimais.

Grandes exames nacionais

Por sua semelhança com o ENEM, em forma e contexto, apresentam-se dois exemplos de grandes provas nacionais.

I. La Prueba Saber, na Colômbia, é realizada pelo Ministério da Educação (MINEDUCACIÓN), se iniciou em 1968 e tem sido reformulada desde então. Atualmente, se presta aos seguintes objetivos: a) selecionar os alunos para a educação superior; b) monitorar a qualidade da formação dos estabelecimentos que oferecem educação de nível médio; e c) produzir medidas para a estimação do valor agregado da educação superior. Atualmente, o exame de Estado Saber 11º se compõe de cinco provas: Matemática, Leitura Crítica, Sociais e cidadania, Ciências Naturais e inglês. Além dos escores nestas competências, se produz medidas de duas subcompetências: cidadania e raciocínio quantitativo. Os escores são computados por meio do uso de modelos de Rasch, da TRI. Cada prova é computada numa escala de 0 a 100 e tem como referência os participantes do teste em 2014, para os quais a média foi estabelecida em 50 pontos e o desvio padrão em 10 pontos da escala. Também são apresentados os percentis em que as pontuações dos alunos se encontram (Instituto Colombiano para la Evaluación de la Educación, 2014).

II. La Prueba de Selección Universitaria (**PSU**), no Chile, contém testes de Matemática e Linguagem e comunicação, que são obrigatórios, e testes de ciências (biologia, física, química e técnico-profissional) e história, geografia e ciências sociais, que são eletivos. Os alunos escolhem as eletivas de acordo com os requisitos das universidades. A prova é coordenada pelo *Departamento de Evaluación, Medición Y Registro Educacional* (DEMRE) da *Univesidad de Chile*. A prova é pontuada classicamente da seguinte forma: calcula-se a pontuação bruta e esta é transformada (escalonamento) para uma escala padrão com média 500 e desvio padrão 110, por meio da padronização de cada escore, tomando como valores, para a padronização, a média e o desvio-padrão dos pontos brutos dos que fizeram a prova naquele ano. Um cálculo adicional não explicitado faz com que os valores sejam fixados entre 150 e 850 pontos. Percentis dos

escores padronizados também são fornecidos às universidades (UNIVERSIDAD DE CHILE, [c2017]a). O processo de equalização aqui empregado pressupõe que os grupos que realizam os testes em cada ano são muito semelhantes segundo sua distribuição de escores de proficiências.

O ACESSO AO ENSINO SUPERIOR NO BRASIL E O ENEM

O sistema de ingresso na educação superior no Brasil é, naturalmente à semelhança de países com igual nível de desenvolvimento econômico, um sistema bastante seletivo. No entanto, até pouco tempo atrás, era bastante descentralizado.

Em 1998, o governo federal do Brasil criou o Exame Nacional do Ensino Médio como um instrumento para avaliar o desempenho dos estudantes no término da educação básica. A finalidade do exame era auxiliar o Ministério da Educação (MEC) na elaboração de políticas pontuais e estruturais de melhoria do ensino brasileiro por meio dos Parâmetros Curriculares Nacionais (PCNs) do ensino médio e fundamental. Durante mais de dez anos, este exame foi usado única e exclusivamente para avaliar as habilidades e competências de concluintes do ensino médio, sem o objetivo de selecionar para o ensino superior. Os exames de seleção para o ensino superior, os concursos vestibulares, eram formulados por equipes locais país afora e formatos diferentes ocorriam nas diversas universidades e centros de ensino superior.

O ingresso no ensino superior no Brasil, particularmente o ingresso nas escolas públicas, sofreu mudanças substanciais a partir de 2009, quando o ENEM passou por uma reformulação metodológica, com a finalidade de se incentivar a sua utilização como exame de seleção unificada nos processos seletivos, principalmente, das universidades públicas federais. Além disso, o nível de centralização do sistema de ingresso se elevou significativamente com a implantação do Sistema de Seleção Unificada (Sisu), que consiste em uma plataforma online por meio da qual as instituições de ensino superior ofertam vagas em cursos de graduação a estudantes que serão selecionados com base na nota obtida no último Enem.

Assim, em poucos anos, o ensino superior brasileiro transitou de um sistema de seleção bastante descentralizado, em que alunos se aplicavam diretamente e de forma independente para cada instituição, para um sistema razoavelmente centralizado, em que parte expressiva deles se submete a algumas

opções de curso para um regulador central e o critério de seleção é baseado nos resultados de um exame de ingresso comum, ainda que parcialmente, no caso de algumas instituições.

Genericamente, o objetivo da política apresentado pelo Ministério da Educação seria a democratização do acesso ao ensino superior de qualidade, possibilitando igualdade na diversificação de escolha por parte de todos os candidatos e acesso a um rol maior de instituições. Apontava-se, ainda, por conta da redução de diversos custos incorridos pelo aluno com a implantação do novo ENEM e do SISU que uma das consequências dessa mudança do sistema de seleção seria o aumento da mobilidade estudantil interna.

Por outro lado, esperava-se a melhoria dos níveis de habilidades dos alunos ingressantes nos diferentes cursos e uma repercussão positiva no ensino médio em geral, pois traria, em certo sentido, um padrão de qualidade a ser perseguido pelas escolas. Em contrapartida, como consequência adversa, esperava-se um aumento do dispêndio com o exame por parte do governo federal e o aumento da evasão nos primeiros anos do Ensino Superior. Há alguma evidência de que houve aumento nas taxas de mobilidade estudantil, mas também nos níveis de evasão de diversos cursos (ver, LI, 2016). No entanto, até agora, realizaram-se poucos estudos para uma avaliação mais abrangente dos impactos do ENEM.

CONCLUSÃO

A atual prova do ENEM, em parte, copia os testes dos vestibulares tradicionais que o antecederam, no que se refere à concepção do que se deve medir, com foco na realização ou êxito (*achievement*) relacionado a conhecimentos específicos (*subjects*) organizados nas quatro disciplinas, avaliadas por questões objetivas de múltipla escolha. Além disso, contém uma prova de redação mensurada a partir de critérios e procedimento padronizados por avaliadores.

Embora o ENEM original fosse concebido para ser uma prova que mensurava mais habilidades cognitivas do que conhecimentos específicos, com o tempo ele se tornou mais parecido com os testes dos vestibulares tradicionais. Outras características psicométricas também se alteraram ao longo dos anos, como o nível de dificuldade dos itens. Essas modificações, em nossa opinião, ocorrem, pelo menos, por dois fatores importantes. O primeiro diz respeito à concepção e à prática avaliativa dos profissionais de certos cursos

das Universidades, mais acostumados em focar a avaliação no conhecimento do que nas habilidades. Logo, para facilitar a adesão ao uso dos resultados dos testes, nada mais natural que aproximar os testes do ENEM dos testes de vestibular tradicionais. Em segundo, também para atender à necessidade das universidades que utilizam o ENEM, e tendo em vista a acirrada disputa por uma vaga em determinados cursos, especialmente nas universidades públicas, o teste passa a introduzir itens mais difíceis numa tentativa de aumentar a precisão das medidas nos níveis mais elevados de proficiências, o que de fato vem ocorrendo e se reflete nas características dos testes que serão discutidas oportunamente neste e no outro documento a ser elaborado.

No entanto, por ser um grande exame nacional, com uma abrangência territorial imensa, o ENEM não poderia utilizar o sistema de pontuação habitualmente utilizado nos vestibulares tradicionais. Isso ocorre, por exemplo, como já foi inclusive observado em edições de anos anteriores do ENEM, pela ocorrência de determinados eventos não previsíveis, como intempéries e catástrofes, que naturalmente demandam a aplicação de novos testes ao grupo de indivíduos afetados pelos eventos, quer seja porque não tenham podido realizar as provas como os demais, quer seja porque tivessem suas provas anuladas por algum motivo. Naturalmente, pelos custos envolvidos, a anulação de todo o ENEM é, senão infactível, pelo menos irracional. Portanto, o ENEM, como grande prova nacional que é, necessita naturalmente de um sistema de pontuação que permita a comparabilidade das medidas produzidas em testes diferentes, do contrário se tornaria inviável a sua realização.

A experiência brasileira em mensuração e escalonamentos de testes admissionais era pouco expressiva até o novo ENEM. Por outro lado, havia considerável experiência na mensuração de testes de avaliação educacional em larga escala, sendo o exemplo mais paradigmático, a mensuração dos resultados do SAEB e da Prova Brasil, pelo menos desde meados da década de 90.

Em contrapartida, embora exista legislação sobre os concursos públicos de admissão, que de certa forma abrange os concursos vestibulares, não há especificidade da lei no que tange ao processo de mensuração em si. Por outro lado, o escopo da lei existente sobre concursos parece não atingir plenamente um teste com a dimensão, necessidades e tecnologia do ENEM.

O sistema de pontuação dos resultados do novo ENEM, a partir de 2009, então deriva dos sistemas de mensuração utilizados nas avaliações em larga escala empregados no Brasil e no mundo para a avaliação em larga escala de sistemas educacionais. A escolha, com pouco debate, recaiu sobre os mesmos métodos de cálculo das medidas e equalização dos resultados empre-

gados no caso da mensuração do SAEB/Prova Brasil e de outras avaliações de sistemas educacionais no país, muito provavelmente porque essa era a *expertise* dos especialistas contratados pelo INEP. Quando se analisam os testes utilizados na admissão ao ensino superior no mundo (ver seção 5,) verifica-se a singularidade dos métodos utilizados na pontuação do ENEM.

Isso não significa, *a priori*, que as medidas do ENEM são piores ou melhores do que as empregadas no SAT ou no LSAT, por exemplo. Ambos os sistemas, SAT e LSAT, utilizam métodos diferentes para o escalonamento de seus resultados. Mas ambos utilizam métodos que permitem a transformação de seus escores brutos para uma escala comparável pelo método conhecido como *IRT True Score Equating,* conforme apresentado na seção 5. A tabela de transformação é fornecida aos candidatos de tal forma que, por meio dela, se pode conferir sua pontuação final. Não está explícito, mas parece que essa metodologia tem sido utilizada, em parte, para satisfazer a legislação americana para testes admissionais de alguns estados (Greer, 1984), como a dos estados da Califórnia e Nova York. De qualquer forma, esse expediente permite maior controle do candidato de sua pontuação final.

REFERÊNCIAS

BAKER, F. B.; KIM, S. **Item Response Theory**. 2 ed. Statistics: Textbooks and Monographs, v. 129. New York, USA: Marcel Dekker, Inc., 2004.

BIRNBAUM, A. Some Latent Trait Models and Their Use in Inferring an Examinee´s ability. In: LORD, F. M.; NOVICK, M. R. **Statistical Theories of Mental Test Scores**. Reading, MA: Addison- Wesley Pub, 1968. p. 397-472.

COURVILLE, T. G. **An empirical comparison of item response theory and classical test theory item/person statistics**. (Unpublished doctoral dissertation). Texas: A&M University, 2005. Disponível em: <http://txspace.tamu.edu/bitstream/ handle/1969.1/1064/etd-tamu-2004B-EPSY-Courville-2.pdf?sequence=1>. Acesso em: 05 nov. 2007.

EDUCATIONAL TESTING SERVICE. **2014 ETS Standards for Quality and Fairness**. Princeton, NJ: ETS, 2015.

_____. **How the Test Is Scored**. c2017. Disponível em: <https://www.ets.org/gre/revised_general/scores/how/>. Acesso em: 04 out. 2017.

_____. Reliability and Comparability of TOEFL iBT™ Scores. **TOEFL iBT Research**, Princeton, NJ, Series 1, v. 3, 2011.

EDWARDS, D.; COATES, H.; FRIEDMAN, T. A survey of international practice

in university admissions testing. **Higher Education Management and Policy**, OECD, v. 24/1, 2012.

FAN, X. Item response theory and classical test theory: an empirical comparison of their item/person statistics. **Educational and Psychological Measurement**, Santa Barbara, CA, v. 58, n. 3, p. 357-381, June 1998.

GREER, D. G. "**Thuth-in-testing legislation**" An Analysis of political and legal consequences, and prospects. Houston TX: Institute for Higher Education Law and Governance, 1984.

HAMBLETON, R. K.; JONES, R. W. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. **Educational Measurement: Issues and Practice**, v. 12(3), p. 38-47, 1993.

INSTITUTO COLOMBIANO PARA LA EVALUACIÓN DE LA EDUCACIÓN. Sistema Nacional de Evaluación Estandarizada de la Educación. Alineación del examen SABER 11°. **Lineamientos generales 2014 – 2**. 2014. Disponível em: <http://www.icfes.gov.co/docman/instituciones-educativas-y-secretarias/saber-11/novedades/650-guia-lineamientos-generales-saber-11-2014-2/file?force-download=1>. Acesso em: 15 out. 2017.

JABRAYILOV, R.; EMONS, W. H. M.; SIJTSMA, K. Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. **Applied Psychological Measurement**, v. 40(8), p. 559-572, 2016.

KOLEN M.; BRENNAN R. L.; **Test Equating, scaling, and linking**. New York: Springer-Verlag, 2004.

LAW SCHOOL ADMISSION COUNCIL. Law School Admission Test (LSAT). Score Bands. c2017b. Disponível em: <https://www.lsac.org/jd/lsat/your-score/score-band>. Acesso em: 04 out. 2017.

LI, D. L. **O novo Enem e a plataforma Sisu**: efeitos sobre a migração e a evasão estudantil. 2016. Dissertação (Mestrado em Ciências) – Programa de Pós-Graduação em Economia do Departamento de Economia da Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2016.

LIU, J.; HARRIS, D. J.; SCHIMIDT, A. Statistical Procedures Used in College Admissions Testing. In: RAO, C. R.; SINHARAY, S. **Handbook of Statistics 26**: Psychometrics. Amsterdam: Elsevier, 2007, p. 1057-91.

LORD, F. M. **Applications of item response theory to practical testing problems**. Hillsdale: Lawrence Erlbaum, New York, 1980.

LORD, F. M.; NOVICK, M. R. **Statistical Theories of Mental Test Scores**. Reading, MA: Addison-Wesley, 1968.

MACDONALD, P.; PAUNONEN, S. V. A Monte Carlo comparison of item and person parameters based on item response theory versus classical test theory. **Educational and Psychological Measurement**, v. 62, p. 921-43, 2002.

MCGRATH, C. H. et al. **Higher Education Entrance Qualifications and Exams in Europe**: A Comparison. Brussels: European Union, 2014.

McKINLEY, R.; KINGSTON, N**. Exploring The Use of IRT Equating for the GRE Subject Test in Mathematics**. ETS Report, RR-87-2, 1987.

MISLEV R. et al. Concepts, Terminology and Basic Models of Evidence-Centered Design. In: WILLIAMSON D.; MISLEV R.; BEJAR I. **Automated Scoring of Complex Tasks in Computer-based Testing**. London: Ed. Lawrence Erlbaum, 2006, p. 15-48.

PALMER, N.; BEXLEY, E.; JAME, R. **Selection and participation in higher**

education university selection in support of student success and diversity of participation. Melbourne: CSHE, 2011.

SALLE, C. M. Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. **The International Journal of Educational and Psychological Assessment**, v. 1, Issue 1, p. 1-11, April 2009.

SARGEANT, C. et al. **INCA comparative tables**. London: International Review of Curriculum and Assessment Frameworks Internet Archive, 2012. Disponível em: https://www.nfer.ac.uk/what-we-do/information-and-reviews/inca/INCA-comparativetablesMarch2012.pdf. Acesso em: ?????

THE COLLEGE BOARD. **At a glance**. c2017a. Disponível em: <https://collegereadiness.collegeboard.org/sat-subject-tests/about/at-a-glance>. Acesso em: 04 out. 2017.

_____. **Score Choice**. c2017b. Disponível em: <https://collegereadiness.collegeboard.org/sat-subject-tests/scores/sending-scores/score-choice>. Acesso em: 04 out. 2017.

_____. **Score Structure**. c2017c. Disponível em: <https://collegereadiness.collegeboard.org/sat/scores/understanding-scores/structure>. Acesso em: 04 out. 2017.

_____. **Understanding Scores**. c2017d. Disponível em: https://collegereadiness.collegeboard.org/sat/scores/understanding-scores/interpreting. Acesso em: 04 out. 2017.

THISSEN, D.; WAINER, H. **Test Scoring**. Mahwah, New Jersey: Lawrence Erlbaum Associates Pub., 2001.

UNIVERSIDAD DE CHILE. Departamento de evaluación, medición y registro educacional. Prueba de Selección Universitaria. **¿Cómo se calcula el puntaje?** [c2017]a. Disponível em: http://www.psu.demre.cl/la-prueba/que-es-la-psu/calculo-puntaje-psu. Acesso em: 20 out. 2017.

UNIVERSIDAD DE CHILE. Departamento de evaluación, medición y registro educacional. Prueba de Selección Universitaria. **El proceso de admisión**. [c2017]b. Disponível em: http://www.psu.demre.cl/proceso-admision/. Acesso em: 24 out. 2017.

VAN DER LINDEN, W. J.; PASHLEY, P. J. Item Selection and Ability Estimation in Adaptive Testing. In: VAN DER LINDEN, W. J.; GLAS, C. A. W. (Eds.). **Elements of adaptive testing**. London: Kluwer Academic Pub., 2013. p. 2-26.

WAY, C.; REESE, C. An Investigation of the Use of Simplified IRT Models for Scaling and Equating the TOEFL Test. **ETS Research Report**, RR-90-29, 1990.

WEISSMAN, A. Optimizing Information Using the Expectation-Maximization Algorithm in Item Response Theory. **Law School Admission Council Research Report,** 11-01, March 2011.

Recebido em: 10 de dezembro de 2019
Aceito em: 19 de junho de 2020
Publicado em: 30 de junho de 2020