Yara da Silva Padilha

EVIDÊNCIAS DE VALIDADE PARA A BATERIA DE PROVAS DE RACIOCÍNIO - ELETRÔNICA

Apoio:

CAMPINAS
2021

Yara da Silva Padilha

EVIDÊNCIAS DE VALIDADE PARA A BATERIA DE PROVAS DE RACIOCÍNIO – ELETRÔNICA

Dissertação apresentada ao Programa de Pós-Graduação *Stricto Sensu* em Psicologia da Universidade São Francisco, Área de Concentração - Avaliação Psicológica, para obtenção do título de Mestre.

ORIENTADOR(A): RICARDO PRIMI

CAMPINAS
2021

157.93 Padilha, Yara da Silva.
P134e Evidências de validade para a Bateria de Provas de Raciocínio - eletrônica / Yara da Silva Padilha. – Campinas, 2021.
71 p.

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

1. Psicometria. 2. Testagem adaptativa computadorizada.
3. Teoria de resposta ao item. I. Primi, Ricardo. II. Título.

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

Educando para a paz

PROGRAMA DE PÓS-GRADUAÇÃO STRICTO SENSU EM PSICOLOGIA

Credenciado pela CAPES
Portaria n.º 177, de 25 de janeiro de 2002.
D.O.U. de 29 de janeiro de 2002, págs. 51 a 60

ATA DO EXAME DE ARGUIÇÃO FINAL DA ALUNA YARA DA SILVA PADILHA – RA 004201805647 – DO PROGRAMA DE PÓS-GRADUAÇÃO STRICTO SENSU EM PSICOLOGIA – MESTRADO DA UNIVERSIDADE SÃO FRANCISCO – USF.

Aos vinte e dois dias do mês de outubro do ano de dois mil e vinte e um, às oito horas e quinze minutos, por videoconferência, reuniu-se a Comissão da Banca Examinadora de Pós-Graduação do Programa de Pós-Graduação Stricto Sensu em Psicologia – Mestrado – da Universidade São Francisco para avaliação do Relatório de Atividades e Pesquisa intitulado **"EVIDÊNCIAS DE VALIDADE PARA A BATERIA DE PROVAS DE RACIOCÍNIO - ELETRÔNICA"** apresentado pela aluna Yara da Silva Padilha, graduada em Psicologia pela Universidade Federal Fluminense que concluiu os créditos exigidos para a obtenção do Grau de "Mestre em Psicologia" segundo os registros constantes no Núcleo de Registro e Controle Acadêmico. Os trabalhos foram iniciados às oito horas e quinze minutos pelo Prof. Dr. Ricardo Primi Orientador da Candidata e Presidente da Banca Examinadora, constituída pelos seguintes Professores: Ricardo Primi Doutor em Psicologia Escolar e do Desenvolvimento Humano pela Universidade de São Paulo – USP; Carlos Henrique Sancineto da Silva Nunes Doutor em Psicologia pela Universidade Federal do Rio Grande do Sul – UFRGS e Felipe Valentini Doutor em Psicologia Social, do Trabalho e das Organizações pela Universidade de Brasília - UnB. A Banca Examinadora encerrou os trabalhos às dez horas e vinte minutos, e considerou a candidata aprovada. E, para constar, eu, Prof. Dr. Ricardo Primi lavrei a presente Ata, que após ser lida e aprovada, segue assinada eletronicamente por mim e pelos demais membros da Banca Examinadora, nos termos da MP nº 2.200-2 de 24/08/2001, que institui a Infraestrutura de Chaves Públicas Brasileira - ICP-Brasil.

Parecer da banca:

Campinas/SP, 22 de outubro de 2021.

Prof. Dr. Ricardo Primi - Orientador e Presidente
Universidade São Francisco

Prof. Dr. Carlos Henrique Sancineto da Silva Nunes
Universidade Federal de Santa Catarina

Prof. Dr. Felipe Valentini
Universidade São Francisco

Agradecimentos

Agradeço a Deus pela oportunidade de realizar este sonho, que me acompanha desde a graduação.

Agradeço aos meus familiares, em especial ao meu pai, João, que acreditou em mim e não mediu esforços para tornar esse sonho possível.

Agradeço ao meu professor orientador Ricardo Primi, pela atenção e pela paciência ao longo do caminho, para além de todo conhecimento compartilhado.

Agradeço aos professores doutores que integraram a banca de defesa de mestrado, Felipe Valentini e Carlos Nunes, pelas contribuições a esta dissertação. Também sou grata a todo o corpo docente do Programa de Pós-Graduação em Psicologia da USF, que colaboraram para minha formação.

Agradeço aos meus colegas da USF, em especial a minha primeira turma de mestrado: Leonardo, Ana, Mayara, Gustavo, Marcela, Fernanda, Érica e Andreza, por todo o suporte social, para além das contribuições acadêmicas.

Por fim, agradeço pelo apoio inicial da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento (processo nº 2019/00022-0), o qual foi essencial para a realização dessa pesquisa.

Resumo

Padilha, Y. S. (2021). *Evidências de validade para a Bateria de Provas de Raciocínio – eletrônica.* Dissertação de Mestrado, Programa de Pós-Graduação Stricto Sensu em Psicologia, Universidade São Francisco, Campinas.

O uso de testes psicológicos é adotado em diferentes contextos, colaborando para a prática profissional do psicólogo. Embora o formato convencional de testes prevaleça em processos de avaliação psicológica, o avanço tecnológico possibilitou o desenvolvimento da Testagem Adaptativa Informatizada (CAT) por meio da Teoria de Resposta ao Item (TRI), que torna a testagem mais breve, mantendo a sua eficiência. CATs já são empregados em âmbito internacional de forma expressiva, mas ainda são recentes no Brasil. Nesse sentido, esta dissertação apresenta a Bateria de Provas de Raciocínio – eletrônica (BPRe), instrumento que avalia as capacidades cognitivas em versão adaptativa computadorizada, que tem como referência a Bateria de Provas de Raciocínio (BPR-5), aplicada no formato lápis e papel. Sendo assim, esta pesquisa teve como objetivos: 1) buscar evidências de validade para a BPRe; 2) verificar a equivalência do funcionamento dos itens em formato lápis e papel (BPR-5) e computadorizado (BPRe), bem como analisar a estrutura interna da BPRe. Para isso, foram realizados dois estudos. O primeiro teve sua amostra composta por 52 universitários com idades entre 19 e 55 anos, de ambos os sexos. Os instrumentos utilizados foram o questionário sociodemográfico, Bateria de Provas de Raciocínio (BPR-5) e Bateria de Provas de Raciocínio (BPRe). Para análise de dados, foram realizadas estatísticas descritivas com a finalidade de caracterizar a amostra e análise de correlação de Pearson com os escores da BPR-5 e da BPRe em busca de evidência de validade convergente. Os resultados indicaram correlações de magnitude pequenas à grandes, as quais podem ter influência do tamanho amostral, além de uma redução no tempo dispensado na aplicação do instrumento. O segundo estudo contou com 13232 participantes, de ambos os sexos, com idades variando entre 11 e 23 anos, dos níveis de ensino fundamental, médio e superior. Parte da amostra respondeu a BPR-5 (n = 12545), enquanto outra parte respondeu a BPRe (n = 687). Foi utilizado o método de regressão logística para a detecção de itens com DIF e análise fatorial exploratória para identificar o número de fatores subjacentes aos itens da BPRe. Os resultados demonstraram a presença de apenas um item com funcionamento diferencial na BPRe. A sua estrutura interna demonstrou estabilidade, sendo a maior parte dos itens com carga fatorial acima de 0,30. Portanto, conclui-se de forma preliminar que a BPRe possui boas perspectivas de uso na avaliação das capacidades cognitivas, apresentando pequenas diferenças entre os formatos em lápis e papel e computador.

Palavras-chave: teoria de resposta ao item, testagem adaptativa computadorizada, psicometria.

Abstract

Padilha, Y. S. (2021). *Validity evidences for the Battery of Reasoning Tests – eletronic.* Master's Thesis, Post-Graduate Studies in Phychology, University San Francisco, Campinas, São Paulo.

The use of psychological tests is adopted in different contexts, contributing to the professional practice of the psychologist. Although the conventional test format prevails in psychological assessment processes, technological advances have enabled the development of Computerized Adaptive Testing (CAT) through the Item Response Theory (IRT), which makes testing shorter while maintaining its efficiency. The use of CATs is already widely adopted internationally, but they are still recent in Brazil. In this sense, this dissertation presents the *Battery of Reasoning Tests* - electronic (BPRe), an instrument that assesses cognitive abilities in a computerized adaptive version, which has as reference the *Battery of Reasoning Tests* (BPR-5), represented in pencil and paper. Therefore, this research aimed to: 1) seek evidence of validity for the BPRe; 2) verify the equivalence of the functioning of items in pencil and paper (BPR-5) and computer (BPRe) format, as well as analyze the internal structure of BPRe. For this, two studies were carried out. The first had its sample composed of 52 university students aged between 19 and 55 years, of both sexes. The instruments used were the sociodemographic questionnaire, the *Battery of Reasoning Tests* (BPR-5) and the *Battery of Reasoning Tests – eletronic* (BPRe). For data analysis, descriptive statistics were performed in order to characterize the sample and Pearson's correlation analysis with the BPR-5 and BPRe scores in search of evidence of convergent validity. The results indicated small to large magnitude correlations, which may have an influence on the sample size, in addition to a reduction in the time spent in applying the instrument. The second study included 13232 participants, of both sexes, aged between 11 and 23 years, from elementary, secondary and higher education levels. Part of the sample responded to BPR-5 (n = 12545), while another part responded to BPRe (n = 687). The logistic regression method was used to detect items with DIF and exploratory factor analysis to identify the number of factors underlying the BPRe items. The results showed the presence of only one item with differential functioning in BPRe. Its internal structure showed stability, with most items with a factor loading above 0.30. Therefore, it is preliminary concluded that the BPRe has good perspectives for use in the assessment of cognitive abilities, showing small differences between the formats in pencil and paper and computer

Keywords: item response theory, computerized adaptive testing, psychometry.

Apoio Financeiro

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior- Brasil (CAPES) – Código de Financiamento 001.

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) no âmbito do convênio FAPESP/CAPES (nº processo de bolsa no país 2019/00022-0).

Sumário

Lista de figuras

Introdução

Estudo 1

Estudo 2

Lista de tabelas

Introdução

Estudo 1

Estudo 2

Lista de abreviaturas e siglas

AERA	*American Education Research Association*
APA	*American Psychology Association*
BI	Banco de Itens
BPR-5	Bateria de Provas de Raciocínio
BPRe	Bateria de Provas de Raciocínio – eletrônica
BPRi	Bateria de Provas de Raciocínio infantil
CAT	*Computerized Adaptive Testing*
CCI	Curva Característica do Item
CHC	Cattell-Horn-Carroll
DIF	*Differential Item Functioning*
EaD	Educação â Distância
ENADE	Exame Nacional de Desempenho dos Estudantes
ENEM	Exame Nacional do Ensino Médio
EPN	Escore Padrão Normalizado
Gc	Inteligência Cristalizada
Gf	Raciocínio fluido
Gkn	Conhecimento de domínios específicos
Gq	Conhecimento Quantitativo
Gv	Processamento Visual
Gwm	Memória de trabalho de curto prazo
ID	Índice de Dificuldade
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LabAPE	Laboratório de Avaliação Psicológica e Educacional
JLME	*Joint Maximum Likelihood Estimation*
MOODLE	*Modular Object-Oriented Dynamic Learning Environment*
NCME	*National Council on Measurement in Education*
RA	Raciocínio Abstrato
RE	Raciocínio Espacial
RM	Raciocínio Mecânico
RN	Raciocínio Numérico
RV	Raciocínio Verbal
TAI-PI	Teste Adaptativo Informatizado para Proficiência em Inglês
TCLE	Termo de Consentimento Livre e Esclarecido
TCT	Teoria Clássica dos Testes
TRI	Teoria de Resposta ao Item
TU	Testagem Universal
WLSMV	*Weighted Least Squares Estimator*

Introdução

A avaliação de aptidões por meio do uso de testes de inteligência é hoje adotada em diferentes contextos. Ainda que o formato lápis e papel tenha destaque na área de avaliação psicológica, a evolução da tecnologia possibilitou o desenvolvimento de instrumentos que tornam a testagem objetiva e breve. A implementação da testagem adaptativa computadorizada (CAT – do inglês *Computerized Adaptive Testing*) foi consideravelmente ampliada nos últimos anos e a tendência de avaliação por meio do uso de testes computadorizados já é adotada em âmbito internacional. No entanto, o uso de CAT ainda é recente no Brasil (Nunes, Spenassato, Oliveira, Bornia, & Primi, 2015). Sendo assim, esse estudo tem como objetivo buscar evidências de validade para a Bateria de Provas de Raciocínio – eletrônica (BPRe) (Primi, 2013), instrumento que avalia habilidades cognitivas em versão adaptativa computadorizada. Ademais, é verificada a equivalência dos itens do instrumento nos diferentes formatos (lápis e papel e computador) e analisada a sua estrutura interna.

Testes psicológicos e inteligência

Os testes psicológicos são ferramentas que visam oferecer informações válidas e precisas para tomada de decisões significativas e contam hoje com sofisticados métodos, tornando-os reconhecidamente fontes de dados minuciosos e consistentes, o que contribui para a prática profissional do psicólogo (Primi, Almeida, Nakano, & Campos, 2018). Desde os primórdios da psicologia científica, o uso de instrumentos de medida esteve presente na mensuração de aptidões para a investigação de diferenças individuais. Os estudos voltados para a estrutura e definição das capacidades intelectuais possuem uma história que evolui gradativamente (Urbina, 2007).

O acúmulo e integração de teorias concebidas ao longo de mais de um século faz da inteligência um dos temas mais pesquisados na ciência psicológica – especialmente os estudos que fazem uso do modelo psicométrico (Caemmerer, Keith, & Reynolds, 2020; Campos & Nakano, 2012; Primi, McGrew, Schneider, Nakano, & Dias, 2017). De modo geral, seu exercício envolve a capacidade de raciocínio, planejamento, resolução de problemas, pensamento abstrato, compreensão de ideias complexas e aprendizagem por meio de experiências (Gottfredson, 1997).

A partir da psicometria, a inteligência é estruturada como um construto multidimensional e hierárquico, tendo como um dos modelos de referência a teoria Cattell-Horn-Carroll de Inteligência – CHC (Carroll, 1993; 1997; 2005). Considerado como o mais completo disponível e, possivelmente o mais empregado atualmente, o desenvolvimento do modelo CHC tornou-se um marco principal nos estudos sobre inteligência e sua taxonomia é adotada como padrão entre pesquisadores dedicados ao tema (McGrew, 2009; Primi et al., 2017).

Dentre os três estratos estabelecidos neste modelo, o primeiro deles corresponde ao fator *g* de Spearman, definido como a associação geral de todas as capacidades cognitivas (Schneider & McGrew, 2018). Essas capacidades são distribuídas em uma segunda camada e o aprimoramento do modelo teórico demonstra hoje a existência de dezessete fatores amplos (Primi, Correia, & Almeida, 2018). Uma terceira camada subdivide os fatores amplos em, aproximadamente, setenta fatores específicos e, nesse sentido, o caminho entre o nível mais alto e o nível mais baixo revela o desenvolvimento de especificidades dentre as capacidades cognitivas, avaliadas por meio de tarefas que compõem os testes psicológicos (Primi & Nakano, 2015). A Figura 1 apresenta a representação simplificada de sua estrutura hierárquica.

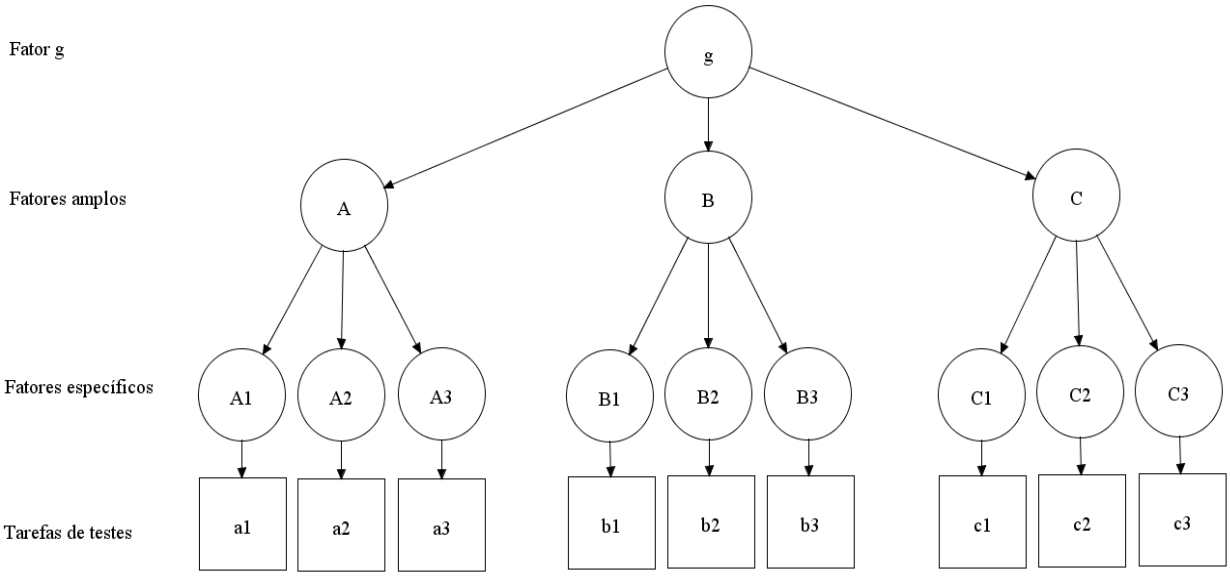

Figura 1. Estrutura hierárquica do modelo CHC.
Adaptado de "*The Cattell-Horn-Carroll Theory of Cognitive Abilities*", de W. J. Schneider e K. S. McGrew, 2018.

A definição da estrutura e organização da inteligência nesse sistema taxonômico abrangente parte do uso de análise fatorial para classificação das habilidades cognitivas primárias. A análise fatorial avalia, por meio de correlações, a causa latente entre os comportamentos manifestos em diferentes testes cognitivos, identificando e agrupando as dimensões da inteligência. Esse processo possibilita a compreensão de capacidades comuns envolvidas na execução das tarefas de cada teste como reflexo dos fatores cognitivos, inferidos por meio de correlações (Primi et al., 2017). Assim, nos estudos sobre inteligência, compreende-se que se diferentes instrumentos requerem a mesma habilidade cognitiva, ou seja, a causa comum, desempenhos correlacionados serão encontrados nos resultados de um mesmo indivíduo (Primi & Nakano, 2015).

Atualmente, é encontrado um número expressivo de instrumentos psicológicos disponíveis para a avaliação das capacidades cognitivas. Um instrumento amplamente utilizado no que se refere à avaliação das capacidades cognitivas é a Bateria de Provas de Raciocínio – BPR-5 (Almeida & Primi, 2000). Seu uso é apontado em estudos de revisão científica como um

dos instrumentos nacionais utilizados com mais frequência quando se trata de avaliação de habilidades cognitivas (Alves, Rosa, da Silva, & Sardinha, 2016; Campos & Nakano, 2012; Lima, Cunha, & Suehiro, 2019; Souza, 2018; Suehiro, Benfica, & Cardim, 2015).

Em uma busca realizada no mês de dezembro de 2020 sobre o instrumento, foram consultadas cinco bases de dados eletrônicas (*LILACS,* PePSIC, *PsycINFO*, *Redalyc,* SciElo), adotando como descritores os termos "bateria de provas de raciocínio" e 'BPR-5", sem delimitação que especificasse os anos de publicação. Com a finalidade de refinar a busca, foi adotado como critério de exclusão a publicação que não apresentasse ao menos um de seus subtestes como parte do delineamento metodológico do estudo. Com isso, foram encontradas 72 publicações que remetessem ao uso da BPR-5 em estudo empírico entre os anos de 2000 e 2019 em formas de artigos científicos, dissertações e teses. Ademais, a literatura apresenta 9 artigos que fazem referência ao instrumento, estimando-o como um dos testes mais utilizados na avaliação de aptidões, atribuindo-lhe uma posição de destaque em âmbito nacional.

Além disso, foram identificados os diferentes contextos nos quais a BPR-5 tem sido empregada, demonstrando a aplicabilidade do teste em diferentes áreas. A busca apontou o seu uso em contextos de rendimento e aprendizagem escolar (Almeida, Guisande, Primi, & Lemos, 2008; Almeida, Lemos, Guisande, & Primi, 2008; Brito et al., 2000; Elói & Candelas, 2010; Lemos, Almeida, Guisande, & Primi, 2008; Lemos et al., 2010; Oliveira & Soares, 2011; Pocinho, 2010; Ponczek & Pinto, 2016; Primi & Almeida, 2010; Primi, Couto, Almeida, Guisande, & Miguel, 2012; Santos et al., 2010), organizacional (Baumgartl & Nascimento, 2004; Baumgartl & Primi, 2006; Cobêro, Primi, & Muniz, 2006; Filizatti, 2004; Madeira, Wainer, Verdin, Alchieri, & Diehl, 2002; Souza, Primi, & Miguel, 2007), clínico (Monteiro & Andrade, 2005; Oliveira & Barbosa, 2015; Passos & Barbosa, 2011) e em processos de orientação profissional (Ambiel, 2010; Godoy, Noronha, Ambiel, & Nunes, 2008; Mansão, 2005; Nunes &

Noronha, 2009; Primi et al., 2002). Com isso, são apresentados indícios positivos quanto ao uso da BPR-5 como um instrumento pertinente para a avaliação de capacidades cognitivas em contextos variados.

Características da Bateria de Provas de Raciocínio (BPR-5)

A BPR-5, aplicada no formato lápis e papel, é organizada em duas formas (A e B). A Forma A é destinada a sujeitos com escolaridade do 7° ao 9° ano do Ensino Fundamental, enquanto a Forma B aplica-se a sujeitos com escolaridade do 1° ao 3° ano do Ensino Médio. A bateria é constituída por cinco subtestes, a saber: Prova de Raciocínio Verbal (RV), Prova de Raciocínio Espacial (RE), Prova de Raciocínio Numérico (RN), Prova de Raciocínio Mecânico (RM) e Prova de Raciocínio Abstrato (RA). A execução das provas requer diferentes tarefas de raciocínio com base no modelo CHC de inteligência, apresentadas na Tabela 1.

Tabela 1. *Relação entre subtestes da BPR-5 e fatores amplos do modelo CHC.*

Subteste	Fatores amplos	Definição
Prova de Raciocínio Verbal (RV)	Inteligência Cristalizada (Gc)	Capacidade ligada à amplitude e profundidade de conhecimentos gerais valorizados pela cultura.
Prova de Raciocínio Espacial (RE)	Processamento Visual (Gv)	Capacidade de gerar, armazenar, evocar e modificar imagens visuais
Prova de Raciocínio Numérico (RN)	Conhecimento Quantitativo (Gq)	Capacidade ligada à amplitude e profundidade de conhecimentos sobre a linguagem matemática, quantitativa e numérica armazenados de forma declarativa e procedural
	Processamento Visual (Gv)	Capacidade de gerar, armazenar, evocar e modificar imagens visuais
	Memória de trabalho de curto prazo (Gwm)	Capacidade de apreender, manter e processar informações na consciência em uma situação imediata
Prova de Raciocínio Mecânico (RM)	Conhecimento de domínios específicos (Gkn)	Capacidade ligada à amplitude, profundidade e domínio de conhecimentos especializados em determinado assunto ou disciplina
	Processamento Visual (Gv)	Capacidade de gerar, armazenar, evocar e modificar imagens visuais
Prova de Raciocínio Abstrato (RA)	Raciocínio fluido (Gf)	Capacidade ligada ao uso deliberado e flexível de operações mentais para a resolução de novos problemas que não podem ser executados de forma automática. Engloba fatores específicos de raciocínio indutivo e dedutivo

Nota: Tabela elaborada com base em Primi et al. (2018).

Conforme o modelo CHC, a Prova de Raciocínio Verbal (RV) é associada à inteligência cristalizada (Gc), assim como a Prova de Raciocínio Espacial (RE) é relacionada com a capacidade de processamento visual (Gv). A Prova de Raciocínio Numérico (RN) está ligada com a habilidade quantitativa (Gq), com o processamento visual e com a memória de trabalho (Gwm), enquanto a Prova de Raciocínio Mecânico (RM) se relaciona com os conhecimentos práticos mecânicos (Gkn) e com o processamento visual. Por fim, a Prova de Raciocínio Abstrato (RA) é associada principalmente ao raciocínio fluido (Gf) (Almeida & Primi, 2000; Primi, Silva, Santana, Muniz, & Almeida, 2013).

Estudos de evidências de validade e estimativas de fidedignidade da BPR-5 demonstram a qualidade psicométrica do instrumento, com evidências baseadas no conteúdo (Campos, 2005; Couto, 2007), na estrutura interna (da Silva-Junior, Nascimento, & Roazzi, 2019; Primi, Silva, Santana, Muniz, & Almeida, 2013), e na relação com variáveis externas (Almeida, Guisande, Primi, & Lemos, 2008; Almeida & Primi, 2004; Guise & Wechsler, 2018; Primi & Almeida, 2000; Primi, Bueno, & Muniz, 2006; Primi, Ferrão, & Almeida, 2010; Souza, 2018). Adicionalmente, os coeficientes de precisão Alfa de Cronbach das provas de raciocínio, que indicam a estabilidade dos escores em razão dos erros de medida, variam de 0,70 (RM) a 0,91 (RN), com média de 0,83. A precisão dos escores globais é situada entre 0,93 e 0,96, demonstrando alta consistência interna, com estimativa de erro de medida nos escores gerais de 3,3 pontos, conforme o Escore Padrão Normalizado – EPN (Primi & Almeida, 2000; Primi et al., 2018).

Ainda assim, é indispensável a realização de novos estudos de validade, que caracterizam-se como um processo contínuo e cumulativo. O estabelecimento de diretrizes internacionais, como o *Standards for Educational and Psychological Testing* (American Education Research Association [AERA], American Psychology Association [APA], & National Council on

Measurement in Education [NCME], 2014), fornecem as bases do conhecimento para a realização de estudos sobre a qualidade psicométrica de testes psicológicos, a qual está condicionada à demonstração de evidências de validade e estimativas de precisão. Dessa forma, são estabelecidos os critérios e requisitos mínimos não somente para a construção, validação e padronização de instrumentos de medida confiáveis, mas também para a avaliação dos testes psicológicos, bem como orientações para as práticas de testagem e os efeitos dos seus usos.

A fim de proceder conforme as diretrizes nacionais (Nunes & Primi, 2010) e internacionais (AERA, APA, & NCME, 2014), novos estudos acerca da BPR-5 têm sido realizados, de forma a aprimorar suas propriedades psicométricas. O desenvolvimento de estudos acerca da BPR-5 quanto aos diferentes públicos visa possibilitar seu uso na avaliação de indivíduos em diferentes faixas etárias. Além das formas A e B já mencionadas, também tem sido estudada a Bateria de Provas de Raciocínio infantil (BPRi), com cadernos de provas direcionados para o uso em crianças do 2º ao 6º ano do ensino fundamental, com faixa etária entre 8 e 12 anos. A atualização das normas contidas no manual da BPR-5 da forma B também visa contemplar a avaliação de universitários e adultos, com tabelas de conversão considerando diferentes níveis de escolaridade e idade (Primi et al., 2018).

Contudo, a abrangência de diferentes faixas etárias e formas do instrumento implica uma questão métrica quanto à comparabilidade dos resultados. Isto porque devem ser considerados os índices de dificuldade dos itens no cálculo de escores, quando comparados os sujeitos quanto ao seu nível de habilidade. Ainda que os escores da BPR-5 apresentem escore padrão normalizado e percentil em suas normas, isso não permite que sejam comparados os sujeitos em outras condições ou em diferentes níveis desenvolvimentais, por exemplo. Assim como a construção de uma escala comum entre as formas da BPR-5, a testagem de equivalência entre os diferentes

formatos, visa permitir esse tipo de comparação. Para isso, contamos com o emprego metodológico da Teoria de Resposta ao Item – TRI (Ayala, 2009; Primi et al., 2018).

Teoria de Resposta ao Item (TRI)

O avanço teórico da psicometria, em consonância com o avanço tecnológico, permitiu que limitações, como a comparabilidade dos resultados obtidos em testes de desempenho, fossem suplantadas. O uso de computadores e *softwares* gerou impactos nos métodos psicométricos e colaborou expressivamente para o incremento de análises. Por muito tempo acreditou-se que um demasiado número de itens estaria atribuindo mais qualidade aos testes psicológicos, por melhor representarem um traço, o que também contribuía para os cálculos estatísticos. Todavia, as modificações do último século permitiram a realização de cálculos matemáticos mais complexos de modo útil e prático e, por conseguinte, o desenvolvimento de instrumentos de medida menos extensos e mais precisos, ao se usarem conjuntos de itens menores e mais discriminativos (Embretson, 1996; Pasquali & Primi, 2003; Streiner, 2010).

Embora a Teoria Clássica dos Testes (TCT) tenha sido essencial para o desenvolvimento de testes psicológicos e ainda seja amplamente utilizada, são assinaladas diversas limitações quanto ao seu uso frente à avaliação de aptidões. Dentre elas, estão a relação de dependência da dificuldade dos itens da amostra, das habilidades dos sujeitos e da precisão da medida de valor único (Sartes & Souza-Formigoni, 2013).

A elaboração da TRI, fomentada pelo avanço tecnológico, possibilitou a superação de limitações apesentadas pela TCT. Oriunda dos trabalhos de Richardson (1936), Lawley (1943), Tucker (1946) e Lord (1952,1953), a TRI teve sua difusão ampliada com os estudos de Rasch (1960) e com a popularização do uso de computadores (Couto & Primi, 2011). Com sua aplicação, cada item do teste é avaliado separadamente, e não apenas o escore total obtido, como ocorre na TCT, o que possibilita avaliar o desempenho dos sujeitos em cada item

especificamente. Assim, se na TCT dois indivíduos com escore igual a 5 seriam considerados como tendo o mesmo nível de habilidade, na TRI o cálculo dos escores é mais complexo considerando as propriedades dos itens como capacidade de discriminação e/ou a sua probabilidade de acerto ao acaso e dificuldade no caso de formas de teste diferentes, o que poderia resultar escores finais distintos (Nakano et al., 2015).

Diferentes modelos matemáticos de TRI foram propostos na literatura, oferecendo informações psicométricas acerca dos itens (Couto & Primi, 2011). Dentre eles, estão os que avaliam apenas um traço latente, definidos como modelos unidimensionais, assim como outros mais complexos que consideram mais de um construto, denominados modelos multidimensionais. A verificação da dimensionalidade é baseada na análise da estrutura fatorial de matrizes de correlação entre os itens, na qual cada conjunto inter-relacionado demonstra a existência de um traço latente comum (Ayala, 2009; Hambleton & Swaminatham, 1985; Primi, 2012). A presença de um traço latente dominante possibilita o ajuste na estimação dos parâmetros e redução de erros padrão (Andrade, Tavares, & Valle, 2000) em um modelo unidimensional.

Um estudo realizado por Primi et al. (2013) testou a unidimensionalidade dos itens da BPR-5 por meio de ajuste do modelo bi-fatorial, visto que a verificação deste pressuposto é de suma importância no emprego da TRI. Foi utilizada uma base de dados com um total de 2.763 estudantes, oriundos de escolas públicas e privadas de três estados brasileiros, divididos entre BPR-5 Forma A (n = 603), BPR-5 Forma B (n = 1.748) e BPRi (n = 412). Os resultados obtidos em análise fatorial e análise do modelo bi-fator demonstram a presença de um fator dominante no instrumento que, nesse caso, pode ser interpretado como a inteligência fluida (Gf), o que explica a maior parte de sua variância comum. Este é apontado como um fator subjacente aos fatores específicos, associados aos diferentes conteúdos presentes nas provas de raciocínio (verbal, abstrato, numérico, espacial e mecânico). A unidimensionalidade apresentada pela BPR-5 por

meio de um fator geral (Gf), também encontrada em seus subtestes, contribui para o estabelecimento de uma base de aplicabilidade do instrumento, com formulários equalizados quanto aos seus itens que, nesse caso, corresponde a 54%. Dessa forma, este trabalho possui enfoque no modelo unidimensional.

A TRI para itens dicotômicos considera modelos logísticos de um, dois e três parâmetros, a saber: (a) a dificuldade, (b) o poder discriminativo e (c) a probabilidade de acerto ao acaso. Com isso, ela permite a formulação adequada de itens para o cálculo de escores latentes, considerando a habilidade do sujeito – *theta* (θ), a dificuldade do item apresentado e a probabilidade (P) de acerto (De Alaya, 2008). Tal modelo pode ser descrito por meio da Curva Característica do Item – CCI (Figura 2), que representa a relação entre a habilidade do sujeito (eixo horizontal) e a probabilidade de um indivíduo com determinada habilidade, atribuir a resposta correta ao item (eixo vertical). A função que descreve essa relação é expressa por P_i (θ), onde a probabilidade de acerto de um item i é dada em função do valor de (θ) (Ayala, 2009; Couto & Primi, 2011; Nakano et al., 2015; Pasquali & Primi, 2003).

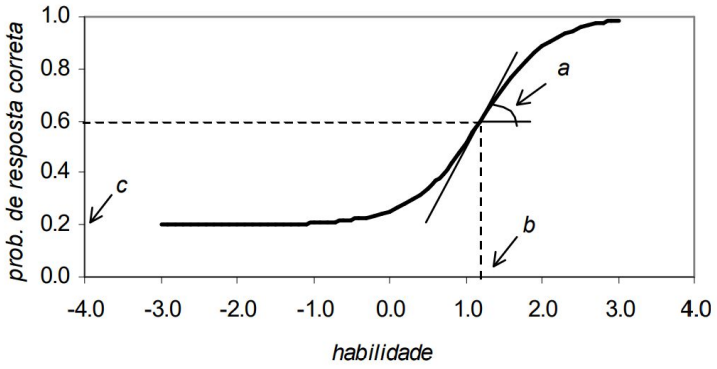

Figura 2. Curva Característica do Item (CCI).
Adaptado de *"Teoria de Resposta ao Item - Conceitos e Aplicações",* de D. F. Andrade e H. R. Tavares e R. C. Valle (2000).

A quantificação do nível de aptidão (θ) dos sujeitos em um traço latente e sua relação com cada item permite avaliar a probabilidade (P) de obtenção de pontos mais altos ou mais baixos, conforme os itens de um teste que avaliem aquele traço. Na TRI, o nível de habilidade do sujeito

é estimado conforme o valor de *theta,* que varia habitualmente entre -4 e +4, embora possam ser encontrados outros valores, sendo os positivos os mais intensos em termos de aptidão. Feita a mensuração de *theta*, é possível prever o desempenho de um sujeito em itens específicos do teste, com o cálculo da probabilidade de acerto.

Além disso, a TRI também permite a realização de análise do Funcionamento Diferencial dos Itens (DIF, do inglês *Differential Item Functioning*). A análise de DIF pode ser utilizada para a verificação de possíveis vieses quando um item se comporta de maneira diferente entre grupos de indivíduos com o mesmo nível de habilidade. Dessa forma, podem ser observadas tendências com relação a outras variáveis, como sexo, escolaridade, diferenças étnicas, entre outras. Na prática, um item que apresente DIF para sexo, considera as diferenças na probabilidade de respostas dadas por homens e por mulheres que possuem o mesmo nível de habilidade, por exemplo (Andrade, Laros, & Gouveia, 2010; Sisto, 2006; Pires, Filgueiras, Ribas, & Santana, 2013; Primi, Carvalho, Miguel, & Silva, 2010). Assim a análise DIF permite avaliar se um determinado item funciona de forma diferente nos grupos considerados em termos de nível que mede o construto e como se relaciona ao construto.

Ao verificar as diferenças de traços latentes entre grupos em testes de desempenho, por exemplo, estamos testando a invariância dos itens, o que visa garantir que a medida seja equivalente para ambos. Dessa forma, quando há invariância, ou seja, o item se comporta da mesma forma em grupos distintos, pode-se concluir que não ocorre o DIF (Valentini, Franco, & Iglesias, 2017). Ressalta-se a importância desse processo de controle, visto que buscam evitar injustiças em processos avaliativos/seletivos, proporcionando as mesmas oportunidades para os indivíduos na apresentação de suas habilidades. Caso contrário, a existência de DIF indica que sujeitos com o mesmo nível de traço latente possuem probabilidades diferentes de acerto. Caso

isso não seja controlado essas diferenças podem adicionar erro sistemático na medida (Andriola, 2006; Jaloto, 2021).

Este é um processo que pode colaborar para melhoria de parâmetros psicométricos de um teste, em especial para o procedimento de normatização (Nunes & Primi, 2010). A ausência de DIF demonstra equivalência da medida, atestando que o teste mede com o mesmo nível os diferentes grupos, não necessitando em princípio de tabelas separadas (Everson & Osterlind, 2009; Linacre, 2014). Para este estudo, o conceito de grupo é um pouco diferente do usual, pois buscamos verificar se os diferentes grupos em razão do meio e sequenciamento dos itens os quais diferem em razão dos diferentes formatos do teste (lápis e papel e CAT). A nossa questão central não tem relação com viés favorecendo um determinado grupo, mas sim uma questão de validade, isso é, se os itens mantêm suas propriedades quando aplicados em meios distintos (lápis e papel *versus* CAT).

Outra contribuição da TRI para a mensuração de traços é a Testagem Adaptativa Computadorizada (CAT), um formato de administração de testes que tem seu uso intensificado nas últimas décadas em âmbito internacional. A partir dos novos modelos psicométricos de TRI, os CATs surgem como uma possibilidade inovadora para as avaliações educacionais e psicológicas (Peres, 2019; Reppold & Gurgel, 2017).

Testagem Adaptativa Computadorizada (CAT)

Embora diferentes formas de testagem tenham sido elaboradas durante o desenvolvimento da psicometria, somente com o avanço tecnológico foi possível ampliar de forma considerável a implementação de CATs a partir da década de 1990. O uso de testes de nivelamento e avaliação do desempenho tem destaque na inserção da testagem adaptativa computadorizada nos contextos educacionais, de certificação, em testagem psicológica e na área da saúde (Nunes et al., 2015).

Em geral, os CATs são constituídos por cinco componentes, a saber: (1) um banco de itens com parâmetros estimados pela TRI, (2) método de início do teste, (3) método de seleção dos itens, (4) método de estimação de proficiência e (5) critério de encerramento do teste. Para maior controle, são aplicadas restrições quanto ao balanceamento do conteúdo e controle da exposição de itens, bem como do tempo de resposta. Esses componentes são desenvolvidos durante a realização de várias etapas, como a elaboração do banco de itens, a sua pré-testagem, a avaliação de sua dimensionalidade, entre outros, para que seja feita a sua implementação (Nunes et. al, 2015).

O funcionamento de um CAT está fundamentado na seleção de itens apropriados especificamente para cada sujeito pela predição do seu desempenho, com base em algoritmos computacionais pré-definidos, o que torna a testagem personalizada. A estimativa de proficiência dada em etapa inicial é provisória, com a apresentação de um ou mais itens baseados no conjunto de regras do sistema elaborado (Nunes et al., 2015). A etapa posterior consiste na apresentação de itens mais informativos para a proficiência, reestimando-a em um ciclo (Figura 3).

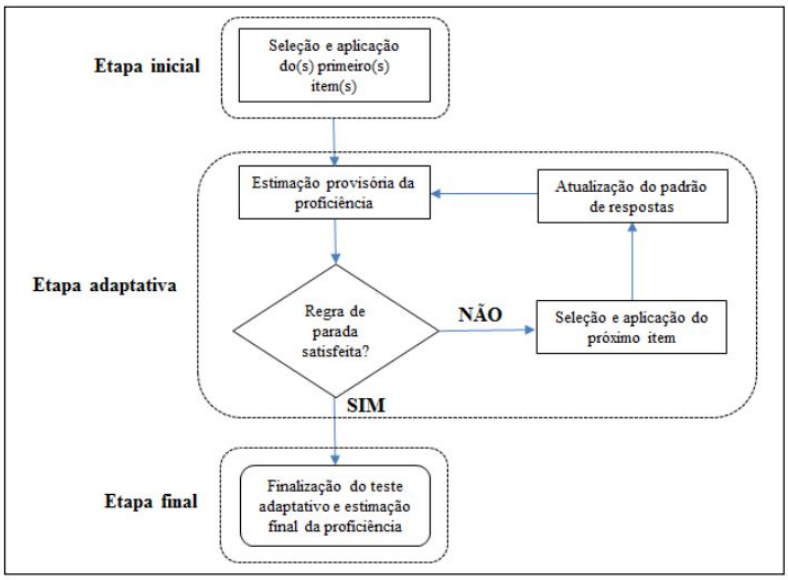

Figura 3. Algoritmo geral de um CAT.
Adaptado de "*Open-source CAT software: R packages and Concerto",* de D. Magis, & J. R. Barrada, 2014.

Quando um indivíduo acerta um item, o próximo será mais difícil, assim como quando o indivíduo erra um item, o próximo será mais fácil. Por fim, a etapa de resultado final da estimativa encerra a avaliação quando a regra de parada do sistema é satisfeita (Nunes et al., 2015). A Figura 4 ilustra a aplicação de um CAT em que a regra de parada pré-estabelecida é 15 itens, sendo os três primeiros itens de treino.

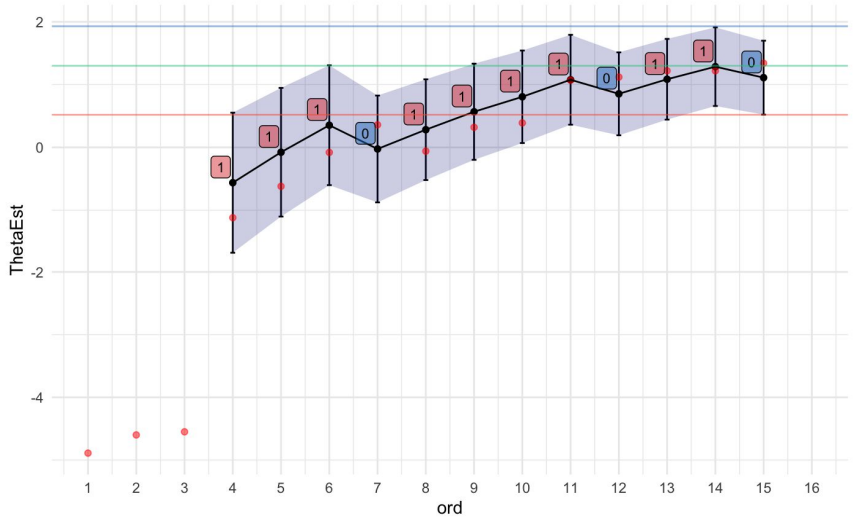

Figura 4. Padrão de respostas de um indivíduo em um CAT.

Os itens numerados como zero correspondem a uma resposta incorreta, enquanto os demarcados pelo número um correspondem a uma resposta correta de um indivíduo, cuja a habilidade se aproxima do nível de dificuldade dos itens, sendo reestimada a cada rodada. Dessa forma, sua habilidade é estimada com maior eficiência e precisão, comparada aos modelos tradicionais de aplicação (e. g., lápis e papel) em função da seleção de itens com maior poder discriminativo sobre medida para o sujeito que está respondendo (Kopec et al., 2008; Nunes & Primi, 2009; Nunes et al., 2015). Para além da melhoria de propriedades psicométricas de instrumentos, outro benefício da implementação de CATs está na brevidade de execução da testagem. Ao empregar este formato de teste, são reduzidos os números de itens necessários para estimar a habilidade dos sujeitos, causando menos fadiga, sem comprometer seu desempenho ou

a precisão do instrumento (Bjorner, Chang, Thissen, & Reeve, 2007; Nunes & Primi, 2009). Com isso, também é reduzida uma possível frustração do testando frente a itens consideravelmente acima ou abaixo do nível de suas habilidades (Urbina, 2007).

A disposição de itens de um CAT é dada a partir de um Banco de Itens (BI) que, por meio de controle de exposição, garante com segurança a distribuição de itens durante a aplicação do teste, diferentemente do que ocorre em instrumentos com formato lápis e papel, os quais têm todos os seus itens expostos (Wainer, 2000). O controle de exposição de itens, que pode ser feito por diferentes métodos, evita que os itens de um teste tornem-se conhecidos em função de uma frequente utilização (Georgiadou, Triantafillou, & Economides, 2007). Além disso, os dados do BI de um sistema CAT são dispostos em uma mesma métrica, o que permite a comparação entre indivíduos, mesmo que respondam a itens diferentes (Bjorner et al., 2007; Nakano, Primi, & Nunes, 2015; Nunes et al., 2015; Pasquali & Primi, 2003).

Adicionalmente, assim como em testes informatizados, a apresentação de itens em CATs é considerada como um diferencial no processo de testagem, com a possibilidade do uso de multimídias tanto na exibição de itens (e. g., áudios, vídeos, gráficos), quanto nas opções de resposta do testando (e. g., realce em textos, seleção de figuras, reordenação de séries, movimentação de objetos) (Huang, Lin, & Cheng, 2009; Nunes et al., 2015; Parshall, Harmes, Davey, & Pashley, 2010; Wainer, 2000). Outra vantagem em comum com os testes informatizados está no desenvolvimento de instrumentos de medida de Testagem Universal (TU), que primam pela inclusão. O emprego de tecnologia potencializa a acessibilidade, viabilizando a realização de testes com sujeitos que possuem deficiência por meio de recursos adicionais, garantindo equidade com relação às pessoas que não possuem deficiência (Almond, 2010; Ketterlin-Geller, 2005).

A administração de testes computadorizados colabora de forma considerável para o processo de testagem, uma vez que evitam erros de aplicação por meio de uniformização de procedimentos e correção automatizada, para além da redução de custos com material impresso. Conforme a programação de cada CAT, também é possível que o sistema forneça dados sobre o controle e avaliação do tempo de resposta aos itens por parte dos indivíduos. Apesar do potencial apresentado pela testagem adaptativa computadorizada para suplementar métodos tradicionais de testagem, seu uso ainda é considerado reduzido e em processo inicial no Brasil, quando comparado a outros países (Nunes et al., 2015; Peres, 2019).

Em uma metaanálise, Mead e Drasgow (1993) buscaram examinar a equivalência entre testes de habilidades cognitivas em lápis e papel e CAT, considerando a hipótese de que a informatização de um instrumento poderia mudar sua escala de pontuação ou mesmo que o construto psicológico poderia ser afetado pelo modo de administração. Para isso, observaram que a maior parte dos estudos usou métodos estatísticos de correlação entre os escores obtidos nas diferentes formas dos instrumentos e concluíram que, nas 159 correlações analisadas, os coeficientes variaram entre 0,72 e 0,97, no que se refere à comparabilidade dos testes.

Do mesmo modo, Peres (2019) analisou a produção científica a respeito de CATs publicados no Brasil. Para isso, realizou uma busca em três bases de dados nacionais (BVS- PSI, SciElo e BDTD), utilizando os termos “testagem adaptativa” e “teste adaptativo”, o que resultou em 8 trabalhos, dentre artigos, dissertações e teses. Utilizando o mecanismo de busca do *Google* com os mesmos termos, foram encontradas 11 dissertações de mestrado e 4 teses de doutorado relacionadas ao desenvolvimento e implementação de CATs. Tais resultados, ainda que com limitações quanto às poucas bases de dados e publicações encontradas somente em revistas brasileiras, permitem concluir que a pesquisa em CAT ainda é incipiente no Brasil.

Ao revisar pesquisas relacionadas a CATs e TRI, Spenassato, Bornia e Tezza (2015) buscaram artigos em 10 bases de dados, mapeando perfil das produções científicas o tema até o ano de 2014. Foram encontrados 935 artigos que, refinados, resultaram em 182 para uma análise mais completa. Dentre os resultados, foi possível constatar que estudos sobre CATs cresceram consideravelmente nos últimos anos e têm sido desenvolvidos em diversos âmbitos, como em saúde, avaliação da personalidade, atitudes e comportamento (42%), seguido por uso de dados simulados (25%), testes para avaliação educacional, aptidão e raciocínio (23%) e avaliação de satisfação (10%). Outros estudos com simulações com dados reais (chamadas de post-hoc) para testar um CAT foram realizados, passo importante para sua aplicação em contextos reais como forma de estudo piloto. Contudo, apenas 28% desses artigos chegaram a etapas de validação, implementação e aplicação do teste em situações reais, sendo a maioria dos estudos restritos às simulações e sugerem que trabalhos futuros apliquem testes adaptativos computadorizados a situações reais, investiguem sua estrutura dimensional, métodos de detecção de DIF em CATs, entre outras.

A implementação de CATs em contextos educacionais pode colaborar para o aumento da eficiência de processos de avaliação. Karay, Schauber, Stosch e Schüttpelz-Brauns (2015) examinaram as diferenças no desempenho de 226 estudantes do curso de medicina da Universidade de Bologna em um teste de progresso interdisciplinar em versão lápis e papel e CAT, comparando os seus respectivos escores. Os resultados demonstram equivalência entre as medidas, com a média de 72,56 acertos no teste em lápis e papel e de 75,18 acertos no teste em versão computadorizada, sugerindo que as duas formas de avaliação não apresentaram diferenças significativas em relação ao desempenho dos alunos. Ademais, foram encontradas diferenças no tempo de testagem, sendo o formato CAT mais breve em sua administração, fornecendo também um feedback imediato para o testando, o que foi apontado pelos alunos como um aspecto

motivador. A proposta de exames computadorizados colaborou ainda para a diminuição da carga de trabalho dos funcionários da universidade, visto que o número de alunos no ensino superior tem crescido e a demanda de exames também. Assim, a eficiência de CATs não se restringiu à administração, mas também à correção e divulgação dos resultados, para além da redução de custos com o material impresso da versão lápis e papel.

No Brasil, uma simulação feita por Spenassato, Trierweiller, Andrade e Bornia (2016) com dados do Exame Nacional do Ensino Médio (ENEM) na área de Matemática e suas Tecnologias do ano de 2012 resultou em uma redução dos 45 itens da versão lápis e papel do exame em 26,6%, definindo o número de 33 itens para a estimação da proficiência de forma eficaz. Esses resultados demonstram a vantagem do uso da tecnologia em procedimentos avaliativos, especialmente os de larga escala, considerando que o uso de testes adaptativos demonstra ser promissor no aumento da eficiência desses processos.

Ao verificar a possibilidade de redução de itens apresentados na prova de Ciências da Natureza e suas Tecnologias do ENEM do ano de 2015, Jaloto (2018) estimou a proficiência dos participantes por meio de dados divulgados pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira). Após a calibração dos itens em uma amostra de 10000 indivíduos, estratificados conforme o percentil da soma de acertos, foi possível observar que para participantes com notas mais baixas, foram necessários os 45 itens para a estimação de proficiência. Entretanto, para participantes com notas altas o número de itens apresentados foi de até 26, com média de 7,2. A simulação da testagem adaptativa computadorizada demonstra a potencialidade do seu uso, que diminui o número de itens sem perder a precisão na estimação do traço latente. Contudo, a implementação desse tipo de testagem no ENEM requer um número expressivo na composição do banco de itens para evitar uma exposição demasiada, ou seja, que um item seja apresentado excessivamente para diferentes sujeitos.

A substituição de uma versão de um exame de proficiência em inglês acadêmico em lápis e papel por uma versão CAT, denominado Teste Adaptativo Informatizado para Proficiência em Inglês (TAI-PI), foi testada por Silva (2015). Desenvolvido para avaliar a proficiência dos alunos de pós-graduação do Instituto de Ciências Matemáticas e da Computação da Universidade de São Paulo, o TAI-PI funciona a partir de um modelo de resposta gradual unidimensional de TRI. O resultado apresentado em sua aplicação demonstrou-se viável para mensurar a proficiência, além de tornar este processo ágil, informando imediatamente o resultado final do testando, classificando-o como aprovado ou reprovado. A qualidade do instrumento adaptativo desenvolvido expressa as futuras perspectivas de implementação do TAI-PI no programa de pós-graduação de forma gratuita, isentando os alunos de apresentarem certificados de abrangência internacional devido ao custo considerável para a realização desses exames.

A possibilidade de administração de um teste adaptativo computadorizado no Exame Nacional de Desempenho dos Estudantes (ENADE) foi estudada por Santana et al. (2017), ao relacionar os dados de 92 estudantes universitários em uma avaliação em formato lápis e papel e uma versão CAT do exame desenvolvido em ambiente *MOODLE (Modular Object-Oriented Dynamic Learning Environment)* com base no modelo de TRI de Rasch (os parâmetros foram estimados em estudo prévio). Comparando os níveis de dificuldade das provas em formato lápis e papel e em CAT, obtiveram a correlação de 0,82, demonstrando que ambas mantiveram equivalência quanto à complexidade dos itens. Contudo, a habilidade dos participantes teve correlação de 0,40, sugerindo a existência de uma diferença no desempenho dos participantes quando comparados os diferentes formatos. Segundo os autores, isso pode ser explicado pelas limitações apresentadas pela plataforma *MOODLE* no que se refere aos seus algoritmos, com critérios de parada e seleção restritos e também ao pequeno número de itens que compôs o banco (n = 26). Ainda assim, o formato CAT foi considerado rápido e eficiente, oferecendo

possibilidades de redução do número de itens do exame, agilidade no processo de avaliação em larga escala e redução de custos com a administração da prova, além de fornecer evidências de validade de construto para o teste.

Da mesma forma, outros estudos como os de Khoshsima e Toroujeni (2017), Oz e Ozturan (2018), Tseng, (2016) buscaram a comparabilidade do desempenho de alunos em testes de proficiência em línguas nos formatos lápis e papel e CAT. Os resultados encontrados corroboram com a literatura científica, a qual aponta resultados positivos quanto ao uso de testes computadorizados, com uma testagem mais breve e eficiente, evidenciando o seu potencial.

Um projeto que visa a implementação de CAT com possibilidade de aplicação na Educação à Distância (EaD) em uma instituição de ensino superior foi desenvolvido por Manseira e Misaghi (2015). A aplicação *on-line* dos exames e a correção automatizada permitem um feedback imediato do aluno, bem como a comparação do seu desempenho com testagens anteriores, para a verificação do seu progresso. Com isso, a participação ativa do aluno torna-o protagonista em sua formação, para além da eficiência da testagem com a melhora dos serviços educacionais prestados.

Dadas as suas condições, a BPR-5 oferece atualmente a apresentação de resultados em Escores Padrão Normalizados (EPN), comparando o desempenho de sujeitos com o mesmo nível de escolaridade (Almeida & Primi, 2000). Estudos para readequar os dados normativos das provas da BPR-5 com base nos procedimentos de linkagem/equalização da TRI têm sido realizados por Primi et al. (2018), o que permite a comparabilidade do desempenho entre diferentes formas da bateria.

Considerando a importância da apresentação de evidências de validade dos instrumentos de medida em avaliação psicológica e visando o desenvolvimento do processo de testagem, o presente estudo tem como objetivos: 1) buscar evidências de validade convergente baseada na

relação da BPRe com a BPR-5 versão lápis e papel; 2) verificar a equivalência do funcionamento dos itens, investigando a presença de DIF comparando-se se os itens funcionam da mesma maneira (dificuldade e discriminação) independentemente do meio de aplicação (lápis e papel ou computador). Por último pretende-se analisar a dimensionalidade de cada subteste da BPRe.

A hipótese é de que sejam encontrados escores equivalentes, sugerindo que os diferentes formatos do instrumento (lápis e papel e computador) estejam medindo o mesmo construto (Pasquali, 2017a; Primi, Muniz, & Nunes, 2009). Dessa forma, espera-se que os índices obtidos demonstrem correlações de magnitudes moderadas à altas, entre 0,50 e 0,90 (Cohen, 1988). Além disso, pretende-se alcançar uma redução do tempo de aplicação do instrumento por meio de avaliação adaptativa. Sugere-se também que os itens permanecerão funcionando de modo semelhante quando apresentados em computador, não havendo a presença de DIF. Por fim, espera-se que a estrutura interna da BPRe apresente os mesmos fatores da BPR-5.

Método

Estudo 1: Evidências de validade baseada na relação com variáveis externas

Participantes

Participaram deste estudo 52 alunos com idades entre 19 e 55 anos (M = 24,23; DP = 9,91) de ambos os sexos, dos quais 86.53% são do sexo feminino (n = 45). Os participantes foram recrutados em universidades privadas do interior do estado de São Paulo, sendo a amostragem por conveniência. Cada participante respondeu ao Questionário Sociodemográfico, um dos subtestes da BPR-5 definidos aleatoriamente e a BPRe. Da amostra, 16 casos foram excluídos por conterem respostas em branco em subtestes da BPR-5, inviabilizando a análise. Assim, 36 protocolos foram contabilizados para análise, das quais 88,88% são do sexo feminino (n = 32). A idade da amostra variou entre 19 e 55 anos (M = 25,59; DP = 9,45).

Instrumentos

Questionário sociodemográfico

Elaborado para a presente pesquisa, o Questionário Sociodemográfico (Anexo 1) foi composto por questões relativas à caracterização da amostra. As perguntas contemplam aspectos como sexo, idade, escolaridade e ocorrência de reprovação durante a vida escolar.

Bateria de Provas de Raciocínio – BPR-5 (Almeida & Primi, 2000)

O instrumento tem como finalidade a avaliação das habilidades cognitivas, especialmente a capacidade de raciocínio. É constituída por cinco provas de raciocínio, a saber: Raciocínio Abstrato (RA), Raciocínio Verbal (RV), Raciocínio Espacial (RE), Raciocínio Numérico (RN) e Raciocínio Mecânico (RM). Para o presente estudo, foram utilizadas as provas RA, RV, RE e RN, descritas abaixo. Em função do caráter opcional atribuído à prova de RM, esta não será utilizada nesse estudo, considerando que requer conhecimentos físico-mecânicos, comumente utilizada em processos avaliativos que buscam verificar especificamente essas habilidades.

a) RA, composta por 25 itens, nos quais é necessário que se descubra a relação existente entre os dois primeiros termos e aplica-la ao terceiro, para se identificar a quarta figura entre as cinco alternativas de resposta, com limite de tempo de 12 minutos. O índice de precisão da prova é de 0,82;

b) RV, composta por 25 itens nos quais a relação analógica existente entre um primeiro par de palavras deverá ser descoberta e aplicada de forma que se identifique a quarta palavra entre as cinco alternativas de resposta que mantenha a mesma relação com uma terceira apresentada, com limite de tempo de 10 minutos. O índice de precisão da prova é de 0,74;

c) RE, composta por 20 itens que implicam descobrir o movimento de cubos, por meio da análise das diferentes faces, escolhendo entre as alternativas de resposta a representação do cubo

que se seguiria se o movimento descoberto fosse aplicado ao último cubo da série, com limite de tempo de 18 minutos. O índice de precisão da prova é de 0,62;

d) RN, composta por 20 itens nos quais deve-se identificar a relação aritmética que rege as progressões nas séries e aplicá-la respondendo quais seriam os dois últimos números que completariam a série, com limite de tempo de 18 minutos. O índice de precisão da prova é de 0,75.

Bateria de Provas de Raciocínio – BPRe – sistema CAT (Primi, 2013).

O sistema BPRe CAT consiste em uma versão adaptativa computadorizada via *web* da BPR-5. O Banco de Itens (BI) do sistema é composto por 185 itens, calibrados por meio da TRI a partir do modelo de Rasch, pelo software Winsteps, com o procedimento de estimação JLME (*Joint Maximum Likelihood Estimation*). O teste é dividido entre quatro provas de raciocínio: Raciocínio Abstrato (RA, 52 itens), Raciocínio Verbal (RV, 57 itens), Raciocínio Espacial (RE, 28 itens) e Raciocínio Numérico (RN, 48 itens). A apresentação dos itens inclui recursos audiovisuais, sendo opção do testando a leitura e/ou escuta de enunciados e opções de resposta. Na etapa inicial da testagem, o sistema apresenta um item com dificuldade baixa. Conforme a resposta é emitida, o item é automaticamente corrigido e obtém-se a estimativa de *theta* a partir dessa informação. Em seguida, o *software* calcula o erro de medida. Com um único item, o erro de medida será muito alto, acima do que seria tolerável, portanto, escolhe-se um novo item cuja informação seja máxima para o *theta* atual e repete-se o processo. A cada ciclo, à medida que se aumenta o número de itens aplicados, o erro vai diminuindo até que atinja um valor aceitável (por exemplo, precisão de 0,80), momento em que se conclui a avaliação. Este processo é repetido para as provas RA, RV, RE e RN, o que permite obter um escore para cada umas delas, assim como um escore geral.

Procedimentos

Inicialmente foram contatadas instituições de ensino superior que disponibilizassem seus laboratórios de informática para a coleta de dados e, com a devida autorização, o projeto foi submetido ao Comitê de Ética em Pesquisa da Universidade São Francisco (CAAE 20181619.5.0000.5514). Mediante a aprovação, foram definidos os dias e horários de coleta juntamente com as instituições, de modo a reservar os laboratórios de informática para a realização das tarefas. Cabe ressaltar que a aplicação da BPRe foi feita via *web*, o que demandou que os computadores possuíssem acesso à internet. A participação no estudo demandou a assinatura do Termo de Consentimento Livre e Esclarecido – TCLE (Anexo 2) por parte dos sujeitos.

A coleta de dados configurou a amostra em quatro grupos, divididos em função da prova a ser realizada. Cada participante executou uma das provas de raciocínio da BPR-5 (RA, RV, RN ou RE), distribuídos aleatoriamente, seguida pela BPRe. As aplicações foram realizadas de modo coletivo, com número de participantes definido conforme a capacidade dos laboratórios. Cada grupo realizou as duas etapas em torno de 50 minutos.

Análise de dados

Os dados coletados foram analisados no *software* R (R Core Team, 2020). Inicialmente, foram realizadas estatísticas descritivas para caracterização da amostra. Em seguida, foi realizada a análise de correlação de *Pearson* os escores da BPR-5 e da BPRe, em busca de evidência de validade convergente.

Resultados e Discussão

Para atender a finalidade deste estudo, o qual buscou evidências de validade baseada na relação com variáveis externas para a BPRe, os escores de suas provas deste instrumento foram

relacionados com os escores das provas da BPR-5, por meio da correlação de Pearson. Os resultados são apresentados na Tabela 1.

Tabela 1.
Coeficientes de correlação entre provas da BPR-5 e BPRe

	N	*r*	*p*
BPR-5 – RA - BPRe – RA	**14**	**0,28**	**0.324**
BPR-5 – RA - BPRe – RN	14	0,50	0.066
BPR-5 – RA - BPRe – RV	14	0,00	0.992
BPR-5 – RA - BPRe – RE	10	-0,23	0.512
BPR-5 – RN - BPRe – RN	**5**	**0,77**	**0.123**
BPR-5 – RV - BPRe – RV	**7**	**0,94**	**0.001**
BPR-5 – RE - BPRe – RA	11	0,67	0.023
BPR-5 – RE - BPRe – RN	11	0,43	0.185
BPR-5 – RE - BPRe – RV	11	0,21	0.532
BPR-5 – RE - BPRe – RE	**10**	**0,18**	**0.612**
BPRe – RA - BPRe – RN	51	0,42	0.002
BPRe – RA - BPRe – RV	52	0,36	0.008
BPRe – RA - BPRe – RE	43	-0,04	0.773
BPRe – RN - BPRe – RV	51	0,43	0.002
BPRe – RN - BPRe – RE	43	0,22	0.142
BPRe – RV - BPRe – RE	43	-0,02	0.899

Em geral, os resultados obtidos demostram magnitudes de correlação entre pequenas e grandes (r = -0,23 e r = 0,94), das quais destacam-se: BPR-5 – RA e BPRe – RA (N= 14; r = 0,28; p = 0,324), BPR-5 – RN e BPRe – RN (N = 5; r = 0,776; e p = 0,123), BPR-5 – RV e BPRe – RV (N = 7; r = 0,946; p = 0.001) e BPR-5 – RE e BPRe – RE (N = 10; ; r = 0,183; p = 0,612). A provas de RN e RV apresentaram magnitudes de correlações grandes e positivas, conforme o esperado para estudos de evidências de validade convergente (Cohen, 1993). Entretanto, para as provas de RA e RE, esperavam-se que as correlações fossem maiores. Sobre isso, o pequeno número amostral (passível de coleta em contexto de isolamento social) pode explicar as correlações de pequena magnitude, sua variação e ausência de significância estatística. A seguir são apresentados seus respectivos diagramas de dispersão, na Figura 1.

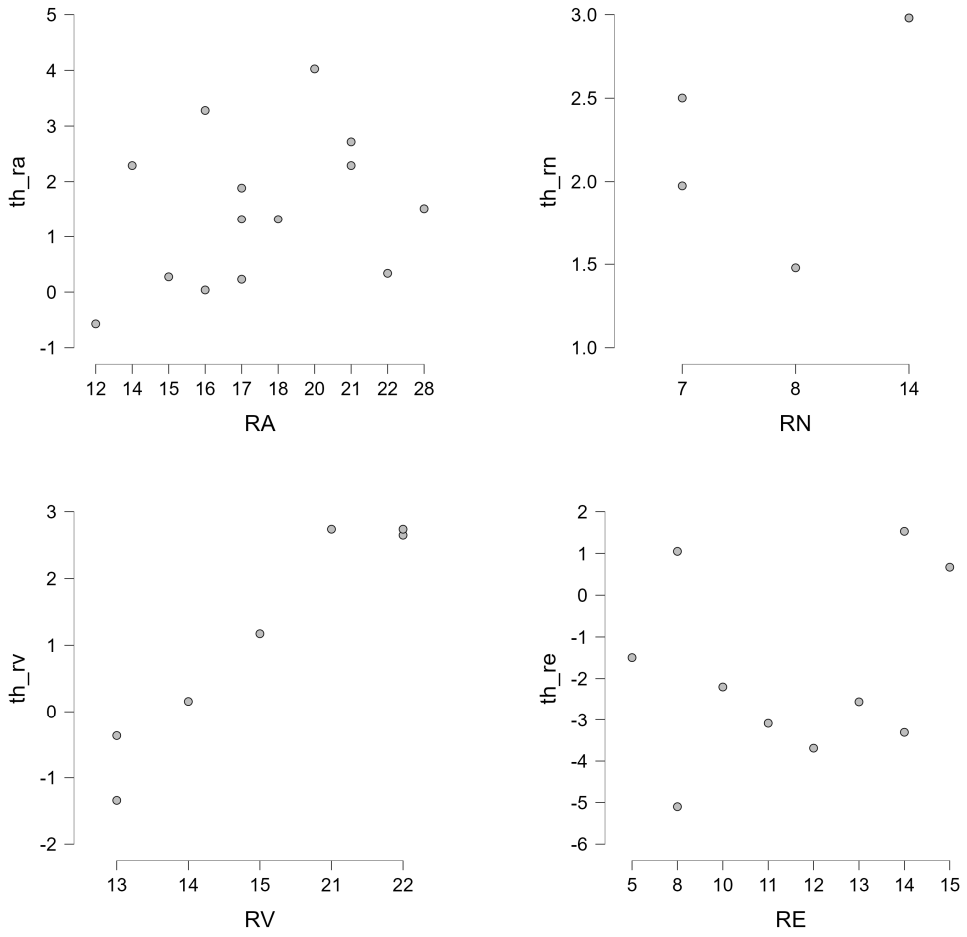

Figura 1. Diagramas de dispersão entre provas de raciocínio da BPR-5 e BPRe.

Observa-se no quadrante superior esquerdo o gráfico de dispersão dos escores das provas de RA, no qual é possível considerar uma magnitude pequena positiva, de 0,28. O quadrante superior direito apresenta o gráfico de dispersão entre os escores das provas de RN, onde a correlação é grande positiva, de 0,77. Nota-se que no quadrante inferior esquerdo, a correlação grande positiva entre as provas de RV, de 0,94. No quadrante inferior direito, a correlação encontrada entre as provas de RE foi pequena positiva, de 0,18. Apesar das correlações apresentadas, apenas a prova de RV apresentou significância estatística (p = 0,001) (Cohen, 1988). Tais resultados não permitem concluir que tratam-se de evidências de validade, como nos

resultados obtivos por Karay et al., (2015), Santana et al. (2017), mas parecem estar de acordo com o esperado. Para estudos posteriores, devem ser consideradas amostras mais consistentes, visto que neste momento não foi possível, em função do isolamento social. A Figura 2 ilustra a matriz de correlações aplicada ao *heatmap* (mapa de calor), ressaltando a magnitude e a direção das associações entre as provas de raciocínio da BPR-5 e da BPRe.

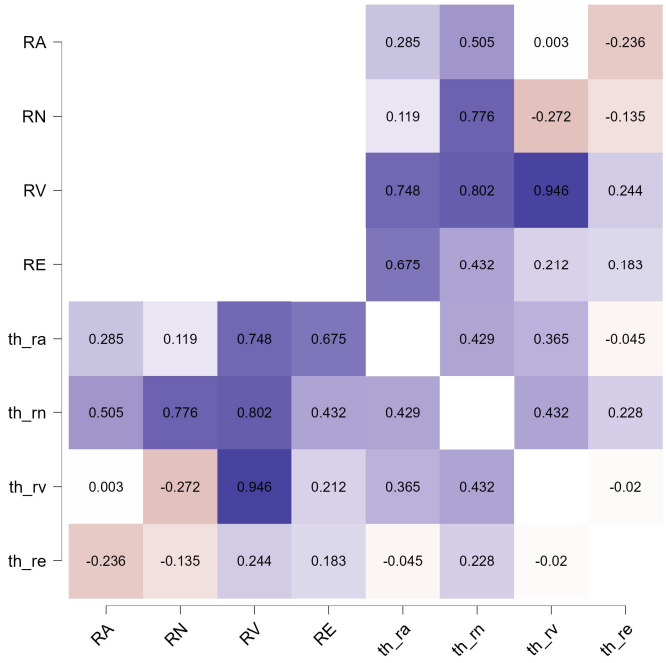

Figura 2. Matriz de correlações entre escores das provas de raciocínio da BPR-5 e BPRe.

A partir do *hetapmap* é possível perceber a diferença no padrão de cada uma das provas de forma visual, onde quanto maior a correlação, mais escura a coloração. Dessa forma, correlacionaram-se em maior magnitude a provas de RV. Em seguida, as prova de RN. Por outro lado, as provas de RA e RE também apresentaram correlações positivas, porém de magnitude pequena (Cohen, 1988). A Tabela 2 expõe o tempo de execução das provas de raciocínios da BPRe.

Tabela 2.
Tempo de execução das provas de raciocínio da BPRe

	RA – tempo total	RN – tempo total	RV – tempo total	RE – tempo total
Validos	52	51	52	44
Ausentes	2	3	2	10
Média	7.46	6.29	3.99	3.39
Desvio Padrão	3.79	3.71	1.47	3.45
Mínimo	0.48	1.08	1.66	0.40
Máximo	15.36	18.26	9.96	14.68

No que se refere ao tempo de aplicação dos instrumentos, foi alcançada uma redução, com diferenças significativas entre os dois grupos, considerando o tempo de aplicação de cada uma das provas da BPR-5, conforme o manual do instrumento. Em média, todas a provas da BPRe foram executadas em tempo inferior às da BPR-5, o que resultou em menor tempo de testagem, sendo: RA de 12 minutos para 7,46 minutos; RN de 18 minutos para 6,29 minutos; RV de 10 minutos para 3,99 minutos; RE de 18 minutos para 3,39 minutos. Tais resultados evidenciam que a implementação de CATs tornam a testagem mais breve, causando menos fadiga nos testandos, sem comprometer a precisão do instrumento (Bjorner et al., 2007; Jaloto, 2018; Nunes & Primi, 2009; Spenassato et al., 2016; Urbina, 2007). Da mesma forma, corroboram com os achados do estudo de Karay et al. (2015), o qual aponta uma redução no tempo de aplicação do instrumento a partir do uso de teste adaptativo computadorizado.

Estudo 2: Análise do funcionamento diferencial dos itens e estrutura interna

Participantes

A amostra deste estudo foi composta no total por 13232 participantes, de ambos os sexos, com idades variando entre 11 e 23 anos (M = 18,24; DP = 2,95). Parte da amostra respondeu a BPR-5 (n = 12545), enquanto outra parte respondeu a BPRe (n = 687). Os dados são provenientes de contextos escolares, sendo 14,56% da amostra com escolaridade em nível de Ensino Fundamental II (n = 1927), 42,56% de Ensino Médio (n = 5631) e 42,82% de Ensino Superior (n

= 5666). Não declararam o nível de escolaridade 0,06% dos participantes (*n* = 8). Essa amostra é parte da base de dados dos estudos de validade e normatização da BPR-5 e BPRe mantidos no Laboratório de Avaliação Psicológica e Educacional (LabAPE) da USF.

Instrumentos

Bateria de Provas de Raciocínio – BPR-5 (Almeida & Primi, 2000)

O instrumento tem como finalidade a avaliação das habilidades cognitivas, especialmente a capacidade de raciocínio. É constituída por cinco provas de raciocínio, a saber: Raciocínio Abstrato (RA), Raciocínio Verbal (RV), Raciocínio Espacial (RE), Raciocínio Numérico (RN) e Raciocínio Mecânico (RM). Para o presente estudo, foram utilizadas as provas RA, RV, RE e RN, descritas abaixo. Em função do caráter opcional atribuído à prova de RM, esta não será utilizada nesse estudo, considerando que requer conhecimentos físico-mecânicos, comumente utilizada em processos avaliativos que buscam verificar especificamente essas habilidades.

a) RA, composta por 25 itens, nos quais é necessário que se descubra a relação existente entre os dois primeiros termos e aplica-la ao terceiro, para se identificar a quarta figura entre as cinco alternativas de resposta, com limite de tempo de 12 minutos. O índice de precisão da prova é de 0,82;

b) RV, composta por 25 itens nos quais a relação analógica existente entre um primeiro par de palavras deverá ser descoberta e aplicada de forma que se identifique a quarta palavra entre as cinco alternativas de resposta que mantenha a mesma relação com uma terceira apresentada, com limite de tempo de 10 minutos. O índice de precisão da prova é de 0,74;

c) RE, composta por 20 itens que implicam descobrir o movimento de cubos, por meio da análise das diferentes faces, escolhendo entre as alternativas de resposta a representação do cubo que se seguiria se o movimento descoberto fosse aplicado ao último cubo da série, com limite de tempo de 18 minutos. O índice de precisão da prova é de 0,62;

d) RN, composta por 20 itens nos quais deve-se identificar a relação aritmética que rege as progressões nas séries e aplicá-la respondendo quais seriam os dois últimos números que completariam a série, com limite de tempo de 18 minutos. O índice de precisão da prova é de 0,75.

Questionário Sóciodemográfico

Elaborado para cadastro como pré-requisito de acesso a BPRe foram coletadas informações de caracterização da amostra. Os itens comtemplam data de nascimento, sexo e instituição de ensino ao qual o participante é vinculado.

Bateria de Provas de Raciocínio – BPRe – sistema CAT (Primi, 2013)

O sistema BPRe CAT consiste em uma versão adaptativa computadorizada via *web* da BPR-5. O Banco de Itens (BI) do sistema é composto por 185 itens, calibrados por meio da TRI a partir do modelo de Rasch, pelo software Winsteps, com o procedimento de estimação JLME (*Joint Maximum Likelihood Estimation*). O teste é dividido entre quatro provas de raciocínio: Raciocínio Abstrato (RA, 52 itens), Raciocínio Verbal (RV, 57 itens), Raciocínio Espacial (RE, 28 itens) e Raciocínio Numérico (RN, 48 itens). Apresentação dos itens inclui recursos audiovisuais, sendo opção do testando a leitura e/ou escuta de enunciados e opções de resposta. Na etapa inicial da testagem, o sistema apresenta um item com dificuldade baixa. Conforme a resposta é emitida, o item é automaticamente corrigido e obtém-se a estimativa de *theta* a partir dessa informação. Em seguida, o *software* calcula o erro de medida. Com um único item, o erro de medida será muito alto, acima do que seria tolerável, portanto, escolhe-se um novo item cuja informação seja máxima para o *theta* atual e repete-se o processo. A cada ciclo, à medida que se aumenta o número de itens aplicados, o erro vai diminuindo até que atinja um valor aceitável (por exemplo, precisão de 0,80), momento em que se conclui a avaliação. Este processo é repetido

para as provas RA, RV, RE e RN, o que permite obter um escore para cada umas delas, assim como um escore geral.

Procedimentos

Para esta etapa, foram utilizados dois bancos de dados pré-existentes. O primeiro, referente a BPR-5, teve seus dados coletados em estudo de normatização do instrumento (Almeida & Primi, 2000) e foi disponibilizado pelos pesquisadores. O tempo estimado para aplicação do instrumento dispensa aproximadamente 60 minutos.

O segundo banco de dados foi construído pela associação de coletas realizadas em estudos preliminares pelo Laboratório de Avaliação Psicológica e Educacional (LabAPE) e coletas *on-line* realizadas durante o ensino à distância, como ferramenta de aprendizagem para alunos do curso de Psicologia de uma universidade privada do interior do estado de São Paulo. Para a aplicação coletiva, os estudantes receberam o link de acesso ao instrumento durante o período regular de aula e foram acompanhados em tempo real pelo docente, pelo responsável da pesquisa e pela equipe técnica que dá suporte ao instrumento. O tempo para responder a bateria de provas foi, em média, de 30 minutos.

As instituições autorizaram a coleta nas respectivas escolas e o projeto foi aprovado pelo Comitê de Ética em Pesquisa da Universidade São Francisco (USF), de acordo com o CAAE nº 20181619.5.0000.5514. Todos os participantes atestaram sua participação voluntária na pesquisa mediante a concordância com o Termo de Consentimento Livre e Esclarecido (TCLE).

Análise de dados

Os dados foram analisados em *software* R (R Core Team, 2020). Primeiramente, foram realizadas análises descritivas para a caracterização da amostra, considerando variáveis como sexo, idade e escolaridade dos participantes. Após, foi realizada aplicação do método de regressão logística (Swaminathan & Rogers, 1990) para identificar itens com DIF. Para isso, são

considerados dois grupos na amostra: indivíduos que responderam a BPR-5 (grupo de referência) e os que responderam a BPRe (grupo focal) (Ayala, 2009). testando o seguinte modelo:

$$P(Y_{pi} = 1|Z_p, G_p) = \frac{e^{\beta_{i0} + \beta_{i1} Z_p + \beta_{i2} G_p + \beta_{i3} Z_p \cdot G_p}}{1 + e^{\beta_{i0} + \beta_{i1} Z_p + \beta_{i2} G_p + \beta_{i3} Z_p \cdot G_p}}$$

Onde:

$P(Y_{pi} = 1|Z_p, G_p)$: probabilidade da pessoa *p* do grupo G_p com escore TRI Z_p acertar o item.

β_{i0}: dificuldade do item *i* (termo constante da regressão).

β_{i1}: coeficiente associado ao escore TRI Z_p, isto é, habilidade do sujeito. Note que esse escore foi estimado conjuntamente nos dois grupos por isso é equalizado está na mesma métrica seja a forma em papel ou informatizada.

β_{i2} : coeficiente associado ao grupo G_p. Esse efeito representa o DIF, isto é, mudanças na dificuldade devido ao grupo (computadorizado – focal vs papel - referência) depois de controlada a habilidade (DIF uniforme).

β_{i3} : coeficiente associado à interação grupo com habilidade $Z_p : G_p$ indicando o quando a relação habilidade/acerto muda nos grupos. Isso indica DIF para discriminação do item (DIF não uniforme).

Em seguida, a matriz de dados foi analisada por meio de análise fatorial confirmatória via TRI implementada pelo pacote *mirt* usando o método *Full-Information maximum likelihood item Factor Analysis* FIFA (Chalmers, 2012) tendo a finalidade identificar a unidimensionalidade de cada teste da BPRe. Considerou-se valores de carga fatorial acima de 0,30 para pertinência aos fatores (Pasquali, 2017b).

Resultados e Discussão

Sendo a BPR-5 calibrada com os mesmos parâmetros da BPRe, considera-se que ambas possuem o nível de *theta* em uma mesma escala. Dessa forma, buscou-se comparar a distribuição do *theta* dos alunos na prova de RA, divididos por nível de ensino (fundamental, médio e superior). Os resultados são apresentados na Figura 1.

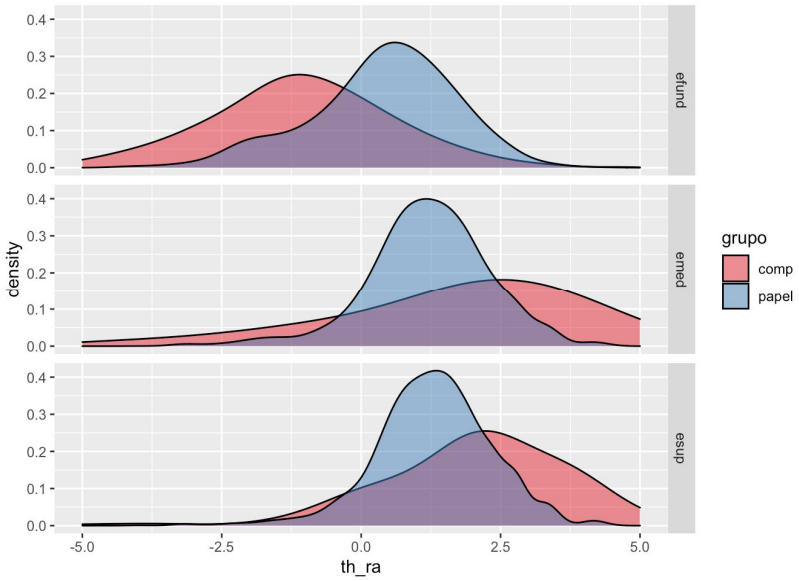

Figura 1. Distribuição de *theta* em RA na BPR-5 e BPRe por nível de ensino.

Observa-se que há uma diferença nos níveis de habilidade entre os grupos que executaram a prova de RA no ensino fundamental. Para este nível de ensino, os alunos que realizaram a prova em lápis e papel (BPR-5) apresentaram um *theta* maior do que os alunos os que fizeram a prova em computador (BPRe). A amostra em lápis e papel (BPR-5) inclui alunos de ensino fundamental, que utilizam a Forma B do instrumento, com itens conforme o nível de habilidade esperado para sua faixa etária. Dessa forma, a prova de RA demonstrou maior nível de dificuldade na BPRe, aplicada em computador para este grupo. Sendo as provas de RA equalizadas, essa diferença pode estar relacionada com características da amostra, a qual é composta por diferentes séries de ensino fundamental para os diferentes grupos (lápis e papel e

computador). Entre alunos do ensino médio e superior, os sujeitos demostraram maior nível de habilidade na execução da BPRe.

Quanto à regressão logística, a qual prediz a probabilidade de acerto aos itens da prova de RA, foram considerados os níveis *theta* dos sujeitos (Beta1) e o nível de dificuldade dos itens e seus escores em relação ao grupo (Beta2), indicando a presença de DIF. Na Tabela 1 são apresentados os resultados da análise.

Tabela 1.
Regressão logística com modelo de dois parâmetros em itens da prova de RA

Item	LR	*P*	sig	R2	ZT	JZ	(Intercept)	SCORE	GROUP	SCORE:GROUP
RA0026	7.89	0.00	**	0.01	A	A	-11.79	0.73	1.03	0
RA0001	10.91	0.00	***	0.00	A	A	-11.74	0.76	0.91	0
RA0003	7.38	0.01	**	0.00	A	A	-13.11	0.79	0.93	0
RA0027	4.02	0.04	*	0.00	A	A	-13.80	0.84	0.53	0
RA0005	39.06	0.00	***	0.00	A	A	-12.92	0.75	-1.20	0
RA0009	2.14	0.14		0.00	A	A	-12.55	0.71	0.00	0
RA0006	0.17	0.68		0.00	A	A	-11.89	0.68	0.00	0
RA0007	4.29	0.04	*	0.00	A	A	-11.87	0.62	0.55	0
RA0028	0.39	0.53		0.00	A	A	-8.37	0.45	0.00	0
RA0029	34.02	0.00	***	0.00	A	A	-10.34	0.56	0.91	0
RA0030	2.59	0.11		0.00	A	A	-10.94	0.58	0.00	0
RA0031	0.81	0.37		0.00	A	A	-7.94	0.45	0.00	0
RA0010	0.08	0.78		0.00	A	A	-9.08	0.57	0.00	0
RA0008	0.04	0.85		0.00	A	A	-10.04	0.54	0.00	0
RA0012	3.64	0.06	.	0.00	A	A	-9.73	0.52	0.00	0
RA0013	1.96	0.16		0.00	A	A	-8.25	0.45	0.00	0
RA0016	0.00	0.94		0.00	A	A	-8.08	0.45	0.00	0
RA0014	23.30	0.00	***	0.00	A	A	-10.96	0.56	-0.76	0
RA0015	7.38	0.01	**	0.00	A	A	-10.53	0.50	-0.43	0
RA0018	16.39	0.00	***	0.00	A	A	-7.14	0.29	1.07	0
RA0019	6.11	0.01	*	0.00	A	A	-8.01	0.40	-0.36	0
RA0024	504.02	0.00	***	0.90	C	C	-10.63	0.50	-4.29	0
RA0022	6.24	0.01	*	0.00	A	A	-11.12	0.51	-0.42	0
RA0023	62.66	0.00	***	0.00	A	A	-7.21	0.27	1.50	0
RA0002	10.34	0.00	**	0.00	A	A	-14.16	0.85	0.98	0
RA0011	1.34	0.25		0.00	A	A	-3.65	0.27	0.00	0
RA0017	0.25	0.62		0.00	A	A	-4.53	0.15	0.00	0
RA0020	39.66	0.00	***	0.00	A	A	-12.53	0.61	-1.05	0
RA0021	2.45	0.12		0.00	A	A	-9.56	0.45	0.00	0
RA0025	0.00	0.98		0.00	A	A	-10.25	0.46	0.00	0

Nota. *** $p < 0$, ** $p < 0{,}001$, * $p < 0{,}05$, ‘ ’ $p < 1$. Tamanho do efeito: A: insignificante, B: moderado, C: grande, conforme Zumbo & Thomas (ZT): 0 'A' 0,13 'B' 0,26 'C' 1 e Jodoin & Gierl (JG): 0 'A' 0,035 'B' 0,07 'C' 1.

Dentre os 30 itens analisados na prova de RA, o método de regressão logística detectou a presença de DIF uniforme em 9, a saber: RA0001, RA0005, RA0029, RA0014, RA0018, RA0024, RA0023, RA0002 e RA0020, os quais foram estatisticamente significativos ($p < 0$). Destes, apenas RA0024 apresentou efeito de magnitude grande conforme Nagelkerke's ($R^2 = 0,90$; $p < 0$), classificado como C tanto nos parâmetros estabelecidos por Zumbo e Thomas (ZT), quanto para Jodoin e Gierl (JZ). Para os demais itens, a pequena magnitude encontrada é negligenciável. Dessa forma, nota-se que o item RA0024 apresenta maior nível de dificuldade na BPRe, realizado em computador.

Posteriormente, foi incluído um parâmetro no modelo, o Beta3, com a finalidade de testar diferenças na capacidade de discriminação dos itens entre os grupos. Assim, os dados foram analisados considerando os níveis *theta* dos sujeitos (Beta1) e o nível de dificuldade dos itens, seus escores em relação ao grupo (Beta2) e a interação theta *vs* grupo indicando DIF na capacidade de discriminação dos itens (Beta3). Os resultados são apresentados na Tabela 2.

Tabela 2.
Regressão logística com modelo de três parâmetros em itens da prova de RA

Item	LR	*P*	Sig	R2	ZT	JZ	(Intercept)	SCORE	GROUP	SCORE:GROUP
RA0026	23.18	0.00	***	0.03	A	A	-11.26	0.70	-41.08	2.74
RA0001	14.80	0.00	***	0.00	A	A	-11.66	0.75	-11.34	0.75
RA0003	11.57	0.00	**	0.00	A	A	-13.05	0.79	-13.74	0.96
RA0027	7.52	0.02	*	0.01	A	A	-13.26	0.81	-7.47	0.49
RA0005	64.68	0.00	***	0.00	A	A	-13.31	0.77	8.64	-0.57
RA0009	14.30	0.00	***	0.00	A	A	-12.33	0.70	-20.48	1.13
RA0006	11.77	0.00	**	0.00	A	A	-11.71	0.67	-12.75	0.74
RA0007	12.75	0.00	**	0.00	A	A	-11.82	0.62	-20.88	1.14
RA0028	9.90	0.01	**	0.01	A	A	-7.89	0.42	-12.44	0.63
RA0029	35.70	0.00	***	0.00	A	A	-9.73	0.53	-0.87	0.10
RA0030	17.54	0.00	***	0.01	A	A	-10.55	0.56	-53.74	2.78
RA0031	6.37	0.04	*	0.00	A	A	-7.63	0.43	-14.95	0.81
RA0010	9.07	0.01	*	0.00	A	A	-9.01	0.57	-13.43	0.83
RA0008	7.93	0.02	*	0.00	A	A	-10.00	0.53	-23.31	1.18
RA0012	22.44	0.00	***	0.00	A	A	-9.56	0.51	-14.19	0.69
RA0013	18.25	0.00	***	0.89	A	A	-8.15	0.45	-30.39	1.62
RA0016	6.19	0.05	*	0.00	A	A	-8.06	0.45	-20.00	1.07
RA0014	23.91	0.00	***	0.00	A	A	-10.92	0.56	-2.17	0.07
RA0015	11.11	0.00	**	0.00	A	A	-10.46	0.50	-4.75	0.19
RA0018	33.44	0.00	***	0.00	A	A	-7.10	0.29	-30.12	1.58
RA0019	10.22	0.01	**	0.00	A	A	-7.91	0.40	-3.46	0.14

RA0024	505.81	0.00	***	0.90	C	C	-10.67	0.50	0.76	-0.20
RA0022	8.71	0.01	*	0.00	A	A	-11.06	0.51	-4.54	0.17
RA0023	83.04	0.00	***	0.01	A	A	-7.10	0.27	-20.71	0.90
RA0002	18.60	0.00	***	0.00	A	A	-13.98	0.84	-15.08	1.06
RA0011	23.95	0.00	***	0.00	A	A	-3.56	0.26	-62.16	3.50
RA0017	3.97	0.14		0.00	A	A	-4.53	0.15	0.00	0.00
RA0020	39.89	0.00	***	0.00	A	A	-12.50	0.61	-2.10	0.05
RA0021	5.70	0.06	.	0.00	A	A	-9.56	0.45	0.00	0.00
RA0025	38.51	0.00	***	0.00	A	A	-10.07	0.45	-42.11	1.75

Nota. *** $p < 0$, ** $p < 0{,}001$, * $p < 0{,}01$, . $p < 0, 05$, ‘ ’ $p < 1$. Tamanho do efeito: A: insignificante, B: moderado, C: grande, conforme Zumbo & Thomas (ZT): 0 'A' 0,13 'B' 0,26 'C' 1 e Jodoin & Gierl (JG): 0 'A' 0,035 'B' 0,07 'C' 1.

Após a inclusão de um terceiro parâmetro (Beta3), foram identificados 16 dentre os 30 itens com DIF na prova de RA, a saber: RA0026, RA0001, RA0005, RA0009, RA0029, RA0030, RA0012, RA0013, RA0014, RA0018, RA0024, RA0023, RA0002, RA0011, RA0020, RA0025, os quais obtiveram nível de significância estatística ($p < 0$). Todavia, apresentaram magnitudes de efeito grande apenas os itens RA0013 ($R^2 = 0{,}89$; $p < 0$; ZT = A; JZ = A) e RA0024 ($R^2 = 0{,}90$; $p < 0$; ZT = C; JZ = C), sendo este classificado como C, conforme Zumbo e Thomas (ZT) e Jodoin e Gierl (JZ). Para os demais itens, a magnitude do efeito foi pequena. Assim, a Figura 2 demonstra a CCI de RA0024, que apresentou maior nível de dificuldade na BPRe, em relação à BPR-5.

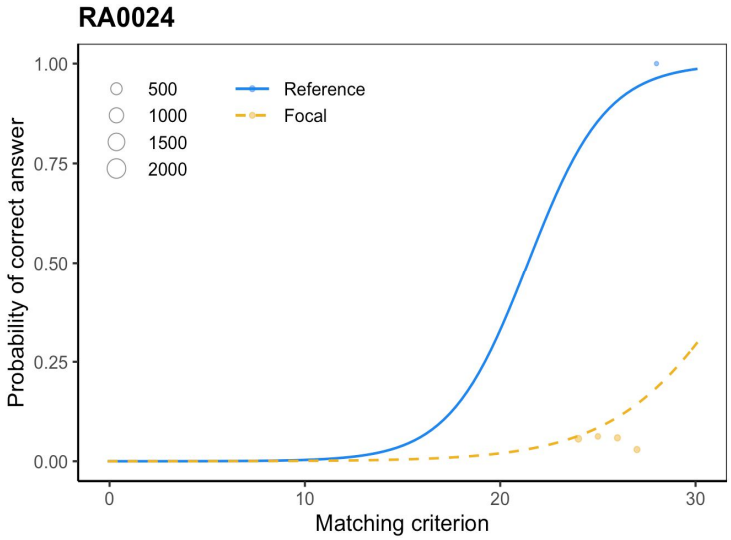

Figura 2. Curva Característica do Item RA0024.
Nota. Linha contínua: grupo de referência (BPR-5); linha tracejada: grupo focal (BPRe).

Para o grupo focal, o qual realizou a BPRe, o item RA0024 apresentou maior nível de dificuldade do que para o grupo de referência, que realizou a BPR-5 ainda que possuindo habilidades estimadas semelhantes. A diferença observada no parâmetro de discriminação (Beta3) caracteriza este item com DIF como não uniforme, que pode ser observado pela diferença na inclinação da curva modelada nos grupos. Em seguida, foi comparada a distribuição *theta* dos alunos na prova de RV, agrupados por nível de ensino (fundamental, médio e superior). Os resultados são apresentados na Figura 3.

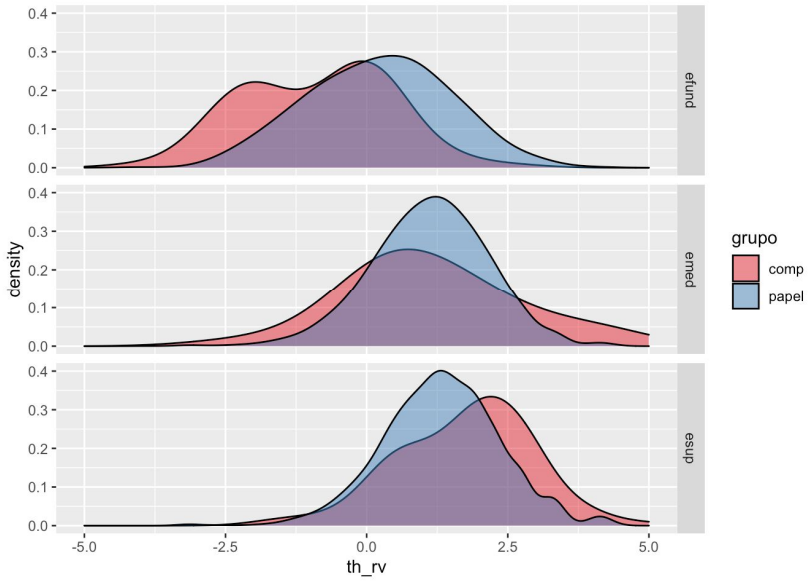

Figura 3. Distribuição de *theta* em RV na BPR-5 e BPRe por nível de ensino.

Os resultados indicaram diferenças nos níveis de habilidade na prova de RV entre alunos do ensino fundamental, os quais apresentaram um *theta* maior no teste em formato lápis e papel (BPR-5), assim como alunos de ensino médio, de forma menos acentuada. Considerando a equalização dos níveis de dificuldade das provas, essa diferença pode ser explicada pelo fato de os sujeitos que compõem a amostra estarem cursando diferentes séries de ensino. Para alunos de ensino superior, a os níveis de *theta* são maiores na BPRe, executada em computador.

Posteriormente, foi realizada a regressão logística com a finalidade de verificar a existência de itens com DIF. Na Tabela 3 são apresentados os resultados da análise.

Tabela 3.
Regressão logística com modelo de dois parâmetros em itens da prova de RV

Item	LR	*p*	Sig	R2	ZT	JZ	(Intercept)	SCORE	GROUP	SCORE:GROUP
RV0001	41.65	0.00	***	0.00	A	A	-10.23	0.70	1.71	0
RV0002	3.40	0.07	.	0.00	A	A	-12.88	0.79	0.00	0
RV0026	0.93	0.34		0.00	A	A	-11.21	0.71	0.00	0
RV0006	4.20	0.04	*	0.00	A	A	-11.34	0.73	0.35	0
RV0007	0.84	0.36		0.00	A	A	-10.12	0.66	0.00	0
RV0027	16.68	0.00	***	0.01	A	A	-7.07	0.43	1.16	0
RV0009	45.34	0.00	***	0.00	A	A	-8.89	0.54	1.24	0
RV0028	17.08	0.00	***	0.01	A	A	-10.26	0.55	-0.87	0
RV0029	2.27	0.13		0.00	A	A	-14.09	0.79	0.00	0
RV0030	11.29	0.00	***	0.00	A	A	-10.74	0.61	-0.56	0
RV0031	2.07	0.15		0.00	A	A	-12.12	0.75	0.00	0
RV0014	1.51	0.22		0.00	A	A	-7.10	0.42	0.00	0
RV0008	1.06	0.30		0.00	A	A	-9.10	0.50	0.00	0
RV0016	0.07	0.78		0.00	A	A	-12.02	0.71	0.00	0
RV0017	0.93	0.34		0.00	A	A	-9.42	0.57	0.00	0
RV0019	7.74	0.01	**	0.00	A	A	-13.82	0.74	-0.48	0
RV0020	2.94	0.09	.	0.00	A	A	-8.68	0.45	0.00	0
RV0021	43.18	0.00	***	0.00	A	A	-9.69	0.48	1.12	0
RV0022	4.14	0.04	*	0.00	A	A	-7.63	0.43	-0.37	0
RV0023	17.21	0.00	***	0.00	A	A	-13.30	0.69	-0.74	0
RV0032	2.15	0.14		0.00	A	A	-12.61	0.60	0.00	0
RV0005	26.38	0.00	***	0.00	A	A	-10.84	0.63	0.94	0
RV0010	13.78	0.00	***	0.00	A	A	-6.70	0.41	0.47	0
RV0011	0.01	0.91		0.00	A	A	-7.52	0.33	0.00	0
RV0012	3.66	0.06	.	0.00	A	A	-7.37	0.47	0.00	0
RV0013	32.49	0.00	***	0.00	A	A	-5.84	0.28	-0.95	0
RV0015	78.56	0.00	***	0.00	A	A	-8.40	0.46	1.26	0
RV0024	10.85	0.00	***	0.00	A	A	-11.69	0.54	-0.75	0

Nota. *** $p < 0$, ** $p < 0{,}001$, * $p < 0{,}05$, ‘ ’ $p < 1$. Tamanho do efeito: A: insignificante, B: moderado, C: grande, conforme Zumbo & Thomas (ZT): 0 'A' 0,13 'B' 0,26 'C' 1 e Jodoin & Gierl (JG): 0 'A' 0,035 'B' 0,07 'C' 1.

A regressão logística indicou a presença de DIF em 12 dos 28 itens analisados, a saber: RV0001, RV0027, RV0009, RV0028, RV0030, RV0021, RV0023, RV0005, RV0010, RV0013, RV0015 e RV0024, os quais apresentaram significância estatística ($p < 0$). No entanto, nenhum

dos itens apresentou grande efeito de magnitude, conforme os parâmetros de Zumbo e Thomas (ZT) e Jodoin e Gierl (JZ), classificando-os como A, o que indica uma magnitude de efeito pequena, considerada negligenciável. Assim, a prova de RV não apresenta itens com DIF, sendo a dificuldade dos itens equivalentes, não importando o meio de aplicação. A análise foi complementada com a adição de um terceiro parâmetro, considerando: os níveis *theta* dos sujeitos (Beta1) e o nível de dificuldade dos itens, seus escores em relação ao grupo (Beta2) e a capacidade de discriminação dos itens (Beta3). Na Tabela 4 são apresentados os resultados.

Tabela 4.
Regressão logística com modelo de três parâmetros em itens da prova de RV

Itens	LR	P	sig	R2	ZT	JZ	(Intercept)	SCORE	GROUP	SCORE:GROUP
RV0001	59.69	0.00	***	0.00	A	A	-10.07	0.69	-22.52	1.84
RV0002	3.64	0.16		0.00	A	A	-12.88	0.79	0.00	0.00
RV0026	8.73	0.01	*	0.01	A	A	-10.90	0.69	-13.76	0.90
RV0006	8.32	0.02	*	0.00	A	A	-11.18	0.72	-3.38	0.25
RV0007	26.75	0.00	***	0.00	A	A	-9.87	0.65	-37.31	2.47
RV0027	19.19	0.00	***	0.01	A	A	-6.93	0.42	-7.67	0.52
RV0009	45.34	0.00	***	0.00	A	A	-8.89	0.54	1.35	-0.01
RV0028	51.66	0.00	***	0.04	A	A	-9.48	0.51	-35.67	1.77
RV0029	47.79	0.00	***	0.00	A	A	-12.68	0.71	-35.14	1.87
RV0030	64.96	0.00	***	0.00	A	A	-9.31	0.53	-29.96	1.62
RV0031	17.16	0.00	***	0.01	A	A	-10.75	0.68	-13.16	0.79
RV0014	29.80	0.00	***	0.00	A	A	-7.03	0.42	-252.59	14.02
RV0008	25.80	0.00	***	0.00	A	A	-9.01	0.50	-32.59	1.72
RV0016	23.36	0.00	***	0.00	A	A	-11.94	0.70	-36.22	2.11
RV0017	32.55	0.00	***	0.00	A	A	-9.36	0.57	-63.75	3.50
RV0019	58.47	0.00	***	0.00	A	A	-13.59	0.73	-39.82	1.98
RV0020	9.57	0.01	**	0.00	A	A	-8.64	0.44	-6.86	0.31
RV0021	47.84	0.00	***	0.00	A	A	-9.62	0.48	-4.93	0.29
RV0022	22.57	0.00	***	0.00	A	A	-7.51	0.43	-16.39	0.82
RV0023	42.21	0.00	***	0.00	A	A	-13.11	0.68	-21.08	0.96
RV0032	4.29	0.12		0.00	A	A	-12.61	0.60	0.00	0.00
RV0005	40.13	0.00	***	0.00	A	A	-10.74	0.62	-20.06	1.21
RV0010	13.78	0.00	***	0.00	A	A	-6.71	0.41	0.57	-0.01
RV0011	23.96	0.00	***	0.00	A	A	-7.42	0.33	-24.78	1.09
RV0012	5.78	0.06	.	0.00	A	A	-7.37	0.47	0.00	0.00
RV0013	43.15	0.00	***	0.00	A	A	-5.76	0.27	-11.33	0.46
RV0015	78.76	0.00	***	0.00	A	A	-8.38	0.46	0.45	0.04

RV0024	25.53	0.00	***	0.00	A	A	-11.61	0.54	-33.82	1.44

Nota. *** $p < 0$, ** $p < 0{,}001$, * $p < 0{,}01$, . $p < 0$, 05, ‘ ’ $p < 1$. Tamanho do efeito: A: insignificante, B: moderado, C: grande, conforme Zumbo & Thomas (ZT): 0 'A' 0,13 'B' 0,26 'C' 1 e Jodoin & Gierl (JG): 0 'A' 0,035 'B' 0,07 'C' 1.

Foram detectados 22 itens dentre os 28 analisados com DIF, a saber: RV0001, RV0007, RV0027, RV0009, RV0028, RV0029, RV0030, RV0031, RV0014, RV0008, RV0016, RV0017, RV0019, RV0021, RV0022, RV0023, RV0005, RV0010, RV0011, RV0013, RV0015 e RV0024. Contudo, apesar de apresentarem significância estatística ($p < 0$), a magnitude do efeito encontrada é pequena, sendo classificadas como A, de acordo com os parâmetros estabelecidos por Zumbo e Thomas (ZT) e Jodoin e Gierl (JZ). Assim, pode-se considerar que a prova de RV não apresentou DIF em relação ao nível de dificuldade nos diferentes formatos (lápis e papel e computador). A Figura 4 demonstra a CCI de maior magnitude encontrado em RV.

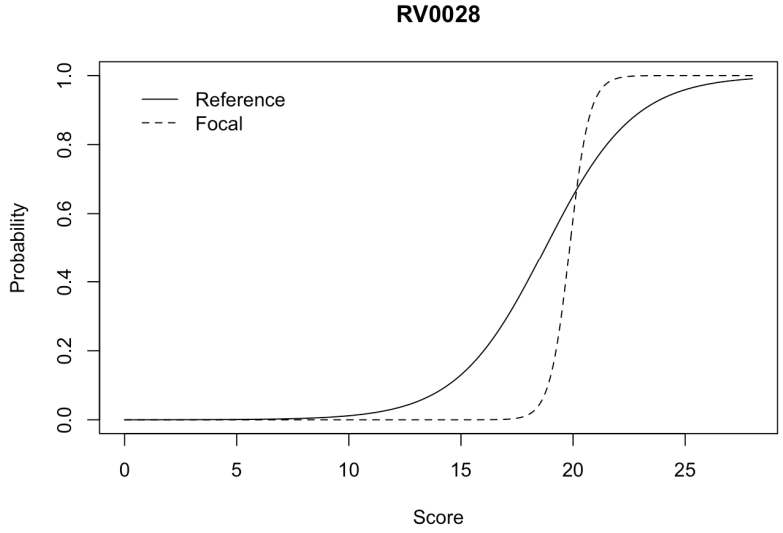

Figura 4. Curva Característica do Item RV0028.
Nota. Linha contínua: grupo de referência (BPR-5); linha tracejada: grupo focal (BPRe).

As diferenças observadas entre as CCI RV0028 demonstram a possibilidade de DIF não uniforme, isto é, o item é mais discriminativo quando aplicado via computador. No entanto, a magnitude do efeito apresentada é pequena ($R^2 = 0{,}04$; $p < 0$; ZT = A; JZ = A). Dessa forma, é considerada negligenciável. Posteriormente, foi comparada a distribuição de *theta* dos alunos na

prova de RN, considerando seus níveis de ensino (fundamental, médio e superior). Os resultados são apresentados na Figura 5.

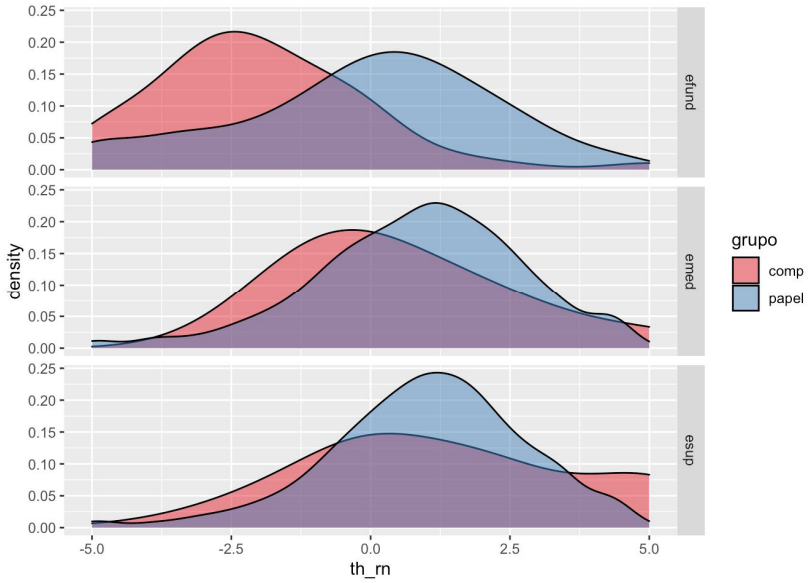

Figura 5. Distribuição de *theta* em RN na BPR-5 e BPRe por nível de ensino.

Os resultados indicaram diferenças nos níveis de habilidade entre os grupos que realizaram a prova de RN em lápis e papel e computador. De forma geral, nota-se que os sujeitos alcançaram um maior nível de *theta* na execução da BPR-5, sendo essa discrepância mais acentuada entre os alunos do ensino fundamental. Considerando a equalização das provas, isso pode ter relação com características da amostra, a qual é composta por grupos de alunos oriundos de diferentes séries do ensino fundamental. Para alunos de ensino médio e superior, essa diferença tem uma proporção relativamente pequena. Para verificar a existência de itens na prova de RV, foi realizada a regressão logística, a qual tem seus resultados apresentados na Tabela 5.

Tabela 5.
Regressão logística com modelo de dois parâmetros em itens da prova de RN

Item	LR	*P*	sig	R2	ZT	JZ	(Intercept)	SCORE	GROUP	SCORE:GROUP
RN0001	40.05	0.00	***	0.00	A	A	-0.20	0.48	-1.73	0
RN0021	1.24	0.27		0.00	A	A	-8.30	1.11	0.00	0
RN0002	0.46	0.50		0.00	A	A	-4.15	0.91	0.00	0
RN0004	0.90	0.34		0.00	A	A	-6.92	1.07	0.00	0

RN0023	2.33	0.13		0.00	A	A	-10.22	1.46	0.00	0
RN0005	17.20	0.00	***	0.00	A	A	-7.06	0.97	-0.55	0
RN0006	12.95	0.00	***	0.00	A	A	-6.87	1.03	-0.82	0
RN0009	8.64	0.00	**	0.00	A	A	-9.46	1.15	-0.50	0
RN0010	2.50	0.11		0.00	A	A	-8.54	1.03	0.00	0
RN0012	5.58	0.02	*	0.00	A	A	-11.41	1.33	-0.76	0
RN0026	11.28	0.00	***	0.02	A	A	-10.02	1.05	-1.13	0
RN0027	4.57	0.03	*	0.01	A	A	-10.13	1.02	-0.93	0
RN0003	0.76	0.38		0.00	A	A	-7.22	1.26	0.00	0
RN0007	0.15	0.70		0.00	A	A	-8.83	1.29	0.00	0
RN0017	12.32	0.00	***	0.00	A	A	-6.40	0.75	-0.74	0

Nota. *** $p < 0$, ** $p < 0{,}001$, * $p < 0{,}05$, ‘ ’ $p < 1$. Tamanho do efeito: A: insignificante, B: moderado, C: grande, conforme Zumbo & Thomas (ZT): 0 'A' 0,13 'B' 0,26 'C' 1 e Jodoin & Gierl (JG): 0 'A' 0,035 'B' 0,07 'C' 1.

O método de regressão logística detectou 6 dentre os 15 itens analisados na prova de RN que apresentaram DIF, sendo estatisticamente significativos ($p < 0$). São eles: RN0001, RN0005, RN0006, RN0009, RN0026 e RN0017. No entanto, a magnitude do efeito demonstrada pelos respectivos itens é pequena, considerando os parâmetros de Zumbo e Thomas (ZT), quanto para Jodoin e Gierl (JZ), classificando-os como A. Ressalta-se que a discrepância entre o número de itens descritos e o número de itnes analisados se deu em função da matriz esparsa de dados. De modo complementar, foi adicionado um terceiro parâmetro ao modelo para realização da regressão logística. A Tabela 6 apresenta os resultados.

Tabela 6.
Regressão logística com modelo de três parâmetros em itens da prova de RN

Itens	**LR**	***P***	**Sig**	**R2**	**ZT**	**JZ**	**(Intercept)**	**SCORE**	**GROUP**	**SCORE:GROUP**
RN0001	88.12	0.00	***	0.01	A	A	0.04	0.44	-66.22	16.00
RN0021	72.12	0.00	***	0.06	A	A	-7.67	1.03	-268.64	35.81
RN0002	33.28	0.00	***	0.00	A	A	-3.86	0.87	-7.21	1.32
RN0004	6.68	0.04	*	0.00	A	A	-6.83	1.06	-1.82	0.29
RN0023	17.70	0.00	***	0.02	A	A	-9.98	1.42	-113.32	16.23
RN0005	35.38	0.00	***	0.00	A	A	-6.88	0.95	-6.17	0.71
RN0006	25.09	0.00	***	0.00	A	A	-6.84	1.03	-112.91	14.03
RN0009	18.27	0.00	***	0.00	A	A	-9.31	1.14	-6.43	0.69
RN0010	27.96	0.00	***	0.00	A	A	-8.41	1.02	-21.43	2.59
RN0012	21.32	0.00	***	0.00	A	A	-11.32	1.32	-119.75	13.23
RN0026	11.46	0.00	**	0.02	A	A	-9.81	1.03	-2.32	0.11
RN0027	28.80	0.00	***	0.05	A	A	-8.60	0.86	-190.06	15.84

RN0003	29.11	0.00	***	0.00	A	A	-7.09	1.24	-17.66	2.97
RN0007	91.42	0.00	***	0.00	A	A	-8.33	1.23	-30.67	4.48
RN0017	58.23	0.00	***	0.00	A	A	-6.21	0.73	-145.59	14.52

Nota. *** $p < 0$, ** $p < 0{,}001$, * $p < 0{,}01$, . $p < 0,\ 05$, ‘ ’ $p < 1$. Tamanho do efeito: A: insignificante, B: moderado, C: grande, conforme Zumbo & Thomas (ZT): 0 'A' 0,13 'B' 0,26 'C' 1 e Jodoin & Gierl (JG): 0 'A' 0,035 'B' 0,07 'C' 1.

Nota-se que 13 dos 15 itens analisados na prova de RN foram apontados com DIF: RN0001, RN0002, RN0023, RN0005, RN0006, RN0009, RN0010, RN0012, RN0026, RN0027, RN0003, RN0007 e RN0017, apresentando significância estatística ($p < 0$). Todavia, a magnitude do efeito observada é pequena, classificando os respectivos itens como A, conforme os parâmetros estabelecidos por Zumbo e Thomas (ZT), quanto para Jodoin e Gierl (JZ). Sendo assim, pode-se considerar que a prova de RN não apresentou itens com DIF. A Figura 6 demonstra a CCI de maior magnitude identificado em RN.

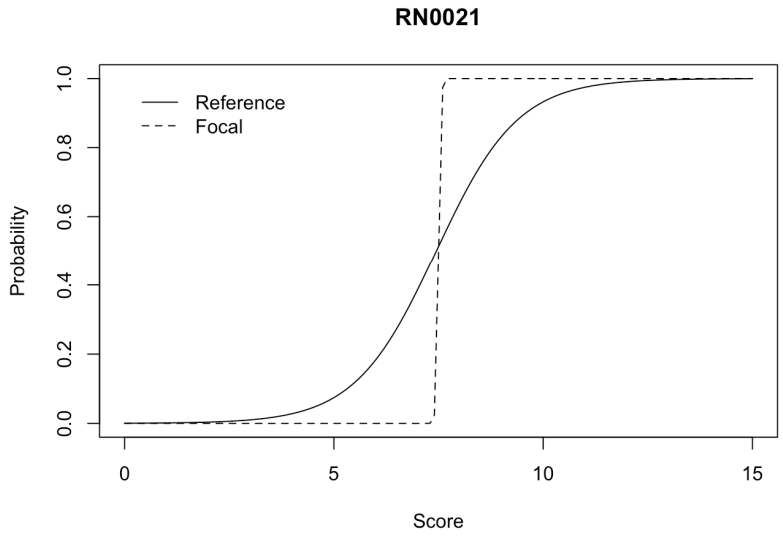

Figura 6. Curva Característica do Item RN0021.
Nota. Linha contínua: grupo de referência (BPR-5); linha tracejada: grupo focal (BPRe).

Observa-se uma diferença na inclinação da CCI RN0021, o que poderia caracterizar a presença de DIF não uniforme. Na Figura 6 o item apresenta uma discriminação quase perfeita quando aplicado no formato eletrônico. Evidentemente esse resultado não pode ser considerado

como uma estimação pontual correta da discriminação do item. Pode ter refletido o alto erro de estimação que temos na amostra que fez o teste no computador. Como nem todos os itens são aplicados em todos os sujeitos na amostra computadorizada, os dados nessa modalidade não são ótimos. Há uma amostra esparsa nos diferentes níveis de habilidade que respondeu no computador. Contudo, apesar da significância estatística, a pequena magnitude do efeito (R^2 = 0,06; $p < 0$; ZT = A; JZ = A) torna a diferença observada negligenciável. Finalmente, foi comparado a desempenho dos alunos por meio da distribuição de *theta* na prova de RE, considerando sues níveis de ensino (fundamental, médio e superior). Os resultados são apresentados na Figura 7.

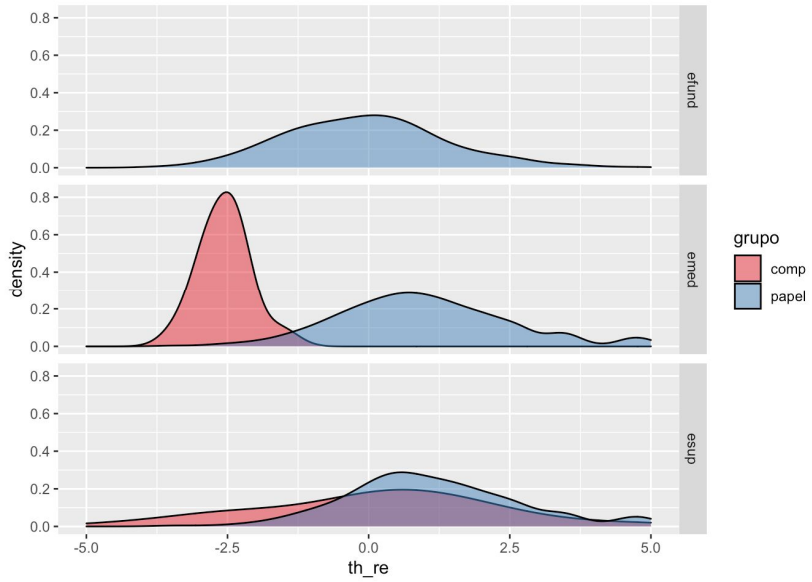

Figura 7. Distribuição de *theta* em RE na BPR-5 e BPRe por nível de ensino.

Para a prova de RE, a amostra em computador não contemplou um número suficiente de alunos do ensino fundamental, impossibilitando a comparação de seus níveis de *theta.* Quanto aos alunos de ensino médio, os que realizaram a prova de RE em computador (BPRe) apresentaram maior habilidade em relação aos que executaram a prova em lápis e papel (BPR-5). Quanto ao grupo de alunos de ensino superior, nota-se que os níveis *theta* são semelhantes, pressupondo que

não há diferenças significativas na execução das provas de RE em lápis e papel e computador. Foi realizada a regressão logística para detecção de itens com DIF, que tens seus resultados apresentados na Tabela 7.

Tabela 7.

Regressão logística com modelo de dois parâmetros em itens da prova de RE

Itens	LR	*P*	Sig	R2	ZT	JZ	(Intercept)	SCORE	GROUP	SCORE:GROUP
RE0009	13.11	0.00	***	0.00	A	A	-10.02	0.58	0.90	0
RE0013	9.62	0.00	**	0.01	A	A	-6.84	0.40	0.96	0
RE0005	0.24	0.62		0.00	A	A	-10.58	0.69	0.00	0
RE0014	2.94	0.09	.	0.00	A	A	-7.55	0.44	0.00	0
RE0015	3.37	0.07	.	0.00	A	A	-8.29	0.48	0.00	0
RE0008	23.86	0.00	***	0.02	A	A	-9.52	0.56	1.40	0
RE0016	0.63	0.43		0.00	A	A	-6.94	0.40	0.00	0
RE0007	2.24	0.13		0.00	A	A	-9.71	0.60	0.00	0
RE0017	11.33	0.00	***	0.00	A	A	-8.37	0.50	0.96	0
RE0018	4.31	0.04	*	0.00	A	A	-8.16	0.47	0.61	0
RE0019	2.56	0.11		0.00	A	A	-9.70	0.55	0.00	0
RE0006	6.58	0.01	*	0.00	A	A	-9.25	0.61	0.70	0
RE0020	3.10	0.08	.	0.00	A	A	-8.10	0.44	0.00	0
RE0021	1.82	0.18		0.00	A	A	-9.53	0.53	0.00	0
RE0022	21.98	0.00	***	0.00	A	A	-9.90	0.57	1.32	0
RE0011	2.18	0.14		0.00	A	A	-7.81	0.40	0.00	0
RE0010	0.10	0.76		0.00	A	A	-10.09	0.57	0.00	0
RE0023	0.32	0.57		0.00	A	A	-8.70	0.46	0.00	0
RE0024	9.90	0.00	**	0.00	A	A	-9.04	0.49	-0.82	0
RE0012	0.34	0.56		0.00	A	A	-7.27	0.37	0.00	0
RE0025	0.22	0.64		0.00	A	A	-8.56	0.54	0.00	0
RE0026	1.95	0.16		0.00	A	A	-10.15	0.51	0.00	0
RE0027	3.96	0.05	*	0.00	A	A	-11.56	0.61	-0.50	0
RE0028	0.19	0.66		0.00	A	A	-10.59	0.53	0.00	0
RE0029	0.16	0.69		0.00	A	A	-9.33	0.48	0.00	0
RE0030	0.66	0.42		0.00	A	A	-12.50	0.61	0.00	0
RE0031	0.66	0.41		0.00	A	A	-10.66	0.53	0.00	0
RE0032	61.24	0.00	***	0.00	A	A	-7.86	0.35	-3.50	0

Nota. *** $p < 0$, ** $p < 0{,}001$, * $p < 0{,}05$, ‘ ’ $p < 1$. Tamanho do efeito: A: insignificante, B: moderado, C: grande, conforme Zumbo & Thomas (ZT): 0 'A' 0,13 'B' 0,26 'C' 1 e Jodoin & Gierl (JG): 0 'A' 0,035 'B' 0,07 'C' 1.

A análise detectou a presença de DIF em 6 dos 28 itens verificados, a saber: RE0009, RE0008, RE0017, RE0022, RE0024 e RE0032, os quais foram estatisticamente significativos ($p < 0$). Entretanto, a magnitude do enfeito apresentada pelos respectivos itens é pequena, sendo

classificados como A, de acordo com os parâmetros de Zumbo e Thomas (ZT) e Jodoin e Gierl (JZ). Para complementar a análise, foi incluso o parâmetro de discriminação (Beta3). Na Tabela 8 são apresentados os resultados.

Tabela 8.

Regressão logística com modelo de três parâmetros em itens da prova de RE

Itens	LR	*P*	Sig	R2	ZT	JZ	(Intercept)	SCORE	GROUP	SCORE:GROUP
RE0009	13.11	0.00	**	0.00	A	A	-10.02	0.58	0.89	0.00
RE0013	11.07	0.00	**	0.01	A	A	-7.03	0.41	3.80	-0.19
RE0005	8.68	0.01	*	0.01	A	A	-9.76	0.64	-8.48	0.56
RE0014	10.56	0.01	**	0.00	A	A	-7.54	0.44	-7.01	0.47
RE0015	7.67	0.02	*	0.00	A	A	-8.33	0.48	-6.06	0.43
RE0008	23.95	0.00	***	0.02	A	A	-9.44	0.56	0.75	0.04
RE0016	10.90	0.00	**	0.00	A	A	-6.91	0.40	-29.23	1.59
RE0007	2.51	0.28		0.00	A	A	-9.71	0.60	0.00	0.00
RE0017	11.47	0.00	**	0.00	A	A	-8.38	0.50	1.80	-0.06
RE0018	4.32	0.12		0.00	A	A	-8.14	0.47	0.00	0.00
RE0019	5.52	0.06	.	0.00	A	A	-9.70	0.55	0.00	0.00
RE0006	6.58	0.04	*	0.00	A	A	-9.25	0.61	0.59	0.01
RE0020	3.21	0.20		0.00	A	A	-8.10	0.44	0.00	0.00
RE0021	2.21	0.33		0.00	A	A	-9.53	0.53	0.00	0.00
RE0022	22.78	0.00	***	0.00	A	A	-9.87	0.57	-0.79	0.14
RE0011	2.70	0.26		0.00	A	A	-7.81	0.40	0.00	0.00
RE0010	0.60	0.74		0.00	A	A	-10.09	0.57	0.00	0.00
RE0023	0.32	0.85		0.00	A	A	-8.70	0.46	0.00	0.00
RE0024	11.51	0.00	**	0.00	A	A	-9.01	0.49	-4.90	0.20
RE0012	0.62	0.73		0.00	A	A	-7.27	0.37	0.00	0.00
RE0025	7.69	0.02	*	0.00	A	A	-8.42	0.53	-5.53	0.35
RE0026	12.56	0.00	**	0.00	A	A	-10.08	0.50	-14.63	0.68
RE0027	16.68	0.00	***	0.00	A	A	-11.45	0.60	-20.61	0.98
RE0028	2.94	0.23		0.00	A	A	-10.59	0.53	0.00	0.00
RE0029	6.35	0.04	*	0.00	A	A	-9.28	0.48	-12.72	0.60
RE0030	2.30	0.32		0.00	A	A	-12.50	0.61	0.00	0.00
RE0031	4.89	0.09	.	0.00	A	A	-10.66	0.53	0.00	0.00
RE0032	71.70	0.00	***	0.00	A	A	-7.83	0.35	-226.90	8.50

Nota. *** $p < 0$, ** $p < 0{,}001$, * $p < 0{,}01$, . $p < 0,\ 05$, ‘ ’ $p < 1$. Tamanho do efeito: A: insignificante, B: moderado, C: grande, conforme Zumbo & Thomas (ZT): 0 'A' 0,13 'B' 0,26 'C' 1 e Jodoin & Gierl (JG): 0 'A' 0,035 'B' 0,07 'C' 1.

Segundo a regressão logística, 5 dos 28 itens analisados apresentaram DIF, sendo: RE0009, RE0008, RE0022, RE0027 e RE0032, os quais são estatisticamente significativos ($p <$ 0). Contudo, a magnitude do efeito encontrada é pequena, conforme os parâmetros estabelecidos

por Zumbo e Thomas (ZT) e Jodoin e Gierl (JZ), classificando-os como A. Dessa forma, pode-se considerar que os itens da prova de RE não apresentaram DIF. A Figura 8 demonstra a CCI de maior magnitude encontrado em RE.

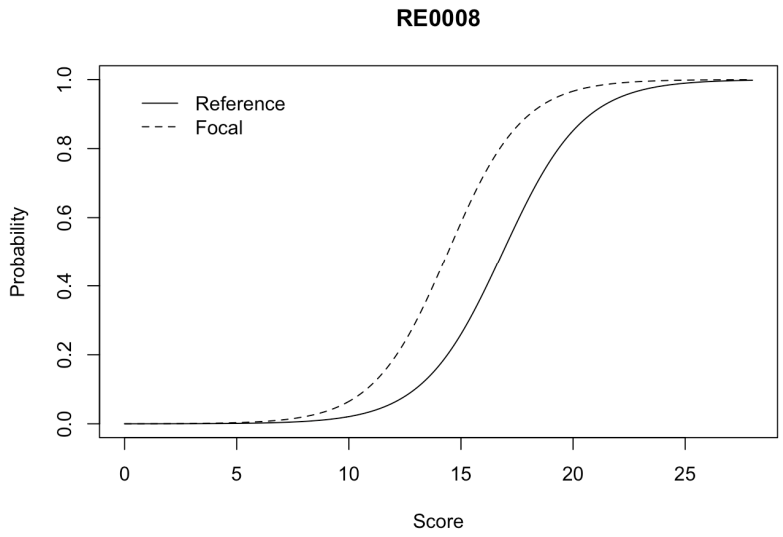

Figura 8. Curva Característica do Item RE0008.
Nota. Linha contínua: grupo de referência (BPR-5); linha tracejada: grupo focal (BPRe).

Nota-se um deslocamento da curva no item RE, o que poderia caracterizar um DIF uniforme. Entretanto, ainda que o item tenha apresentado significância estatística, a magnitude do efeito observada é pequena ($R^2 = 0{,}02$; $p < 0$; ZT = A; JZ = A), sendo a diferença entre os grupos negligenciável. De maneira geral podemos concluir que a maioria dos itens apresentou propriedades psicométricas (dificuldade e discriminação) semelhantes quando transferidos para o meio computadorizado por um algoritmo de avaliação adaptativa. Os itens funcionam da mesma indicando que continuam medindo o mesmo construto que a BPR-5, administrada em lápis e papel.

Em seguida, foi conduzida uma análise fatorial FIFA com os itens das provas da BPRe, visando identificar a unidimensionalidade de cada uma delas. Os resultados da análise da prova de RA são apresentados na Tabela 9.

Tabela 9.
Cargas fatoriais dos itens da prova de RA da BPRe

Item	Carga Fatorial	Item	Carga Fatorial
RA0026	0.75	RA0013	0.48
RA0001	0.75	RA0016	0.48
RA0003	0.76	RA0014	0.60
RA0027	0.78	RA0015	0.53
RA0005	0.76	RA0018	0.21
RA0009	0.71	RA0019	0.40
RA0006	0.70	RA0024	0.48
RA0007	0.64	RA0022	0.52
RA0028	0.45	RA0023	0.20
RA0029	0.59	RA0002	0.79
RA0030	0.58	RA0011	0.42
RA0031	0.48	RA0017	0.06
RA0010	0.64	RA0020	0.57
RA0008	0.56	RA0021	0.44
RA0012	0.56	RA0025	0.44

Nota: Soma das cargas fatoriais ao quadrado: 9,98; Proporção de variância explicada: 0,33; Fator correlacionado: 1.

A análise fatorial apontou um fator explicando 33% da variância, sendo este o raciocínio abstrato (RA). As cargas fatoriais dos itens variaram de 0,06 a 0,79. A maior parte dos itens apresentou cargas acima de 0,30, com exceção dos itens RA0018 (0,21), RA0023 (0,20) e RA0017 (0,06), que obtiveram valores abaixo do recomendado (Pasquali, 2017b).

Da mesma forma, foram verificadas a cargas fatoriais para os itens da prova de RV da BPRe. Os resultados são apresentados na Tabela 10.

Tabela 10.
Cargas fatoriais dos itens da prova de RV da BPRe

Item	Carga Fatorial	Item	Carga Fatorial
RV0001	0.67	RV0017	0.54
RV0002	0.75	RV0019	0.62
RV0026	0.69	RV0020	0.38
RV0006	0.70	RV0021	0.39
RV0007	0.66	RV0022	0.38
RV0027	0.45	RV0023	0.57
RV0009	0.55	RV0032	0.60
RV0028	0.51	RV0005	0.57
RV0029	0.67	RV0010	0.34
RV0030	0.53	RV0011	0.21
RV0031	0.66	RV0012	0.41
RV0014	0.38	RV0013	0.14
RV0008	0.46	RV0015	0.39

Nota: Soma das cargas fatoriais ao quadrado: 8,068; Proporção de variância explicada: 0,288; Fator correlacionado: 1.

Foi apontado um fator explicando 28% da variância, o qual se refere ao raciocínio verbal (RV) com cargas fatoriais dos itens variando entre 0,14 a 0,75. Na prova de RV, dois itens apresentaram valores abaixo do recomendado 0,30 (Pasquali, 2017b), sendo RV0011 (0,21) e RV0013 (0,14).

Na Tabela 11 são apresentadas as cargas fatoriais dos itens da prova de RN.

Tabela 11.
Cargas fatoriais dos itens da prova de RN da BPRe

Item	Carga Fatorial	Item	Carga Fatorial
RN0001	0.59	RN0010	0.67
RN0021	0.76	RN0012	0.77
RN0002	0.72	RN0026	0.72
RN0004	0.76	RN0027	0.70
RN0023	0.84	RN0003	0.83
RN0005	0.72	RN0007	0.80
RN0006	0.75	RN0017	0.50
RN0009	0.75		

Nota: Soma das cargas fatoriais ao quadrado: 8,101; Proporção de variância explicada: 0,54; Fator correlacionado: 1.

A análise apontou um fator explicando 54% da variância e as cargas fatoriais obtidas, variando entre 0,50 a 0,84. Na prova de RN nenhum dos itens apresentou valores abaixo do recomendado, sendo superiores a 0,30 (Pasquali, 2017b). Na prova RN o número de itens com dados suficientes para poder executar a análise foi menor que nas outras provas. A prova RN requer que os alunos digitem os dois números que completam a sequência e por isso é um item de resposta construída e diferentemente das outras provas que são itens de múltipla escolha. Assim a precisão é maior e consequentemente requer um número menor de itens para atingir o critério de encerramento do algoritmo CAT. Consequentemente os itens são selecionados com menor frequência que as demais provas.

Na Tabela 12 são apresentados os resultados análise fatorial exploratória para a prova de RE.

Tabela 12.
Cargas fatoriais dos itens da prova de RE da BPRe

Item	Carga Fatorial	Item	Carga Fatorial
RE0009	0.68	RE0022	0.67
RE0013	0.39	RE0011	0.51
RE0005	0.69	RE0010	0.65
RE0014	0.54	RE0023	0.57
RE0015	0.60	RE0024	0.60
RE0008	0.57	RE0012	0.40
RE0016	0.52	RE0025	0.63
RE0007	0.61	RE0026	0.58
RE0017	0.60	RE0027	0.67
RE0018	0.52	RE0028	0.63
RE0019	0.60	RE0029	0.57
RE0006	0.68	RE0030	0.68
RE0020	0.55	RE0031	0.62
RE0021	0.58	RE0032	0.41

Nota: Soma das cargas fatoriais ao quadrado: 9,857; Proporção de variância explicada: 0,352; Fator correlacionado: 1.

Na prova de RE, a análise apontou um fator explicando 35% da variância e cargas fatoriais dos itens variando entre de 0,39 a 0,68. Todos itens analisados apresentaram cargas fatoriais com valores acima de 0,30 (Pasquali, 2017b). De modo geral, foi possível verificar a unidimensionalidade dos itens de cada uma das provas da BPRe, os quais apresentaram um único fator por prova, de forma a manter uma estrutura interna com estabilidade fatorial.

Considerações finais

O objetivo deste trabalho foi buscar evidências de validade para a Bateria de Provas de Raciocínio – eletrônica (BPRe), bem como verificar o funcionamento diferencial de seus itens e analisar a sua estrutura interna. Para atender tais objetivos, no primeiro estudo realizou-se uma correlação entre os escores das provas da BPR-5 e da BPRe. As análises iniciais demonstram que a BPRe cumpre com a maioria das expectativas em relação a avaliação das capacidades cognitivas, quando comparada a BPR-5.

Ainda que as correlações apresentadas entre os diferentes formatos do teste apresentem magnitudes entre pequenas e grandes, não atendendo as hipóteses estabelecidas, devem ser consideradas as características da amostra. Em função do isolamento social estabelecido em combate ao Covid-19, que coincidiu com o transcorrer da coleta desse estudo, tornou-se inviável a realização de uma coleta de dados mais ampla, com uma amostra mais consistente em razão do fechamento das escolas e consequentemente da impossibilidade de realizar a coleta da versão em lápis e papel. Dessa forma, não é possível considerar que tais resultados forneçam evidências de validade convergente baseada na relação na relação da BPRe com a BPR-5, já que as estimativas de correlação com poucos sujeitos são muito variáveis. Entretanto caminham no sentido correto de correlações positivas.

Um aspecto importante que pode ser observado, ainda que com uma amostra pequena, foi a redução significativa no tempo de execução das provas de raciocínio na BPRe, por meio da testagem adaptativa computadorizada, quando comparadas ao tempo dispensado na realização do teste no formato lápis e papel, o que representa um fator positivo quanto ao seu uso. Ressalta-se ainda que apesar da redução do tempo, em razão da escolha de itens ótimos para cada sujeito a

avaliação CAT tende a ter precisão equivalente a avaliação em tradicional em lápis e papel, fazendo uso de um número menor de itens.

No segundo estudo, foram comparados os níveis de habilidade dos alunos, no qual foram observadas pequenas diferenças quando analisada a distribuição de *theta* das provas de raciocínio nos diferentes formatos do teste. Em geral, a amostra de alunos de ensino fundamental apresentou maior nível de habilidade no formato em lápis e papel. Este resultado pode estar relacionado ao perfil de alunos de ensino fundamental que compõem a amostra, podendo haver divergências nas séries em que se encontram em relação ao grupo que respondeu no computador (em quase sua maioria do 6ª ao 8ª ano). Estudantes de nível médio e superior apresentaram maior nível de habilidade ao realizarem a BPRe na maioria das provas. Quando foram observadas diferenças para estes níveis de ensino, elas foram consideradas pequenas, sendo apresentados níveis de *theta* semelhantes nas provas em lápis e papel e computador. Entretanto há um maior disperssào das notas (desvio padrão). Isso pode ter ocorrido em razão da maior precisão das medidas nesse meio.

Ao verificar o funcionamento diferencial dos itens, apenas na prova de RA foi detectado um item com DIF substancial (RA0024), evidenciando maior dificuldade na sua execução na BPRe, realizada em computador. Dessa forma, faz-se necessário realizar uma nova calibração do item e re-estimação dos escores para nova verificação do seu funcionamento. Para os demais itens que apresentaram significância estatística, a magnitude do efeito encontrada foi pequena, sendo consideradas negligenciáveis.

A análise fatorial FIFA foi conduzida com a finalidade de verificar a unidimensionalidade dos itens na versão eletrônica, examinando-se se formavam uma escala única homogênea. Em cada uma das provas foi observado padrões condizentes com a medida de um único fator. A maior parte dos itens apresentou cargas fatoriais acima de 0,30, com exceção alguns itens da prova de RA (RA0018, RA0023 e RA0017) e de RV (RV0011 e RV0013). Nas provas de RN e

RE, todos os itens apresentaram cargas fatoriais adequadas. Assim, a avaliação da estrutura interna da BPRe apresentou estabilidade fatorial. Sugere-se a realização de uma análise fatorial confirmatória em estudos posteriores considerando-se uma amostra mais adequada.

Uma limitação do presente estudo refere-se a amostra da aplicação adaptativa. Em decorrência dos algoritmos CAT cada pessoa faz uma forma específica do teste, assim a matriz final, considerando o banco de itens, fica com muitos casos de respostas faltantes (matriz esparsa). Isso limita o poder dos estudos DIF e da análise fatorial, pois não há garantia de que o item tenha sido aplicado em toda a amostra. Na verdade, cada item é administrado em uma amostra restrita em termos de amplitude, em pessoas cuja a habilidade seja próxima de sua dificuldade. Uma sugestão seria ter um desenho em que se administrem itens escolhidos de forma aleatória do banco para que possam ser calibrados adequadamente no futuro, após uma coleta mais massiva de dados.

Portanto, os resultados desta pesquisa apresentam dados satisfatórios para a BPRe, ainda de caráter inicial. Destacam-se a importância do desenvolvimento de testes adaptativos computadorizados para a avaliação psicológica, especialmente como ferramenta que pode ser utilizada em avaliações de larga escala, pela brevidade em sua aplicação, imediato feedback dos resultados, bem como sua utilidade em contextos de isolamento social, como o atual cenário. Mediante as limitações apontadas na realização deste estudo, um passo seguinte a ser considerado é a condução de estudos que busquem evidências de validade para a BPRe com amostras mais amplas, que contemplem alunos de instituições públicas e privadas e de diferentes níveis de ensino.

Referências

Alavarse, O. M., Catalani, E. M. T., Menheghetti, D. R., & Travitzki, R. (2018). Teste adaptativo informatizado como recurso tecnológico para alfabetização inicial. *Sistemas, Cibernética e Informática, 15*(3), 68-78. http://www.iiis.org/CDs2017/CD2017Spring/papers/CB368SB.pdf

Ayala, R. J. de. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: Guildford Press.

Almeida, L. S., Guisande, M. A., Primi, R., & Lemos, G. (2008). Contribuciones del factor general y de los factores específicos en la relación entre inteligencia y rendimiento escolar. *European Journal of Education and Psychology, 1*(3), 5-16. https://doi.org/10.30552/ejep.v1i3.13

Almeida, L.S., Lemos, G., Guisande, M.A. & Primi, R. (2008). Inteligência, escolarização e idade: normas por idade ou série escolar?.*Avaliação Psicológica, 7*(2), 117-125. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712008000200002&lng=pt&tlng=pt

Almeida, L. S., & Primi, R. (2000). *Bateria de Provas de Raciocínio*: *Manual Técnico*. são Paulo: Casa do Psicólogo.

Almeida, L. S., & Primi. R. (2004). Perfis de Capacidades Cognitivas na Bateria de Provas de Raciocínio (BPR-5). *Psicologia Escolar e Educacional*, *2*(8), 135-144. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1413-85572004000200002&lng=pt&tlng=pt

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., Torres, C., Haertel, G., Dolan, R., Beddow, P., & Lazarus, S. (2010). Technology-Enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. *Journal of Technology, Learning, and Assessment, 10*(5). https://ejournals.bc.edu/index.php/jtla/article/view/1605

Alves, I. C. B., Rosa, H. R., da Silva, M. A., & Sardinha, L. S. (2016). Avaliação da inteligência: revisão de literatura de 2005 a 2014. *Avaliação Psicológica*, *15*, 89–97. http://www.redalyc.org/articulo.oa?id=335049854010

Ambiel, R. A. M. (2010). Um estudo de caso em Orientação Profissional: Os papéis da avaliação psicológica e da informação profissional. *Revista Brasileira de Orientação Profissional*, *11*(1), 133-143. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1679-33902010000100013&lng=pt&tlng=pt

American Education Research Association [AERA], American Psychology Association [APA], & National Council on Measurement in Education [NCME] (2014). *Standards for Psychology and Educational Testing.* Washington, DC: American Education Research Psychology Association.

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria de Resposta ao Item - Conceitos e Aplicações*. São Paulo, SP: Associação Brasileira de Estatística.

Andrade, J. M., Laros, J. A., & Gouveia, V. V. (2010). O uso da Teoria de Resposta ao Item em avaliações educacionais: diretrizes para pesquisadores. *Avaliação Psicológica, 9*(3), 421-435. https://www.redalyc.org/pdf/3350/335027284009.pdf

Andriola, W. B. (2006). Estudo sobre o viés de itens em testes de rendimento: Uma retrospectiva. *Estudos em Avaliação Educacional*, *17*(35), 115. http://doi.org/10.18222/eae173520062111

Baumgartl, V. O., & Nascimento, E. (2004). A Bateria de Provas de Raciocínio (BPR-5) aplicada a um contexto organizacional. *Psico-USF*, *9*(1), 1-10. https://doi.org/10.1590/S1413-82712004000100002

Baumgartl, V. O., & Primi, R. (2006). Evidências de validade da Bateria de Provas de Raciocínio (BPR-5) para seleção de pessoal. *Psicologia Reflexão e Crítica, 19*(2), 246-251. https://doi.org/10.1590/S0102-79722006000200010

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research, 16*(1), 95–108. https://doi.org/10.1007/s11136-007-9168-6

Brito, M.R.F., Munhoz, A.M.H., Primi, R., Gonçalez, M.H., Rezi, V., Neves, L.F., Sanches, M.H. &Marinheiro, F.B. (2000). Exames nacionais: uma análise do ENEM aplicado à matemática. *Revista Avaliação*, *4*(5), 45-53. http://periodicos.uniso.br/ojs/index.php/avaliacao/article/view/1125

Caemmerer, J. M., Keith, T. Z., & Reynolds, M. R. (2020). *Beyond individual intelligence tests: Application of Cattell-Horn-Carroll Theory. Intelligence, 79, 101433.* https://doi.org/10.1016/j.intell.2020.101433

Campos, C. R., & Nakano, T. C. (2012). Produção científica sobre avaliação de inteligência: O estado da arte. *Interação em Psicologia, 16*(2), 271-282. https://doi.org/10.5380/psi.v16i2.22619

Campos, H. R. (2005). *Análise de Conteúdo e sua Relação com a Dificuldade dos Itens da Bateria de Provas de Raciocínio (BPR-5)* (Dissertação de mestrado). Universidade São Francisco, Programa de Pós-Graduação Stricto Sensu em Psicologia. Itatiba, São Paulo.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press.

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues*. (pp. 122−130). New York: Guilford.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues,* 2nd Ed. (pp. 69−76). New York: Guilford.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Cobêro, C., Primi, R., & Muniz, M. (2006). Inteligência emocional e desempenho no trabalho: um estudo com MSCEIT, BPR-5 e 16PF. *Paidéia (Ribeirão Preto)*, *16*(35), 337-348. https://doi.org/10.1590/S0103-863X2006000300005

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Couto, G. (2007). *Desenvolvimento de Escalas com Interpretação Referenciada nos Itens para a Bateria de Provas de Raciocínio* (Tese de doutorado). Programa de Pós-Graduação em Psicologia, Universidade São Francisco, Itatiba, SP.

Couto, G., & Primi, R. (2011). Teoria de resposta ao item (TRI): Conceitos elementares dos modelos para itens dicotômicos. *Boletim de Psicologia*, *61*(134), 1-15. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S0006-59432011000100002&lng=pt&tlng=pt.

Da Silva-Junior, R. M., Nascimento, A. M., & Roazzi, A. (2019). Bateria de Provas de Raciocínio (BPR-5): Avaliação das qualidades psicométricas em adolescentes do nordeste. *Revista AMAzônica, 23*(1), 264-288. http://periodicos.ufam.edu.br/amazonica/article/view/5174/4135

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press, Inc.

Elói, J., & Candelas, A. A. (2010). Efeitos de um programa de resolução de problemas e de criatividade via e-learning no desempenho da matemática em alunos do 3° ciclo do ensino básico. *International Journal of Developmental and Educational, 2*(1), 95-103. http://www.redalyc.org/articulo.oa?id=349832325009

Embretson, S. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349. http://testolog.narod.ru/Rasch5.pdf

Everson, H., & Osterlind, S. (2009). *Differential Item Functioning*. London: Sage.

Filizatti, R. (2004). *Estudo de validade dos testes 16PF e BPR5 no contexto organizacional* **(**Dissertação de mestrado). Universidade São Francisco, Bragança Paulista.

Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, *5*(8). https://ejournals.bc.edu/index.php/jtla/article/view/1647

Godoy, S., Noronha, A. P. P., Ambiel, R. A. M., Nunes, M. F. O. (2008). Instrumentos de inteligência e interesses em orientação profissional. *Estudos de Psicologia, 13*(1), 75-81. https://doi.org/10.1590/S1413-294X2008000100009

Gottfredson, L. S. (1997). *Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. Intelligence, 24*(1), 13–23. doi:10.1016/s0160-2896(97)90011-8

Guise, Q. G., & Wechsler, S. M. (2018). Avaliação integrada de inteligência e criatividade. *Revista de Psicología (PUCP), 36*(2), 525-548. https://doi.org/10.18800/psico.201802.005

Hambleton, R., & Swaminatham, A. (1985). *Item Response Theory, Principles and Applications*. Boston, MA: Kluwer.

Huang, Y. M., Lin, Y. T., & Cheng, S. C. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers and Education*, *52*(1), 53-67. https://doi.org/10.1016/j.compedu.2008.06.007

Jaloto, A. (2018). É possível reduzir o número de questões do ENEM por meio de uma testagem adaptativa computadorizada? *Revista do Seminário Internacional de Estatística com R, 3*(2), 1-4. http://periodicos.uff.br/anaisdoser/article/view/29248/16960

Jaloto, A. (2021). Funcionamento Diferencial Do Item (DIF) e invariância da medida. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Universidade São Francisco (USF).

Karay, Y., Schauber, S. K., Stosch, C., & Schüttpelz-Brauns, K. (2015). Computer Versus Paper—Does It Make Any Difference in Test Performance?. *Teaching and Learning in Medicine, 27*(1), 57–62. https://doi.org/10.1080/10401334.2014.979175

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *The Journal of Technology, Learningand Assessment*, *4*(2). http://napoleon.bc.edu/ojs/index.php/jtla/article/download/1649/1491

Khoshsima, H., Toroujeni, S. M. H. (2017). Transitioning to an alternative assessment: computer-based testing and key factors related to testing mode. *European Journal of English Language Teaching, 2*(1), 54-73. https://doi.org/10.581/zenodo.268576

Kopec, J. A., Badii, M., McKenna, M., Lima, V. D., Sayre, E. C., & Dvorak, M. (2008). Computerized adaptive testing in back pain - Validation of the CAT-5D-QOL. *Spine*, *33*(12), 1384-1390. https://doi.org/10.1097/BRS.0b013e3181732a3b

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceeding of the Royal Society of Edimburg, 61*, 273-287. https://doi.org/10.1017/S0080454100006282

Lemos, G., Almeida, L. S. A., Guisande, M. A., & Primi, R. (2008). Inteligência e rendimento escolar: análise da sua relação ao longo da escolaridade. *Rev. Port. de Educação, 21*(1), 83-99. http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0871-91872008000100005&lng=pt&tlng=pt

Lemos, G., Almeida, L.S., Guisande, M.A., Primi, R., Martinho, G. & Fortes, I. (2010). Inteligência e rendimento escolar: contingências de relacionamento menos óbvio no final da adolescência. *Revista Galego-Portuguesa de Psicología e Educación*, *18*(1), 1138-1663. https://repositorium.sdum.uminho.pt/bitstream/1822/11598/1/Intelig%C3%AAncia%20e%20Rendimento%20Escolar.pdf

Lima, T. H., Cunha, N. B., & Suehiro, A. C. B. (2019). Produção científica em avaliação psicológica no contexto escolar/educacional. *Psicologia Escolar e Educacional, 23*, 1-9. https://doi.org/10.1590/2175-35392019018897

Linacre, J. M. (2014). *Winsteps® Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com.

Lord, F. (1952). A Theory of Test Scores. *Psychometric Monograph, 7*. https://www.psychometricsociety.org/sites/default/files/pdf/MN07.pdf

Lord, F.M. (1953). An application of confidence intervals of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18,* 57-75. https://doi.org/10.1007/BF02289028

Madeira, M. J.P., Wainer, R., Verdin, R., Alchieri, J. C., & Diehl, E. K. (2002). Geração de estilos cognitivos de aprendizagem de negociadores empresariais para adaptação de ensino tutorializado na web. *Paidéia (Ribeirão Preto)*, *12*(23), 133-147. https://doi.org/10.1590/S0103-863X2002000200010

Magis, D., & Barrada, J. R. (2014). Open-source CAT software: R packages and Concerto. https://orbi.uliege.be/handle/2268/163595

Mansão, C. S. M. (2005). *Interesses profissionais: Validação do Self-Directed Search Career Explorer – SDS* (Tese de doutorado). Pontifícia Universidade Católica de Campinas, São Paulo.

Manseira. P. R. P., & Misaghi, M. (2015). Arquitetura de software para um Sistema de Gestão de Testes Adaptativos Computadorizado. *Produção em foco, 5*(1), 1-25. https://doi.org/10.14521/P2237-5163201500070001

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1-10. https://doi.org/10.1016/j.intell.2008.08.004

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458. https://doi.org/10.1037/0033-2909.114.3.449

Monteiro, J. K., & Andrade, C. G. (2005). Avaliação do raciocínio abstrato, numérico e espacial em adolescentes surdos. *Aletheia*, (21), 93-99.

http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1413-03942005000100009&lng=pt&tlng=pt

Nakano, T. C., Primi, R., Miliani, A. F. M., Martins, A. A., & Abreu, I. C. C. (2015). Bateria para avaliação das altas habilidades/superdotação: análise dos itens via Teoria de Resposta ao Item. *Estudos de Psicologia, 32* (4), 729-741. https://doi.org/10.1590/0103-166X2015000400016

Nakano, T. C., Primi, R., & Nunes, C.H.S.S. (2015). Análise de itens e Teoria de Resposta ao Item. Em C. S. H. Hutz; D. R. Bandeira; C. M. Trentini (Orgs.). *Psicometria* (p. 71-123) Porto Alegre: Artmed.

Nunes, C. H. S. S., & Primi, R. (2009). Teoria de Resposta ao Item: conceitos e aplicações na Psicologia e na Educação. Em C. Hutz (Org), *Avanços e polêmicas em avaliação psicológica* (pp. 25-69). São Paulo: Casa do Psicólogo.

Nunes, C.H.S.S., & Primi, R. (2010). Aspectos técnicos e conceituais da ficha de avaliação dos testes psicológicos. Em CFP, *Avaliação Psicológica: Diretrizes Na Regulamentação Da Profissão* (pp.101–128). Brasília: Conselho Federal de Psicologia.

Nunes, C. H. S. S., Spenassato, D., Oliveira, C. M., Bornia, A. C., & Primi, R. (2015) Testes Adaptativos Computadorizados – CAT. Em C. M. M. Vendramini, D. Bartholomeu, J. M. Montiel, & M. C. R. Silva (Orgs.), *Aplicações de Métodos Estatísticos Avançados à Avaliação Psicológica e Educacional* (pp. 37-76). São Paulo: Vetor Editora.

Nunes, M. F. O., & Noronha, A. P. P. (2009). Relações entre interesses, personalidade e habilidades cognitivas: um estudo com adolescentes. *Psico-USF, 14*(2), 131-141. https://doi.org/10.1590/S1413-82712009000200002

Oliveira, A. M., & Barbosa, A. J. G. (2015). Uma análise bioecológica do baixo desempenho escolar de estudantes com dotação intelectual. *Psicologia Escolar e Educacional*, *19*(3), 585-594. https://doi.org/10.1590/2175-3539/2015/0193910

Oliveira, M. B., & Soares, A.B. (2011). Auto-Eficácia, Raciocínio Verbal e Desempenho Escolar em Estudantes. *Psicologia: Teoria e Pesquisa, 27*(1), 33-39. https://doi.org/10.1590/S0102-37722011000100005

Oz, H., & Ozturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests?. Journal of Language and Linguistic Studies, *14*(1), 67-85. https://search.informit.com.au/documentSummary;dn=865755134110806;res=IELHSS

Parshall, C., Harmes, J. C., Davey, T., & Pashley, P. (2010). Innovative Items for Computerized Testing. Em W. J. van der Linden & C. A. W. Glas (Orgs.), *Elements of Adaptive Testing* (pp. 215-230). Springer New York. https://doi.org/.1007/978-0-387-85461-8_11

Pasquali, L. (2017b). *Psicometria: teoria dos testes na psicologia e na educação.* Editora Vozes Limitada.

Pasquali, L. (2017a). Validade dos testes. *Revista Examen, 1*(1), 14-48. https://examen.emnuvens.com.br/rev/article/view/19/17

Pasquali, L. & Primi, R. (2003). Fundamentos da Teoria de Resposta ao Item - TRI. *Avaliação Psicológica, 2*(2), 99-110. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712003000200002&lng=pt&tlng=pt

Passos, C. S., & Barbosa, A. J. G. (2011). Características de superdotação em um par de gêmeos monozigóticos. *Psico-USF*, *16*(3), 317-326. https://doi.org/10.1590/S1413-82712011000300008

Peres, A. J. S. (2019). Testagem Adaptativa por Computador (CAT): Aspectos conceituais e um panorama da produção brasileira. *Revista Examen, 3*(3), 66-86. https://examen.emnuvens.com.br/rev/article/view/101

Pires, P. Filgueiras, A. Ribas, R., & Santana, C. (2013). Funcionamento Diferencial de Itens (DIF) e experiência de afeto: questões de gênero. *Gerais: Revista Interinstitucional de Psicologia*, *6*(1), 114-126. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1983-82202013000100009&lng=pt&tlng=pt.

Pocinho, M. M. F. D. D. (2010). Psicologia, cognição e sucesso escolar: concepção e validação dum programa de estratégias de aprendizagem. *Psicologia: Reflexão e Crítica*, *23*(2), 362-373. https://doi.org/10.1590/S0102-79722010000200019

Ponczek, V., & Pinto, C. (2016). The Building Blocks of Skill Development. Sao Paulo School of Economics – FGV. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=EEAESEM2017&paper_id=330

Primi, R. (2012). Psicometria: fundamentos matemáticos da teoria clássica dos testes. *Avaliação Psicológica, 11*(2), 297-307. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712012000200015&lng=pt&tlng=pt.

Primi, R. (2013). Psicometria e Tecnologia: Sistema de Testagem Adaptativa Informatizada. São Paulo: Laboratório de Avaliação Psicológica e Educacional (LabAPE).

Primi, R., & Almeida, L. S. (2000). Estudo de validação da bateria de provas de raciocínio: BPR5. *Psicologia: Teoria e Pesquisa, 16*(2), 165173. https://doi.org/10.1590/S0102-37722000000200009

Primi, R., Almeida, L. S., Nakano, T. C., & Campos, C. (2018). *Atualização das normas da Bateria de Provas de Raciocínio (BPR-5) usando a Teoria de Resposta ao Item.* Relatório Técnico não publicado.

Primi, R., Bighetti, C. A., Munhoz, A. H., Noronha, A. P. P., Polydoro, S. A. J., Di Nucci, E. P., & Pelegrini, M. C. K. (2002). Personalidade, interesses e habilidades: um estudo correlacional da BPR-5, LIP e do 16PF. *Avaliação Psicológica*, *1*(1), 61-72. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712002000100007&lng=pt&tlng=

Primi, R., Bueno, J. M. H. & Muniz, M. (2006). Inteligência emocional: validade convergente e discriminante do MSCEIT com a BPR-5 e o 16PF. *Psicologia Ciência e Profissão*, *26* (1), 26-45. https://doi.org/10.1590/S1414-98932006000100004

Primi, R., Carvalho, L. F. D., Miguel, F. K., & Silva, M. C. R. D. (2010). Análise do funcionamento diferencial dos itens do Exame Nacional do Estudante (ENADE) de psicologia de 2006. *Psico-USF*, *26*(3), 379-393. https://doi.org/10.1590/S1413-82712010000300011

Primi, R., Correia, T. A., & Almeida, L. S. (2018). Bateria de Provas de Raciocínio (BPR-5). Em C. H. S. Hutz, D. R. Bandeira, & C. M. Trentini (Orgs.), *Avaliação psicológica da inteligência e da personalidade* (p. 109-122). Porto Alegre: Grupo A Educação.

Primi, R., Couto, G., Almeida, L.S., Guisande, M.A. & Miguel, F.K. (2012). Intelligence, age and schooling: data from the Reasoning Tests Battery (BRT-5). *Psicologia: Reflexão e Crítica*, 25 (1), 79-88. https://doi.org/10.1590/S0102-79722012000100010

Primi, R., Ferrão, M. E., & Almeida, L. S. (2010). Fluid intelligence as a predictor of learning: a longitudinal multilevel approach applied to math. *Learning and Individual Differences, 20*(5), 446-451. https://doi.org/10.1016/j.lindif.2010.05.001

Primi, R., Mansão, C.M., Muniz, M. &Nunes, M.F.O. (2009). *SDS - Questionário de Busca Auto-Dirigida. Manual Técnico da Versão Brasileira.* São Paulo: Casa do Psicólogo.

Primi, R., McGrew, K., Schneider, J., Nakano, T. C., & Dias, N. M. (2017). *Revisão de modelos de inteligência: Como a inteligência é concebida a partir de modelos baseados em evidência?.* São Paulo: Instituto Ayrton Senna.

Primi, R., Muniz, M., & Nunes, C. H. S. S. (2009). Definições Contemporâneas de Validade de Testes Psicológicos. Em C. H. S. Hutz (Org.), *Avanços e polêmicas em avaliação psicológica* (p. 243-265). São Paulo: Casa do Psicólogo.

Primi, R., & Nakano, T. C. (2015). Inteligência. Em F. H. dos Santos, V. M. Andrade, & O. F. A. Boeno (Orgs.), *Neuropsicologia Hoje* (p. 49-58). Porto Alegre: Artmed.

Primi, R., Silva, M. C. R., Santana, P. R., Muniz, M., & Almeida, L. S. (2013). The use of the bi-factor model to test the uni-dimensionality of a battery of reasoning tests. *Psicothema (Oviedo)*, *25*(1), 115-122. https://doi.org/10.7334/psicothema2011.393

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reppold, C. T., & Gurgel, L. G. (2017). Instrumentos psicológicos informatizados. Em M. R. C. Lins, & J. C. Borsa (Orgs.), *Avaliação psicológica: aspectos teóricos e práticos* (p. 77-87). Petrópolis: Vozes.

Richardson, M.W. (1936). The relation between difficulty and the differential validity of a test. *Psychometrika, 1,* 33-49. https://doi.org/10.1007/BF02288003

Santana, L. F., Bartholomeu, D., Montiel, J. M., Couto, G., Berberian, A. A., Pessoto, F. (2017). Avaliação informatiza adaptativa do ENADE pelo MOODLE: evidências de validade.

Informática na Educação: teoria & prática, 20(2), 222-238. https://seer.ufrgs.br/InfEducTeoriaPratica/article/view/69900/43630

Santos, A. A. A., Primi, R., Vendramini, C.M.M., Taxa, F.O.S., Lukjanenko, M.F.S.P., Muller, F., Sampaio I., Andraus Jr, S., Kuse, F.K. & Bueno. C.H. (2000). Habilidades básica de ingressantes universitários. *Revista Avaliação*, *16*(2), 33-45. http://www.scielo.br/pdf/epsic/v7n1/10953.pdf

Sartes, L. M. A. & Souza-Formigoni, M. L. O. (2013). Avanços na Psicometria: Da Teoria Clássica dos Testes à Teoria de Resposta ao Item. *Psicologia: Reflexão e Crítica, 26*(2), 241-250. https://doi.org/10.1590/S0102-79722013000200004

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. Em D. P. Flanagan & E. M. McDanough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4rd ed.) (pp. 73-163). New York: Guilford Press.

Silva, V. R. (2015). *Avaliação da proficiência em inglês acadêmico através de um teste adaptativo informatizado* (Dissertação de mestrado). Universidade de São Paul, Programa De Pós-Graduação em Estatística. Instituto de Ciências Matemáticas e de Computação. São Paulo – SP`. https://teses.usp.br/teses/disponiveis/104/104131/tde-27092017-144727/publico/VanessaRufinodaSilva_revisada.pdf

Sisto, F. F. (2006). O funcionamento diferencial dos itens. *Psico-USF, 11*(1), p. 35-43. http://www.scielo.br/pdf/pusf/v11n1/v11n1a05.pdf

Souza, V. V. (2018). *Construção e evidências de validade de uma bateria brasileira de múltiplas habilidades com base na teoria Cattell–Horn–Carroll* (Dissertação de mestrado). http://repositorio.unb.br/bitstream/10482/32622/1/2018_VictorVasconcelosdeSouza.pdf

Souza, C.V.R., Primi, R. & Miguel, F.K. (2007). Validade do Teste Wartegg: Correlação Com 16PF, BPR-5 e Desempenho Profissional. *Avaliação Psicológica*, *6*(1), 39-49.

http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712007000100006&lng=pt&tlng=pt

Spenassato, D., Bornia, A. C., & Tezza, R. (2015). Computerized adaptive testing: A review of research and technical characteristic. *IEEE Latin America Transactions*, *13*(12), 3890-3898. https://doi.org/10.1109/TLA.2015.7404924

Spenassato, D., Trierweiller, A. C., Andrade, D. F., & Bornia, A. C. (2016). Testes Adaptativos Computadorizados Aplicados em Avaliações Educacionais. *Revista Brasileira de Informática na Educação, 24*(2), 1-12. https://doi.org/10.5753/RBIE.2016.24.02.1

Streiner, D. L. (2010). Measure for measure: New developments in measurement and item response theory. *The Canadian Journal of Psychiatry, 55*(3), 180–186. https://doi.org/10.1177/070674371005500310

Suehiro, A. C. B., Benfica, T. S., & Cardim, N. A. (2015). Avaliação cognitiva infantil nos periódicos científicos brasileiros. *Psicologia: Teoria e Pesquisa, 31*(1), 25-32. https://doi.org/10.1590/0102-37722015011755025032

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://psycnet.apa.org/doi/10.1111/j.1745-3984.1990.tb00754.x

Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education, 97,* 69–85. https://doi.org/10.1016/j.compedu.2016.02.018

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11,* 1-13. https://doi.org/10.1007/BF02288894

Urbina, S. (2007). *Fundamentos da testagem Psicológica.* Porto Alegre: Artmed.

Valentini, F., Franco, V. R., & Iglesias, F. (2017). Introdução à análise de invariância: Influência de variáveis categóricas e intervalares na parametrização dos itens. In B. F. Dmásio, & J.

C. Borsa (Orgs.). *Manual de desenvolvimento de instrumentos psicológicos* (pp. 347–373). Vetor.

Wainer, H. (2000). CATs: Whither and whence. *Psicológica*, 121-133. https://www.uv.es/psicologica/articulos1y2.00/wainer.pdf

Anexos

Anexo 1. Questionário sociodemográfico

Nome:_____

1. Sexo:

( ) Masculino

( ) Feminino

2. Idade: _____

3. Qual o seu nível de escolaridade?

( ) Cursando o ensino superior

( ) Ensino superior completo

4. Você já passou por reprovação durante a vida escolar?

( ) Não

( ) Uma vez

( ) Mais de uma vez

Anexo 2. Termo de Consentimento Livre e Esclarecido (TCLE)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (1ª via)
TÍTULO: Evidências de validade para a Bateria de Provas de Raciocínio – eletrônica

Eu_____, RG_____, abaixo assinado, dou meu consentimento livre e esclarecido para participar como voluntário do projeto de pesquisa supracitado, sob a responsabilidade da pesquisadora Yara da Silva Padilha, aluna do Programa de Pós Graduação – Mestrado em Psicologia da Universidade São Francisco, sob orientação do Prof. Dr. Ricardo Primi. Assinando este termo, estou ciente de que:
1 - O objetivo da pesquisa é buscar evidências de validade para a Bateria de Provas de Raciocínio – eletrônica (BPRe);
2 - Durante o estudo serão aplicados: Questionário Sociodemográfico, subtestes da Bateria de Provas de Raciocínio (BPR-5) e Bateria de Provas de Raciocínio – eletrônica (BPRe). Estima-se que 50 minutos sejam suficientes para responder os instrumentos;
3 - Obtive todas as informações necessárias para poder decidir conscientemente sobre a minha participação na referida pesquisa;
4- As respostas a estes instrumentos não apresentam riscos conhecidos à sua saúde física e mental, mas é possível que causem desconforto emocional;
5 - Estou livre para interromper a qualquer momento minha participação na pesquisa, o que não me causará nenhum prejuízo;
6 – Meus dados pessoais serão mantidos em sigilo e os resultados gerais obtidos na pesquisa serão utilizados apenas para alcançar os objetivos do trabalho, expostos acima, incluída sua publicação na literatura científica especializada;
7 - Poderei contatar o Comitê de Ética em Pesquisa da Universidade São Francisco para apresentar recursos ou reclamações em relação à pesquisa (Av. São Francisco de Assis, 218, sala 35, prédio central - Cidade Universitária CEP: 12916-900, Bragança Paulista – SP, de segunda a sexta, entre 8h e 16h. Telefone: (11)2454-8981/ E-mail: comiteetica@usf.edu.br);
8 - Poderei entrar em contato com o responsável pelo estudo, Yara da Silva Padilha, sempre que julgar necessário pelo *email* yara.padilha@mail.usf.edu.br ou pelo telefone (19) 3236-1314;
9- Este Termo de Consentimento é feito em duas vias, sendo que uma permanecerá em meu poder e outra com o pesquisador responsável.

Local, data: _____, _____

Assinatura do Sujeito de Pesquisa ou Responsável:_____

Assinatura do Pesquisador Responsável: _____

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (via do participante)
TÍTULO: Evidências de validade para a Bateria de Provas de Raciocínio – eletrônica

Eu_____, RG_____, abaixo assinado, dou meu consentimento livre e esclarecido para participar como voluntário do projeto de pesquisa supracitado, sob a responsabilidade da pesquisadora Yara da Silva Padilha, aluna do Programa de Pós Graduação – Mestrado em Psicologia da Universidade São Francisco, sob orientação do Prof. Dr. Ricardo Primi. Assinando este termo, estou ciente de que:
1 - O objetivo da pesquisa é buscar evidências de validade para a Bateria de Provas de Raciocínio – eletrônica (BPRe);
2 - Durante o estudo serão aplicados: Questionário Sociodemográfico, subtestes da Bateria de Provas de Raciocínio (BPR-5) e Bateria de Provas de Raciocínio – eletrônica (BPRe). Estima-se que 50 minutos sejam suficientes para responder os instrumentos;
3 - Obtive todas as informações necessárias para poder decidir conscientemente sobre a minha participação na referida pesquisa;
4- As respostas a estes instrumentos não apresentam riscos conhecidos à sua saúde física e mental, mas é possível que causem desconforto emocional;
5 - Estou livre para interromper a qualquer momento minha participação na pesquisa, o que não me causará nenhum prejuízo;
6 – Meus dados pessoais serão mantidos em sigilo e os resultados gerais obtidos na pesquisa serão utilizados apenas para alcançar os objetivos do trabalho, expostos acima, incluída sua publicação na literatura científica especializada;
7 - Poderei contatar o Comitê de Ética em Pesquisa da Universidade São Francisco para apresentar recursos ou reclamações em relação à pesquisa (Av. São Francisco de Assis, 218, sala 35, prédio central - Cidade Universitária CEP: 12916-900, Bragança Paulista – SP, de segunda a sexta, entre 8h e 16h. Telefone: (11)2454-8981/ E-mail: comiteetica@usf.edu.br);
8 - Poderei entrar em contato com o responsável pelo estudo, Yara da Silva Padilha, sempre que julgar necessário pelo *email* yara.padilha@mail.usf.edu.br ou pelo telefone (19) 3236-1314;
9- Este Termo de Consentimento é feito em duas vias, sendo que uma permanecerá em meu poder e outra com o pesquisador responsável.

Local, data: _____, _____

Assinatura do Sujeito de Pesquisa ou Responsável:_____

Assinatura do Pesquisador Responsável: _____