Universidade de Brasília

Instituto de Psicologia

Departamento de Psicologia Social e do Trabalho

Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

ESTRUTURA FATORIAL E PROPRIEDADES PSICOMÉTRICAS DOS ESCORES OBTIDOS NO SON-R 6-40

Renata Manuelly Feitosa de Lima

Brasília, 2015

Universidade de Brasília

Instituto de Psicologia

Departamento de Psicologia Social e do Trabalho

Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

ESTRUTURA FATORIAL E PROPRIEDADES PSICOMÉTRICAS DOS ESCORES OBTIDOS NO SON-R 6-40

Renata Manuelly Feitosa de Lima

Orientador: Jacob Arie Laros

Dissertação apresentada ao Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações da Universidade de Brasília, como requisito parcial à obtenção do título de Mestre em Psicologia.

Brasília, 2015

Universidade de Brasília

Instituto de Psicologia

Departamento de Psicologia Social e do Trabalho

Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

Banca Examinadora:

_____

Prof. PhD. Jacob Arie Laros (*Orientador*)
Universidade de Brasília - UnB

_____

Profª. Drª. Cristiane Faiad de Moura (*Membro*)
Universidade de Brasília - UnB

_____

Profª. Drª. Cláudia Fukuda (*Membro*)
Universidade Católica de Brasília - UCB

_____

Profª. Drª. Isolda de Araújo Günther (*Suplente*)
Universidade de Brasília - UnB

AGRADECIMENTOS

Gostaria de agradecer a todos que contribuíram com a execução desse trabalho. Porém, as contribuições que recebi foram muitas e é difícil nomear todos que me ajudaram nos últimos anos.

Em primeiro lugar, devo agradecer a Deus por ter permitido a conclusão dessa etapa na minha vida acadêmica.

De modo especial gostaria de agradecer ao meu orientador, Prof. Laros, que me incentivou muuuito no início do mestrado. Quero registrar aqui a profunda admiração que tenho pelo trabalho dele, sua conduta ética e seu comprometimento com a educação e a ciência. Professor, obrigada por todos os seus ensinamentos, por ser tão otimista, alegre, disponível e por construir um ambiente tão agradável, tão amistoso no laboratório.

Agradeço aos membros da banca examinadora, professoras Cristiane Faiad de Moura, Cláudia Fukuda e Isolda de Araújo Günther, que aceitaram gentilmente o convite de avaliar o trabalho. Muito obrigada!

Meu muito, muito obrigada aos amigos do Laboratório Métodos e Técnicas de Avaliação que foram muito generosos comigo e que me ajudaram sempre que precisei: Talita, Gina, Daniel, Luiz, Camila, Fabiana e Alexandre. Obrigada por tudo! Desejo registrar um agradecimento especial ao Felipe Valentini e à Girlene por seus preciosos conselhos e agradecer por todo cuidado ao me receber no Laboratório Meta. Tenham certeza que lembrarei sempre da gentileza, da paciência e da amizade de vocês.

Sou muito grata ao Wlad, Laizza, Annelise e Elisângela pela assistência e auxílio na coleta de dados. Além disto, não poderíamos coletar os dados sem a assistência das escolas. Essas instituições e os profissionais a elas associados foram generosos em termos de tempo e dados. Obrigada também aos alunos que generosamente aceitaram realizar as tarefas propostas tornando possível o trabalho empírico.

Também agradeço a presença constante de minha família, por todo apoio e paciência demonstrados durante essa jornada. Um obrigado mais que especial ao meu companheiro, meu querido esposo Junior que me acompanhou nessa caminhada! Seu

carinho, atenção, compreensão e apoio foram fundamentais para que esse sonho fosse alcançado.

Agradeço ao Instituto de Psicologia da UnB que ofereceu o espaço de reflexão acadêmica para a realização deste trabalho.

Agradeço ao CNPQ que concedeu a bolsa de mestrado e ao Fundo SON de pesquisa pelo apoio financeiro.

Sumário

Lista de Figuras

Manuscrito 2. Resultados preliminares da normatização e validação do SON-R 6-40 para o Brasil

Lista de Tabelas

RESUMO GERAL

A verificação das habilidades cognitivas é provavelmente o procedimento mais empregado no processo de avaliação psicológica de crianças com consequências enormes na escolarização, encaminhamento para programas especiais e tratamento de deficiências específicas. A contribuição dos testes psicológicos neste o processo de avaliação psicológica é fundamental. Levando em consideração as grandes consequências da avaliação de inteligência para a vida das pessoas avaliadas, o nível da qualidade das características psicométricas dos instrumentos psicológicos utilizados nesta avaliação precisa ser alta. Assim, o objetivo dos três manuscritos da presente dissertação foi avaliar as características psicométricas dos testes de inteligência frequentemente usados no Brasil, bem como avaliar as características psicométricas do teste SON-R 6-40, um teste não-verbal de inteligência que está em fase de adaptação para o contexto brasileiro. Os dados sobre as características psicométricas dos testes de inteligência frequentemente usados no Brasil foram obtidos nos manuais dos testes. Os dados em relação às características psicométricas do teste SON-R 6-40 foram baseados no estudo preliminar de normatização e validação do SON-R 6-40 para o contexto brasileiro. Em geral, os resultados sugerem que os instrumentos disponíveis no Brasil destinados à avaliação da inteligência possuem características em sua normatização que não permitem a avaliação de determinados grupos, por exemplo, pessoas analfabetas, com transtornos dentro do espectro do autismo ou com distúrbios de linguagem. Em relação ao SON-R 6-40 os resultados indicam que o instrumento é unidimensional, com qualidade psicométrica e índices de fidedignidade adequados para o uso nas diferentes faixas etárias e com evidências satisfatórias de validade de construto e validade convergente.

Palavras-chave: avaliação de inteligência; características psicométricas; testes de inteligência usados no Brasil.

ABSTRACT

The assessment of cognitive abilities is probably the most used procedure for the psychological diagnostic evaluation of children, with far reaching consequences for schooling, referral to special programs, and treatment of specific handicaps. The contribution of psychological tests in this process of psychological evaluation is fundamental. Taking the considerable consequences of the evaluation of intelligence for the persons in question into account, it seems obvious that a high quality is needed of the psychometric characteristics of the psychological instruments used in this type of evaluation. The general objective of the three manuscripts of the present dissertation was to evaluate the psychometric characteristics of frequently used intelligence tests in Brazil, as well as to evaluate the psychometric characteristics of the SON-R 6-40, a non-verbal test of intelligence which is being adapted for the Brazilian context. The data on the psychometric characteristics of frequently used intelligence tests in Brazil were obtained from the test manuals. The data related to the psychometric characteristics of the SON-R 6-40 test were based on the normatization and validation study of the SON-R 6-40 in Brazil. In general, the results suggest that the available instruments in Brazil designed for the evaluation of intelligence have characteristics that don't allow the evaluation of certain groups, for instance, illiterate people, persons with disorders within the spectrum of autism or with language disturbances. In relation to the SON-R 6-40 the results indicate that the instrument is unidimensional, has high psychometric quality and shows adequate reliability coefficients in the different age groups and satisfactory evidence of the construct and convergent validity.

Keywords: intelligence assessment; psychometric characteristics; frequently used tests in Brazil.

Apresentação

De acordo com o censo de 2010 realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), estima-se que há no Brasil 2.611.536 pessoas que possuem deficiência intelectual/mental. Nesses casos, a avaliação cognitiva é um passo fundamental tanto para diagnosticar quanto para planejar uma intervenção e tomar decisões.

Assim, considerando a pertinência e a contribuição dos testes de inteligência no diagnóstico precoce de deficiências intelectuais e também no campo da seleção profissional, por exemplo, esta pesquisa foi elaborada e organizada em três manuscritos com o objetivo de contribuir com a área da avaliação psicológica no Brasil, mais especificamente na área do desenvolvimento de instrumentos que buscam mensurar a inteligência. A inteligência é um fenômeno complexo e a sua conceituação e modelos sofreram refinamentos e evoluções com o passar dos anos. Para grande parte dos pesquisadores da área, a inteligência está associada à capacidade para aprender relações, utilizando conhecimentos prévios ou apenas o raciocínio.

O manuscrito 1 refere-se à analise dos estudos empíricos descritos nos respectivos manuais de testes de inteligência que são frequentemente utilizados no Brasil. O presente estudo se justifica pela necessidade de avaliar de forma mais precisa os estudos apresentados nos manuais dos testes de inteligência para que haja constante aprimoramento e desenvolvimento desses instrumentos que receberam o parecer favorável do Conselho Federal de Psicologia.

O manuscrito 2 apresenta dados relativos às propriedades psicométricas dos itens e dos escores no SON-R 6-40, um teste não verbal de inteligência para pessoas de 6 a 40 anos de idade. Os dados deste estudo fazem parte da pesquisa de normatização e validação do SON-R 6-40 para o Brasil que está em fase de andamento. De maneira

geral, os resultados embasam o uso do SON-R 6-40 como uma escala de inteligência geral, com qualidade psicométrica e índices de fidedignidade adequados para o uso nas diferentes faixas etárias contempladas.

Por fim, o manuscrito 3 foi organizado com o objetivo de obter evidências de validade convergente dos escores do teste SON-R 6-40 com outro teste de inteligência. Para conseguir isto, o teste foi administrado junto com o WISC-IV em uma amostra de 120 crianças. Os resultados obtidos são muito similares aos resultados encontrados em estudos realizados em outros países e indicam adequada validade convergente dos escores do SON-R 6-40 para a faixa etária investigada.

Os manuscritos serão apresentados a seguir de forma detalhada, descrevendo os procedimentos adotados, seus principais resultados e as limitações de cada estudo, indicando possíveis aprimoramentos em pesquisas futuras.

MANUSCRITO 1

Características psicométricas nos manuais dos testes de inteligência utilizados no Brasil

Título em inglês

Psychometric characteristics reported in manuals of the used intelligence tests in Brazil.

Sugestão de título abreviado:

Características psicométricas nos manuais de testes de inteligência

RESUMO

Para entender melhor as semelhanças e as diferenças nas operacionalizações do construto de inteligência utilizadas nos diferentes testes é necessário avaliar os manuais de forma detalhada. A necessidade de tal avaliação é ainda maior no Brasil com um sistema relativamente novo de avaliação de testes psicológicos. Outra razão de realizar uma avaliação criteriosa dos manuais é a possibilidade de comparar as características dos estudos de normatização e de validação, e obter dados sobre a qualidade desses estudos. Este trabalho apresenta uma análise dos manuais dos testes de inteligência frequentemente utilizados no Brasil com parecer favorável do Conselho Federal de Psicologia, a saber: BPR-5, DFH, Raven, R-1, TONI-3, WAIS-III e WISC-IV. Os resultados indicam que na maioria dos manuais dos testes investigados não foi fornecida informação adequada sobre as fases da construção ou adaptação dos testes, o desenvolvimento das normas e as evidências de validade e fidedignidade.

Palavras-chave: construto de inteligência; estudos de normatização; manuais de teste.

ABSTRACT

To gain a better understanding of the similarities and differences in the operationalizations of the construct of intelligence in different tests, a detailed analysis of the manuals of these tests is necessary. Such an undertaking is even more necessary in Brazil with a relatively new test evaluation system. Another reason to realize a criterious evaluation of test manuals is the possibility to compare the characteristics of the normatization and validation studies and to obtain data on their quality. This study presents an analysis of the manuals of the seven used intelligence tests in Brazil with a positive evaluation of the Federal Council of Psychology: the BPR-5, DFH, Raven, R-1, TONI-3, WAIS-III, and WISC-IV. The results indicate that the majority of the manuals of the investigated tests did not provide sufficient information on the construction or adaptation phase of the test, the development of the norms, and on the evidences of validity and reliability.

Keywords: construct of intelligence; test norms; test manuals

RESUMEN

Para entender mejor las semejanzas y las diferencias en las operacionalizaciónes del constructo de inteligencia que se utiliza en los diferentes instrumentos, es necesario evaluar los manuales de forma detallada. La necesidad de tal tarea es mayor en Brasil, ya que cuenta un sistema de evaluación de instrumentos psicológicos recientemente establecido. Otra razón para realizar una evaluación cuidadosa de los manuales, es la posibilidad de comparar las características dos estudios de normatización y validación, y obtener datos sobre la calidad de las investigaciones. Este trabajo presenta un análisis de los instrumentos de inteligencia frecuentemente utilizados en Brasil que cuentan con la aprobación del Consejo Federal de Psicologia. Estos son: BPR-5, DFH, Raven, R-1, TONI-3, WAIS-III y WISC-IV. Los resultados indican que en el caso de la mayoría de los manuales de los testes investigados, no se ofrece información adecuada sobre las etapas de construcción y adaptación, ni sobre la obtención de normas y de evidencias de validez y confiabilidad.

Palabras clave: el constructo de inteligencia; estudios de normatización; manuales de testes

Alguns testes psicológicos pretendem avaliar aspectos mais gerais como inteligência e personalidade, enquanto outros buscam medir questões mais específicas, como ansiedade em situação de provas. É importante ressaltar que para a utilização adequada de um teste precisa-se de um bom profissional, isto é, um psicólogo competente. Esta competência exige saber selecionar instrumentos e técnicas de avaliação de acordo com os objetivos, público alvo e situação, além de saber quando usar ou não os testes. Nunes et al. (2012) apresentam de forma sucinta uma relação de competências básicas que precisam ser desenvolvidas pelo psicólogo em relação à temática da avaliação psicológica. Unindo uma boa formação acadêmica e a prática, o psicólogo vai acumulando as competências necessárias à realização de boa avaliação, utilizando, entre outras ferramentas, os testes psicológicos (Ambiel, Rabelo, Pacanaro, Alves & Lemes, 2011).

Ao escolher um teste psicológico para auxiliar na avaliação psicológica, é importante que o profissional observe os requisitos mínimos e obrigatórios citados na Resolução 002/2003 do Conselho Federal de Psicologia (CFP), que define e regulamenta o uso, a elaboração e a comercialização de testes psicológicos no Brasil. O CFP avalia e qualifica se os instrumentos são adequados para o uso a partir da constatação de requisitos mínimos. Estes requisitos foram criados a partir da publicação de documentos da *International Test Commission* (ITC), *American Educational Research Association* (AERA), *American Psychological Association* (APA), *National Council on Measurement in Education* (NCME) e *Canadian Psychological Association* (CPA) que desenvolveram diretrizes internacionais relacionados com as exigências técnicas mínimas e o uso correto de testes. Entre os critérios citados na resolução 002/2003, ressalta-se a obrigatoriedade da apresentação de estudos que relatam evidências empíricas de validade e fidedignidade das interpretações propostas para os

escores do teste; apresentação de dados empíricos sobre as propriedades psicométricas dos itens do instrumento; apresentação clara dos procedimentos de aplicação e correção, bem como as condições nas quais o teste deve ser aplicado e descrição das características da amostra de padronização de maneira clara e exaustiva.

A análise das propriedades psicométricas dos itens e das evidências de validade e fidedignidade são fundamentais, pois estas características podem interferir nos resultados de uma avaliação ou pesquisa. Segundo Hogan (2006), a análise dos itens é importante uma vez que a qualidade do item é o alicerce para todas as análises realizadas no nível de escore total e porque a grande maioria dos testes consiste em conjuntos de itens individuais. Assim, é possível controlar as características de um teste por meio do controle dos itens que o compõem.

É possível diferenciar três fases na análise de itens: pré-testagem, análise estatística e seleção de itens. Em relação à análise estatística dos itens, seguindo a Teoria Clássica dos Testes (TCT), os procedimentos tradicionais são o cálculo do índice de dificuldade (a percentagem de acerto) e do índice de discriminação do item (a correlação entre o item e o escore total). Também são utilizados outros índices baseados na Teoria de Resposta ao Item (TRI). A TRI é um conjunto de modelos matemáticos que considera o item como unidade básica de análise e procura representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e do traço latente do indivíduo (Andrade, Laros & Gouveia, 2010; Andrade, Tavares & Valle, 2000). Seguindo a TRI, as análises utilizadas para caracterizar o item são: estimação do parâmetro de dificuldade, do parâmetro de discriminação e do parâmetro de acerto casual (para os itens dicotômicos com múltipla escolha). Os parâmetros estimados têm relação com três modelos teóricos comumente nomeados de modelos de um parâmetro (1PLM), de dois parâmetros (2PLM) e de três

parâmetros (3PLM) (Hogan, 2006). O modelo de um parâmetro leva em conta somente o parâmetro de dificuldade. Já o modelo de dois parâmetros considera tanto a dificuldade quanto a discriminação do item. Por último, o modelo de três parâmetros estima além da dificuldade e da discriminação do item a probabilidade de acerto ao acaso.

Outro critério obrigatório para o CFP é a apresentação de estudos empíricos que revelem evidências de validade e fidedignidade dos escores obtidos nos instrumentos. Os testes devem estar apoiados por evidências de fidedignidade e validade para os grupos para o qual o teste foi construído (International Test Commission [ITC], 2003). A validade é entendida como o grau em que as evidências empíricas alinhadas com uma teoria embasam as inferências e interpretações sobre as características psicológicas das pessoas feitas a partir do comportamento observado (Urbina, 2007). Há diferentes tipos de evidências de validade – evidências baseadas no conteúdo, no processo de resposta, na estrutura interna, baseadas nas relações com variáveis externas e evidências baseadas nas consequências da testagem (AERA, APA, & NCME, 1999). Caso um instrumento não possua evidências de validade, não há garantia de que as interpretações sobre as características psicológicas das pessoas manifestadas pelas suas respostas sejam fundamentadas (Primi, Muniz & Nunes, 2009).

No que se refere às evidências de fidedignidade dos escores de um teste, é necessário constar no manual o método utilizado para estimar a fidedignidade. A fidedignidade refere-se ao grau de precisão, estabilidade dos resultados em diferentes situações (Anastasi & Urbina, 1997). Portanto, todos os testes devem relatar o índice de fidedignidade estimado para que o usuário do teste possa avaliar o grau de precisão dos escores do instrumento em questão. Existem diferentes tipos de fidedignidade, por exemplo, a fidedignidade teste-reteste, a fidedignidade interavaliadores, a fidedignidade

de forma paralela, a fidedignidade baseada na correlação entre as duas metades do teste e a fidedignidade de consistência interna. O último tipo de fidedignidade é o método mais frequentemente utilizado. Todos esses métodos de estimação da fidedignidade fornecem um coeficiente na forma de correlação com valores entre 0 e 1. Urbina (2007) afirma sobre o coeficiente de fidedignidade que estimativas de fidedignidade baixas (menor do que 0,70) sugerem que o escore obtido de um teste pode não ser muito confiável. Segundo a autora a maioria de usuários de testes buscam coeficientes pelo menos da faixa de 0,80 ou mais.

Em relação às características da amostra de normatização, a descrição cuidadosa das características do grupo de referência é indispensável. É de suma importância discutir sobre a representatividade da amostra normativa em relação à população alvo. A determinação da qualidade de um grupo de referência que pretende ser representativo de uma população é uma questão de teoria da amostragem (Richardson, & cols., 1989). A amostragem aleatória é raramente empregada na prática da normatização dos testes. No Brasil as amostras de conveniência escolhidas em base da disponibilidade, são frequentemente utilizadas para a construção das normas. Geralmente, esses grupos provem de uma única localização geográfica, sendo relativamente homogêneos em termos culturais, de faixa etária, de nível de escolaridade e de outras variáveis importantes (Hogan, 2006).

O escore bruto precisa ser convertido em algum tipo de escore normatizado para poder ser interpretado, uma vez que o escore bruto por si só não têm significado. Assim, o escore normatizado situa o escore individual no contexto dos escores obtidos pelos outros examinandos que compuseram o grupo de referência (Hogan, 2006). No que se refere ao tipo de norma, as normas nos testes de inteligência geralmente são calculadas de acordo com a idade dos sujeitos (Almeida, Lemos, Guisande & Primi, 2008). Assim,

o desempenho de uma pessoa é avaliado em comparação com pessoas da mesma idade. A descrição de qual grupo de referência foi utilizado na obtenção dos escores normatizados (baseado na idade, sexo, tipo de escola, etc.) é essencial para a interpretação correta dos resultados do teste. Para cada tipo de norma (percentis, escores padronizados, estaninos) existem pontos fortes e fracos que precisam ser conhecidos e levados em consideração. Para uma discussão mais específica sobre os tipos de normas, sugere-se a leitura dos textos de Anastasi e Urbina (2000), Hogan (2006) e Almeida et al.(2008).

A avaliação psicológica é uma das práticas mais importantes dos psicólogos, pois para que se possa propor qualquer tipo de intervenção em qualquer campo de atuação da Psicologia, faz-se necessário um mínimo de conhecimento sobre os fenômenos e processos psicológicos do objeto de estudo (Conselho Federal de Psicologia [CFP], 2011). A avaliação psicológica é definida por Ambiel et al. (2011) como um processo de construção do conhecimento acerca de questões psicológicas com o objetivo de orientar, sugerir ações e intervenções para a pessoa avaliada. Nesse sentindo, convém ratificar que a avaliação psicológica é um processo técnico e científico praticado com pessoas ou grupo de pessoas, onde se utilizam de diversos métodos, técnicas e instrumentos (CFP, 2011). Entre esses métodos estão os testes psicológicos, aos quais se recorre quando se intenciona avaliar um construto psicológico, por exemplo, inteligência, depressão, ideação suicida, atenção, etc. Em relação à mensuração da inteligência, geralmente ela é realizada com o objetivo de avaliação diagnóstica e os seus resultados têm consequências importantes tanto para a vida escolar como para a formulação de recomendações para a criação de programas especiais de educação e tratamento das desordens (ITC, 2003; Laros, Jesus & Karino, 2013).

De acordo com a pesquisa realizada por Campos e Nakano (2012), os instrumentos tradicionalmente mais utilizados na avaliação da inteligência em pesquisas no período entre 2000 e 2010 no Brasil foram a BPR-5, DFH, MSCEIT, RAVEN, R-1, WAIS-III, WISC-III e Bateria de Habilidades Cognitivas Woodcock-Johnson III. Em um estudo desenvolvido por Alves, Alchieri e Marques (2001) o DFH, Raven, WAIS-III e WISC-III foram destacados como os testes mais ensinados nos cursos de Psicologia. Alguns dos instrumentos utilizados nacionalmente são os mesmos indicados por Flanagan e Harrison (2005) como os mais utilizados no contexto internacional, por exemplo, a Bateria de Habilidades Cognitivas Woodcock-Johnson III, o WAIS-III e o WISC-III. Dos instrumentos supracitados, até o período de elaboração deste artigo, dois ainda não possuem adaptação para o contexto brasileiro: o MSCEIT e a Bateria de Habilidades Cognitivas Woodcock-Johnson III. No entanto, há estudos da Bateria Woodcock-Johnson III no contexto brasileiro (Wechsler & Schelini, 2006; Chiodi & Wechsler, 2012).

Neste artigo buscou-se identificar se os instrumentos utilizados no país, conforme levantamento de Campos e Nakano (2012), apresentavam nos seus manuais os estudos empíricos que a Resolução 002/2003 exige e, especificamente, verificar os estudos desenvolvidos em cada instrumento, no que se refere à validade, à fidedignidade e à normatização do teste. A escolha do manual dos respectivos testes se justifica porque ele deve ser uma das principais fontes de informação sobre o instrumento e sobre a teoria que foi escolhida para a construção deste.

O presente estudo se justifica pela necessidade de avaliar de forma mais precisa os estudos apresentados nos manuais dos testes de inteligência publicados no Brasil para que haja constante aprimoramento e desenvolvimento desses instrumentos que receberam o parecer favorável do CFP. Entende-se que a inclusão do teste psicológico

no rol de testes com parecer favorável não garante uma prática de testagem adequada, mas entende-se como uma medida necessária que auxilia no reconhecimento de instrumentos que atendem critérios mínimos de qualidade (CFP, 2011). Ademais, considera-se importante tal avaliação porque é possível observar os tipos de investigações empregadas, a qualidade dos estudos citados e o reconhecimento das lacunas que demandam estudos mais adequados. Além disso, a necessidade é ainda maior no Brasil porque o sistema de avaliação de testes psicológicos foi recentemente implantado no ano de 2003.

Método

Foram adotados dois critérios para selecionar os testes aqui citados: (1) testes presentes em revisão realizada por Campos e Nakano (2012) sobre os instrumentos mais utilizados no Brasil e (2) testes que tem o parecer favorável do Conselho Federal de Psicologia.

Resultados

A Tabela 1 apresenta os instrumentos analisados e exibe informações sobre os autores, o ano de publicação de cada teste, o tamanho da amostra de normatização e a região onde os dados foram coletados. Essas informações foram retiradas do SATEPSI – Sistema de Avaliação de Testes Psicológicos – serviço mantido pelo CFP e do manual dos respectivos testes.

Tabela 1. Testes selecionados

Teste		Autores	Ano de Publicação	N	Região
BPR-5		Almeida e Primi	2000	1.763	SP e RS
DFH		Sisto	2005	2.750	São Paulo
Raven	Escala Especial	Alves, Duarte, Angelini, Duarte e Custódio	1999	1.547	São Paulo
	Escala Geral	Campos	2001	1.759	-
R1	R-1 Teste não verbal de inteligência	Alves	2002	4.629	SP (900), RJ (363), PR (2.102), ES (495), RN (253)
	R-1- Forma B	Sisto, Santos e Noronha	2004	752	-
TONI-3		Brown, Sherbenou e Johnsen	2006	382	São Paulo
WAIS-III		Nascimento, Silva e Tosi	2004	788	MG
WISC-IV		Castro, Silva, Rueda, Noronha, Sisto e Santos	2011	1.861	SP (650), MG (625), RJ (13), PR (399), SC (48), RS (16), PB (5), RN (104)

Os testes selecionados são descritos abaixo de forma mais detalhada. O levantamento aqui apresentado focou nos estudos de normatização e nos estudos de evidências de validade e precisão dos escores presentes nos manuais de cada teste.

BPR-5: Bateria de Provas de Raciocínio.

De acordo com o manual, o teste auxilia nas atividades relacionadas ao psicodiagnóstico, seleção, orientação profissional e escolar (Almeida & Primi, 2000). A BPR-5 é organizada em duas formas (A e B), com cinco subtestes cada, formando no total 115 itens. A Forma A aplica-se aos estudantes do 7° ao 9° ano do ensino fundamental e a Forma B aos alunos do ensino médio. A aplicação pode ser individual ou coletiva e o tempo total de aplicação, incluindo as instruções, é de cerca de 1 hora e

40 minutos. Há um tempo determinado para a aplicação de cada prova e uma ordem de aplicação dos subtestes.

O estudo de normatização ocorreu nos anos de 1998 e 1999 quando as duas formas foram aplicadas em 1.763 alunos do ensino fundamental e médio residentes em seis cidades do estado de São Paulo e em duas cidades do Rio Grande do Sul. No total, 603 alunos (46,9% do sexo masculino) responderam à Forma A e 1.160 (43,2% do sexo masculino) alunos responderam à Forma B. O manual não informa quantos alunos residiam no estado de São Paulo ou quantos residiam no estado do Rio Grande do Sul. Apesar disso, o manual apresenta informações claras e detalhadas sobre a amostra utilizada no estudo de normatização, declarando informações relacionadas à escolaridade, idade, sexo e variáveis socioeconômicas dos participantes.

No que se refere aos coeficientes de fidedignidade dos escores da BPR-5, o manual apresenta os coeficientes calculados pelo método da consistência interna e pela divisão em duas metades. A consistência interna estimada para os escores de cada subteste variou de 0,70 a 0,91 na Forma A e de 0,80 a 0,88 na Forma B. Utilizando o método das metades, os coeficientes estimados variaram de 0,66 a 0,92 na Forma A e de 0,80 a 0,89 na Forma B.

Com relação à evidência de validade, foi realizada uma análise fatorial de componentes principais dos cinco subtestes e os resultados indicaram a presença de um único fator responsável pela maior parte da variação entre os escores nos cinco subtestes. Além disso, o manual inclui também estudos de correlação entre os resultados da BPR-5 com as notas escolares. Os coeficientes de correlação foram moderados, aumentando à medida que se aproximam o conteúdo das provas e o conteúdo das disciplinas.

DFH: Desenho da Figura Humana – Escala Sisto

O teste do Desenho da Figura Humana – Escala Sisto pretende avaliar o fator *g* de inteligência e o seu público alvo são crianças com idade entre 5 a 10 anos. O teste consiste em um desenho que é realizado pela criança e o sistema de avaliação é composto por 30 itens que quantifica os detalhes que são apresentados na figura ou a ausência de elementos que são considerados importantes. Há escalas diferenciadas de correção para cada sexo: a escala masculina e a escala feminina, porque os autores afirmam que há diferenças entre os sexos na execução do desenho. O teste pode ser aplicado tanto individualmente quanto coletivamente, sem tempo limite para resolução da tarefa proposta. Porém, o tempo médio das aplicações é de 15 minutos para as crianças menores e de 5 a 8 minutos para crianças maiores – o autor não especifica a idade das crianças desses diferentes grupos.

O manual declara que foram investigadas 2.750 crianças na faixa etária de 5 a 10 anos (M = 8,1 e DP = 1,30), sendo que três crianças não informaram o sexo e 239 não registraram a série. Das que forneceram informações, 48,7% eram do sexo masculino, 72,1 % frequentavam escolas públicas e 27,9 % escolas particulares. As escolas onde as crianças foram avaliadas pertenciam a oito diferentes cidades do interior paulista. A aplicação foi coletiva e foi realizada na sala de aula das crianças.

Em relação aos estudos de fidedignidade dos escores, foram realizados estudos utilizando o método das duas metades (N = 2.750), método teste-reteste (N = 390), correlação entre avaliadores com apenas três aplicadores e baseado na consistência interna (N = 2.750). No método das duas metades os coeficientes variaram de 0,74 a 0,81 na escala masculina e 0,71 a 0,80 na escala feminina. No método teste-reteste, o coeficiente apresentou uma variação maior, de 0,69 a 0,90 na escala masculina e 0,64 a

0,90 na escala feminina. Os coeficientes de consistência interna variaram de 0,77 a 0,82 na escala masculina e de 0,74 a 0,83 na escala feminina.

Em relação aos estudos de validade, o manual apresenta três tipos de análises que foram realizadas: (1) correlação entre o escore total do DFH com outros testes de inteligência; (2) análise da evidência da validade interna dos itens; e (3) correlação entre a idade e os escores brutos.

O estudo de evidências de validade convergente entre o DFH- Escala Sisto e o teste das Matrizes Progressivas Coloridas de Raven (CPM) foi realizado com 279 crianças matriculadas nos primeiros anos do ensino fundamental de uma escola pública do interior do estado de São Paulo. As idades variaram entre 7 a 10 anos e 49,1% desta amostra eram do sexo masculino. Os coeficientes de correlação encontrados foram de 0,57 entre o DFH - escala masculina e o Raven total e 0,50 entre o DFH - escala feminina e o Raven total.

Em relação ao estudo da evidência da validade interna dos itens realizada por meio de análises fatoriais e pelo modelo de Rasch, dois estudos são apresentados tanto para a escala masculina quanto para a escala feminina. O primeiro estudo se refere ao ajuste dos itens selecionados para compor as escalas masculina e feminina e o segundo se refere à verificação da unidimensionalidade das escalas.

Na análise da correlação entre aumento da idade e aumento dos escores brutos, o coeficiente de correlação de Pearson encontrado foi de $r = 0,62$ na escala feminina e de $r = 0,64$ na escala masculina. Esses dados informam que há uma tendência de aumento do escore bruto conforme a idade aumenta.

Raven

O teste das Matrizes Progressivas de Raven é formado por três escalas: Geral, Colorida e Avançada. As duas primeiras escalas são utilizadas no Brasil. A escala

Colorida é chamada de Matrizes Progressivas Coloridas de Raven – Escala Especial e a escala Geral é conhecida como Matrizes Progressivas de Raven – Escala Geral. Abaixo as duas versões utilizadas no Brasil são descritas de forma mais específica.

Matrizes Progressivas Coloridas de Raven – Escala Especial

De acordo com o manual, o teste é indicado para avaliação de crianças de cinco anos a onze anos e meio, mas pode ser empregado para deficientes intelectuais e pessoas idosas (Angelini, Alves, Custódio, Duarte & Duarte, 1999). Segundo o manual o teste também é útil para pessoas portadoras de deficiências físicas, afasias, paralisia cerebral e surdez, bem como para aquelas que não dominam a língua nacional. Entretanto, o manual não apresenta estudos com esses grupos.

O teste é composto por apenas um subteste com três séries. Cada série tem 12 itens, totalizando 36 itens. O manual não descreve de forma detalhada e clara a amostra utilizada no estudo de normatização. Em 1987 o teste foi aplicado em 1.417 crianças matriculadas em escolas públicas e particulares. Posteriormente o teste foi aplicado em mais 130 crianças estudantes do 5° e 6° ano. Todos esses dados foram obtidos em 93 escolas localizadas no estado de São Paulo. No total, participaram do estudo 1.547 alunos, sendo 773 crianças do sexo feminino (50%), 715 estudantes de escolas estaduais (46,2%), 487 de escolas municipais (31,5%) e 345 estudantes de escolas particulares (22,3%). Posteriormente, o teste foi aplicado em mais 361 crianças (192 do sexo feminino) provenientes de escolas particulares. A construção das normas para a população geral e para as escolas públicas foram feitas a partir da amostra de 1.547 respondentes. As normas para as escolas particulares levou em conta a amostra com 706 respondentes (345 + 361).

Na pesquisa de normatização a aplicação do teste foi individual até a idade de 7½ anos. Nas demais faixas etárias, o teste foi aplicado em grupo, não excedendo o número de 10 participantes em cada grupo.

Para verificar a fidedignidade dos escores do teste o método das metades foi utilizado. Nas diferentes faixas etárias, o coeficiente de fidedignidade dos escores do teste apresentou grande variação: de 0,59 a 0,93 para o sexo masculino e de 0,41 a 0,94 para o sexo feminino. Esses valores indicam que o teste não é indicado para avaliação nas faixas etárias iniciais, isto é, antes dos sete anos e meio. Outra questão é que não é discutida a razão de estimar a fidedignidade para o grupo masculino e grupo feminino. O manual relata que para os dois sexos reunidos os coeficientes variaram entre 0,52 e 0,93, porém nenhuma correção para a influência da variável idade é citada pelos autores, indicando que o coeficiente pode apresentar superestimação.

Em relação aos estudos que demonstram evidências de validade, o manual apresenta apenas a análise do aumento do escore bruto de acordo com o aumento da idade. Poucos estudos são apresentados no manual e análises essenciais não são discutidas, por exemplo, análise da dimensionalidade do instrumento.

Matrizes Progressivas de Raven – Escala Geral - Séries A, B, C, D e E

O manual afirma que o teste avalia a capacidade que um indivíduo possui para apreender figuras sem significado e descobrir as relações que existem entre elas, imaginar a natureza da figura que completaria o sistema de relações implícito e, ao fazê-lo, desenvolver um método sistemático de raciocínio (Raven, 2008). Segundo o manual o público alvo do teste abrange todas as idades, desde a escola infantil até a idade avançada. Entretanto, o manual não relata estudos com esses grupos específicos.

A escala é formada por 60 itens divididos em cinco séries com 12 itens cada uma. A aplicação pode ser individual ou coletiva.

O estudo de normatização do teste foi realizado em 2002. Participaram 1.759 pessoas, na faixa etária de 13 a 73 anos, dos sexos masculino e feminino, com escolaridade a partir do ensino fundamental incompleto até nível superior completo. O manual não descreve de forma clara o grupo utilizado para o estudo de normatização.

Em relação aos estudos para verificar a fidedignidade dos escores, o manual não apresenta uma descrição clara do método utilizado e da amostra utilizada. Não foi localizado nenhum índice que indica a fidedignidade dos escores obtidos no teste Raven – Escala Geral.

Para a obtenção de evidências de validade dos escores, foram selecionados 351 indivíduos da amostra total que responderam o Raven – Escala Geral e quatro subtestes (Cálculo Numérico, Vocabulário, Percepção de Detalhes e Série de Letras) da bateria BTAG II. Os resultados encontrados para as correlações são descritos a seguir: Cálculo Numérico (r = 0,63), Vocabulário (r = 0,58), Percepção de Detalhes (r = 0,63) e Série de Letras (r = 0,66). O manual, porém, deveria apresentar estudos com testes de inteligência e testes que tem o parecer favorável do Conselho Federal de Psicologia ou testes que são reconhecidos internacionalmente. O manual ainda apresenta uma análise fatorial para indicar a estrutura interna do instrumento.

R-1: Teste não verbal de inteligência

O teste foi criado para o exame psicotécnico de motoristas e pode ser empregado em outras áreas da Psicologia, em especial, na seleção profissional (Alves & Oliveira, 2012). De acordo com o manual, o instrumento tem o objetivo de avaliar a inteligência de adultos e é recomendado para ser usado com pessoas com baixo nível de escolaridade e estrangeiros. O teste é composto por 40 itens, sendo possível realizar a aplicação individual ou coletiva.

Em relação aos estudos relacionados à fidedignidade, o manual relata que os índices foram estimados por meio de dois métodos: o teste-reteste e o das metades. A amostra para o estudo utilizando o método teste-reteste para estimar a fidedignidade foi de 64 adultos, com idade variando entre 18 e 48 anos. O intervalo entre teste e reteste foi de um mês a dezenove meses e a correlação encontrada foi de $r = 0,68$.

A fidedignidade estimada pelo método das metades foi realizada a partir de uma amostra composta de 2.012 sujeitos, a idade variou entre 18 e 65 anos.

R-1 – Forma B: Teste não verbal de inteligência

O teste foi proposto para ser uma forma paralela do R-1: Teste não-verbal de inteligência. Segundo o manual, o teste pode ser empregado no exame psicotécnico de motoristas, bem como em outras áreas que necessitem de um teste alternativo para pessoas analfabetas, com baixa escolaridade ou com dificuldades específicas para a compreensão do português, porém não foi descrito no manual nenhum estudo com esses grupos. O teste pode ser aplicado tanto individualmente quanto coletivamente e o tempo limite de aplicação é de 30 minutos (Sisto, Santos & Noronha, 2004).

A amostra do estudo de normatização foi composta por 752 estudantes de cursos para jovens e adultos, pessoas com defasagem na escolaridade ou em fase de escolarização tardia. Do total de estudantes, 747 forneceram informações sobre o gênero, sendo 50,3% do sexo feminino. Apenas 709 alunos forneceram informações sobre a idade, sendo que a idade mínima foi de 15 anos e a máxima de 76 anos. O manual não apresenta informações suficientes do grupo utilizado no estudo de normatização, por exemplo, faltam informações sobre o estado onde ocorreu a pesquisa, renda e escolaridade dos participantes. A aplicação do instrumento foi coletiva, na maior parte das vezes em grupos de 15 a 20 sujeitos.

Usando o método das metades a fidedignidade estimada dos escores do teste foi de 0,81,e com base na fórmula do alfa de Cronbach a fidedignidade foi de 0,93. Foram calculados também os coeficientes de fidedignidade por faixa etária, sendo que o agrupamento de idade se deu com um intervalo muito grande, por exemplo, faixa etária 1 é formada por sujeitos com 15 a 26 anos, faixa etária 2 é formada por sujeitos com 27 a 37 anos e assim por diante. No método das metades o coeficiente variou de 0,75 a 0,84 e utilizando o alfa de Cronbach, o coeficiente apresentou valores entre 0,90 e 0,92. Porém, é importante ressaltar que não há no manual nenhuma descrição sobre a correção para a influência da variável idade.

Para a análise de evidências de validade convergente, foram realizados dois estudos: o primeiro com o R-1 Forma B e o Teste G-36 e o segundo estudo com o R-1 Forma B e o Teste dos Relógios. No primeiro estudo, os dois testes foram aplicados em 78 estudantes de cursos de Educação de Jovens e Adultos (EJA). A idade mínima foi de 15 e a máxima de 64, sendo 27 (34,6%) no sexo masculino. A correlação encontrada entre o escore total do R-1 Forma B e o escore total do G-36 foi de 0,80. Já no segundo estudo, os dois testes foram aplicados em 68 alunos de cursos EJA, sendo 33 (48,5%) do sexo feminino, com idade mínima variando de 16 anos a 65 anos. A correlação de estimada entre os escores dos dois instrumentos foi de 0,64. O manual também apresenta dados sobre a estrutura fatorial do instrumento e sobre a relação entre o escore bruto e o aumento da idade. Nesse caso, a média dos escores bruto diminuiu conforme aumentou a idade dos grupos etários.

TONI-3 (Forma A): Teste de inteligência não-verbal.

Segundo o manual, o teste é indicado para avaliar a inteligência geral, é de aplicação individual, contém 45 itens e é destinado para crianças de 6 a 10 anos (Brown, Sherbenou & Johnsen, 2006).

O estudo de normatização foi realizado com uma amostra de 382 crianças de 6 a 10 anos, residentes em duas cidades do interior do estado de São Paulo. Faltam informações da amostra utilizada, por exemplo, o tipo de escola que as crianças da amostra normativa frequentavam e nível socioeconômico das crianças.

Para estimar a fidedignidade dos escores foram calculados três índices: alfa de Cronbach (α = 0,83), Spearman-Brown (α = 0,66) e Guttman (α = 0,62), porém não é citado qual índice de Lambda foi utilizado (existem seis índices diferentes de Guttman). É pouco provável que foi utilizado Lambda 2 de Guttman, uma vez que este índice sempre mostra valores maiores do que alfa de Cronbach. Foi utilizado ainda o método teste-reteste para estimar a fidedignidade, com intervalo de 15 dias entre a primeira e última aplicação. Participaram deste estudo 95 crianças com idade entre 6 a 10 anos, (M = 8,13; DP = 1,28) e a correlação encontrada foi de *r* = 0,99. Um valor tão alto para a fidedignidade de teste-reteste (praticamente 1,00) obviamente é uma superestimação do valor real. Os autores não corrigiram a correlação encontrada para a variabilidade da amostra em relação a variável idade. Hogan (2006) tece algumas considerações a respeito das desvantagens do método teste-reteste. Existe sempre uma preocupação quanto ao efeito que o primeiro teste pode exercer no segundo teste. Hogan (2006) afirma que o examinando pode se lembrar das respostas dadas no primeiro teste e responder da mesma maneira no segundo momento da testagem, mesmo que esteja pensando diferente no último momento. Tal fato, diz Hogan, tende a inflacionar o coeficiente de fidedignidade.

Estudos que demonstram evidências de validade do teste com outras medidas são descritos no manual, a saber: Desenho da Figura Humana- Escala Sisto (*N* = 50; *r* = 0,49), Teste Cloze (*N* = 96; *r* = 0,46), Escala de Avaliação da Escrita – Forma A (*N* = 139; *r* = -0,46, era esperado coeficientes negativos) e o instrumento Escala de

Reconhecimento de Palavras (*N* = 136; *r* = 0,31). O manual também descreve análise referente ao crescimento da média do escore bruto conforme o aumento dos grupos etários.

WAIS-III: Escala de inteligência Wechsler para adultos

Esse teste é um instrumento de aplicação individual para avaliação da capacidade intelectual de adultos na faixa etária entre 16 e 89 anos. Contem um total de 14 subtestes e fornece três escores de QI: verbal, de execução e total. A duração da aplicação dos 11 subtestes do WAIS-III que produzem os três escores de QI é de aproximadamente 75 minutos.

O estudo de adaptação e normatização dessa escala para o contexto brasileiro ocorreu entre os anos de 1997 e 2000. A autora esclarece que as normas brasileiras são preliminares porque as normas apresentadas foram derivadas a partir do desempenho de adolescentes e adultos residentes na região metropolitana de Belo Horizonte, Minas Gerais. Participaram do estudo de normatização 788 sujeitos, maiores de 16 anos, sendo que 53,8% da amostra eram do sexo feminino.

Para a estimação do coeficiente de fidedignidade dos escores, foram utilizados dois métodos: consistência interna e teste-reteste. O índice de consistência interna de 11 dos 14 subtestes foi calculado utilizando-se o coeficiente Alfa de Cronbach. O Lambda 2 (λ_2) de Guttman foi utilizado para estimar a fidedignidade dos escores no subteste Armar Objetos.

Os coeficientes de consistência interna foram estimados para cada grupo etário (8) e as médias dos coeficientes dos coeficientes para os subtestes foram obtidas. A média dos coeficientes de consistência interna da maioria dos subtestes oscilou entre 0,82 e 0,92, com exceção do subteste Armar Objetos, λ_2 = 0,66. É importante observar aqui

que os grupos etários ainda consistem de uma variedade relativamente grande em termos de idade.

O método teste-reteste para estimar a fidedignidade foi aplicado em uma parte da amostra total (N = 43), com idades entre 16 e 59 anos. Os participantes foram testados duas vezes, dentro de um intervalo de 2 a 17 semanas. Nesse estudo não houve divisão em faixas etárias. Os coeficientes encontrados para os subtestes Compreensão, Vocabulário e Informação e foram de 0,90, 0,93, 0,95, respectivamente. Já os coeficientes estimados para os subtestes Semelhanças, Aritmética, Completar Figuras, Códigos, Cubos, Raciocínio Matricial e Procurar Símbolos foram 0,89, 0,85, 0,80, 0,85, 0,87, 0,81, 0,89, respectivamente. Os demais subtestes apresentaram os seguintes coeficientes: Dígitos (0,66), Sequência de Números e Letras (0,73) Arranjo de Figuras (0,76) e Armar Objetos (0,65). Cabe alertar que todas estimativas de fidedignidade apresentados acima superestimam a fidedignidade por causa da presença de variância de idade no grupo utilizado. Para obter estimativas corretas as correlações precisam ser corrigidas por causa influência de idade (a fórmula de correção para influência da variável idade: $r_{xx'} = [r_{xx} - r_{xa}^2] \div [1 - r_{xa}^2]$, onde $r_{xx'}$ é a fidedignidade corrigida pela presença de variância de idade, r_{xx} é a fidedignidade não corrigida e r_{xa}^2 é a correlação entre o teste e a variável idade levada ao quadrado (Tellegen & Laros, 2014).

Para a investigação de evidências de validade, foram realizadas várias análises fatoriais e estudos de correlação do WAIS-III com outro teste de inteligência, o teste Matrizes Progressivas de Raven – Escala Geral (N = 53; r = 0,78).

WISC-IV: Escala de Inteligência Wechsler para Crianças – 4ª edição.

De acordo com a pesquisa de Campos e Nakano (2012), o WISC-III é um dos instrumentos mais utilizados na avaliação psicológica. Porém, aqui foi descrita quarta edição da Escala Wechsler de Inteligência para Crianças porque essa é a nova versão

publicada em 2013. O WISC-III sofreu uma série de mudanças, incluindo alterações nos conteúdos dos subtestes, adição de novos subtestes, exclusão de três subtestes e mudanças nos procedimentos de aplicação e pontuação.

O teste foi desenvolvido para avaliar a capacidade intelectual e o processo de resolução de problemas de crianças e adolescentes de 6 anos e 0 meses a 16 anos e 11 meses. É um teste administrado individualmente, com tempo médio de aplicação de 90 minutos.

O manual descreve estudos de fidedignidade dos escores utilizando dois métodos. A correlação estimada entre a correção de quatro avaliadores apresentam valores de 0,88 a 0,99. Por meio do método das metades, os coeficientes de fidedignidade variaram de 0,65 a 0,97.

Em relação a evidências de validade, o manual descreve as análises fatoriais realizadas e análise referente ao aumento do escore bruto conforme o aumento da idade. Correlações com outros testes também foram estimadas. A correlação encontrada entre os subtestes do WISC-IV e o Teste de Cloze: Coisas da Natureza (N = 90) variou de 0,20 a 0,61. Com a Escala de Reconhecimento de Palavras e os subtestes do WISC-IV (N = 69) os coeficientes de correlação apresentaram valores entre 0,39 a 0,75. Com o teste Desenho da Figura Humana, a correlação encontrada entre os subtestes variou de 0,26 a 0,45.

Discussão

O desenvolvimento de um teste psicológico é um processo longo, detalhado e as questões relacionadas ao processo de amostragem, ao estabelecimento de escores normatizados, ao processo de estimar a fidedignidade dos escores e à obtenção de evidências de validade são centrais. Entretanto, nota-se que ainda alguns testes não apresentam de forma detalhada em seus manuais como o instrumento foi construído,

suas características psicométricas e os métodos utilizados para estimá-las e uma descrição minuciosa da amostra de normatização, por exemplo.

Hutz (2009) faz uma discussão pertinente sobre a finalidade de um manual ou o objetivo do manual do teste psicológico. O autor afirma que o manual é uma das principais fontes do teste em questão e sobre a teoria que embasou a construção do instrumento. O mesmo autor ainda afirma que os manuais dos testes brasileiros precisam de constantes atualizações porque a aprovação no SATEPSI é válida por 20 anos. Hutz (2009) sugere que não é razoável utilizar um manual por duas décadas sem revisão. Ele recomenda que os manuais incluíssem adendos com novos estudos realizados, pois as teorias mudam, os métodos de análise ficam mais modernos e também as normas mudam. No caso da mensuração da inteligência, existe a preocupação com o efeito Flynn (Flynn, 2009), que significa um aumento dos escores brutos – cerca de três pontos na escala de QI (100,15) - em dez anos para baterias amplas de inteligência, por exemplo, o WAIS-III. Para testes como os de Raven, o aumento pode chegar a sete pontos em dez anos.

Em relação à estimativa da fidedignidade, o alpha de Cronbach, compreendido ou não, é a medida mais citada nos manuais e periódicos e a mais utilizada pelos construtores dos testes e (Sijtsma, 2009; Maroco & Garcia-Marques, 2006; Ten Berge & Zegers, 1978). Porém existem algumas limitações dessa estimativa de consistência interna que geralmente não são consideradas pelos construtores e usuários dos testes. Ele não é indicado quando os instrumentos contêm poucos itens ou quando a amostra é pequena (Laros, Jesus & Karino, 2013; Sijtsma, 2012; Tellegen & Laros, 2004). Sijtsma (2009) sugere melhores alternativas para estimar a fidedignidade dos testes, tais como, *Lambda 2 de* Guttman e *Greatest Lower Bound* (GLB). O *lambda* 2 de Guttman pode

ser calculado no SPSS e o GLB pode ser estimado utilizando o programa *Factor* (Ten Berge & Kiers, 2003), disponível gratuitamente no site http://www.ppsw.rug.nl/~kiers/.

Ainda em relação à análise da fidedignidade, os manuais precisam descrever de forma detalhada e cuidadosa os coeficientes utilizados nos estudos apresentados. O manual do TONI-3 Forma A, por exemplo, utilizou diferentes índices de consistência interna obtidos em diferentes faixas etárias, mas não cita qual tipo de Lambda adotou, sendo que existem seis diferentes Lambdas de Guttman. E as estimativas de fidedignidade apresentadas nos manuais parecem ser superestimadas porque os autores não citam a correção da influência da variável idade no grupo utilizado. Para obter estimativas corretas as correlações precisam ser corrigidas.

Os *standards* 1999 descrevem duas categorias para análise de estrutura interna de um teste: a consistência interna e a análise fatorial. A consistência interna fornece somente evidências fracas e ambíguas referentes à validade de um teste. O melhor é pensar a fidedignidade como sendo um pré-requisito para a validade (Hogan, 2006). É necessário constar outras evidências que apontem que o construto está sendo mensurado.

De forma geral, os testes analisados apresentam estudos de normatização com amostras específicas e que talvez, não representam o público alvo para o qual o teste é destinado. Apesar disso, poucos manuais apresentaram uma discussão da representatividade da amostra. O manual da BPR-5 apresenta uma discussão sobre a representatividade do grupo normativo, pois os autores do teste afirmam que partir deste grupo irá surgir os parâmetros de comparação. Os autores da BPR-5 relatam ainda que não conseguiram compor uma amostra com representatividade nacional, porém estabeleceram comparações das características socioculturais da amostra com as estimativas nacionais (Almeida & Primi, 2000).

Além de ser representativa do público alvo, a amostra precisa ser descrita cuidadosamente no manual. É essencial informar quem foram os participantes da pesquisa, a faixa etária, o sexo, a escolaridade, o nível socioeconômico, os locais onde os dados foram coletados, o contexto onde o instrumento foi aplicado e qualquer outra variável importante que tem relação com o construto investigado.

Uma alternativa de teste que foi construído utilizando uma amostra com representantes das cinco regiões brasileiras e de diferentes extratos socioeconômico é o teste não-verbal de inteligência SON-R 2½-7[a] (Laros, Jesus & Karino, 2013). O SON-R 2½-7[a] é um teste de inteligência geral para crianças novas, que avalia um espectro largo de habilidades sem envolver o uso da linguagem, podendo ser aplicado também em crianças com problemas auditivos e de linguagem. Ele é uma versão abreviada do SON-R 2½-7, de origem holandesa, que possui estudos de normatização e evidências de validade em alguns países europeus. No Brasil, a amostra de normatização ficou composta por 1.200 crianças, divididas equitativamente quanto à idade e o sexo. A pesquisa foi realizada em 13 estados diferentes, contemplando 36 cidades.

Considerações Finais

O presente estudo teve como objetivo principal analisar os estudos presentes nos manuais de testes psicológicos comumente usados e verificar a qualidade dos estudos desenvolvidos a partir das informações apresentadas nos manuais dos testes. Ele também tentou fornecer uma base para o conhecimento dos princípios fundamentais ao selecionar um teste psicológico. Mas, esta pesquisa não busca encerrar-se em si mesmo e sua brevidade não permite a apresentação completa das questões relacionadas à construção e seleção de testes psicológicos.

A construção e o desenvolvimento de um teste é um processo longo e é importante que as etapas realizadas sejam bem descritas e explicadas nos respectivos

manuais. Primi & Nunes (2010) discutem sobre o nível de exigência atual para aprovação do teste psicológico. Os autores consideram o nível muito baixo e afirmam que os requisitos mínimos declarados na resolução 002/2003 é apenas um grupo pequeno de informações que um manual precisa incluir. Eles indicam que a comissão consultiva do CFP vem discutindo pontos para o aprimoramento das exigências e que duas opções têm sido mais ressaltadas no debate: o aumento dos requisitos mínimos e a elaboração de recomendações. A última proposta tem sido mais acolhida e como consequência houve uma alteração na ficha de avaliação dos instrumentos com o objetivo de caracterizar o teste e elaborar recomendações. Além disso, de acordo com os autores, a comissão citada prepara uma reavaliação dos manuais buscando se preparar para oferecer recomendações em função dos estudos considerando quatro aspectos dos instrumentos: construto, propósito, contexto e validade.

Com as questões apresentada aqui, pode-se concluir a importância do aprimoramento dos manuais dos testes de inteligência. Todos os cuidados e sugestões indicados são propostas para que se reduza a crítica aos testes de inteligência (Flores-Mendoza, Nascimento & Castilho, 2002) e para induzir um aumento na qualidade da prática profissional. Além disso, a proposta do trabalho foi fornecer uma contribuição inicial, porém existem ainda várias possibilidades de investigações acerca do tema tratado.

Referências

Almeida, L. S., & Primi, R. (2000). *BPR-5: Bateria de provas de raciocínio: manual técnico.* São Paulo: Casa do Psicólogo.

Alves, I. C. B., Alchieri, J. C., & Marques, K. (2001). Panorama geral do ensino das técnicas de exame psicológico no Brasil. Em *I Congresso de Psicologia Clínica – Programas e Resumos* (pp. 10-11), Universidade Presbiteriana Mackenzie, São Paulo.

Alves, I. C. B., & Oliveira, R. (2012). *R-1: Teste não verbal de inteligência: manual técnico*. São Paulo: Vetor Editora.

Ambiel, R. A. M., Rabelo, I. S., Pacanaro, S. V., Alves, G. A. S., & Leme, I. F. A. S. (2011). *Avaliação psicológica: guia de consulta para estudantes e profissionais de psicologia.* São Paulo: Casa do Psicólogo.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Anastasi, A., & Urbina, S. (2000). *Testagem psicológica*. Porto Alegre: Artes Médicas Sul LTDA.

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,* 421-435.

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria da resposta ao item: conceitos e aplicações.* São Paulo: ABE – Associação Brasileira de Estatística.

Angelini, A. L., Alves, I. C. B., Custódio, E. M., Duarte, W. F., & Duarte, J. L. M. (1999). *Matrizes progressivas coloridas de Raven – escala especial. Manual técnico.* São Paulo: CETEPP - Centro Editor de Testes e Pesquisas em Psicologia.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (2006). *TONI-3 (forma A). Teste de inteligência não verbal. Manual do examinador.* São Paulo: Vetor Editora.

Campos, C. R., & Nakano, T. C. (2012). Produção científica sobre avaliação da inteligência: o estado da arte. *Interação psicológica, 16*, 271-282.

Chiodi, M. G., & Wechsler, S. M. (2012). Estudo de validade convergente da bateria de habilidades cognitivas Woodcock-Johnson-III – versão ampliada. *Avaliação Psicológica, 11*, 63-75.

Conselho Federal de Psicologia (2003). *Resolução 002/2003*. Retrieved from http://site.cfp.org.br/resolucoes/resolucao-n-2-2003/.

Conselho Federal de Psicologia (2011). *Ano da avaliação psicológica. Textos geradores.* Brasília, DF: CFP.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-307.

Flores-Mendoza, C. E., Nascimento, E., & Castilho, A. V. (2002). A crítica desinformada aos testes de inteligência. *Revista estudos de psicologia, PUC-Campinas, 19*, 17-36.

Flynn, J. R. (2009). *O que é inteligência? Além do efeito Flynn*. Porto Alegre: Artmed.

Hogan, T. P. (2006). *Introdução à prática de testes psicológicos.* Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A.

Hutz, C. S. (2009). *Avanços e polêmicas em avaliação psicológica.* São Paulo: Casa do Psicólogo

Hutz, C. S., & Bandeira, D. (1993). Tendências contemporâneas no uso de testes, uma análise da literatura brasileira e internacional. *Psicologia: Reflexão e Crítica, 6*, 85-101.

International Test Commission (2003). *Diretrizes para o uso de testes: International Test Commission (ITC).* . Retrieved from http://www.ibapnet.org.br/ diretrizesITC.pdf

Laros, J. A., Jesus, G. R., & Karino, C. A. (2013). Validação brasileira do teste não verbal de inteligência SON-R 2½ - 7 [a]. *Avaliação Psicológica,12* ,233-242.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Maroco, J., & Garcia-Marques, T. (2006). Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? *Laboratório de Psicologia, 4,* 65-90.

Nunes, M. F. O., Muniz, M., Reppold, C. T., Faiad, C., Bueno, J. M. H., & Noronha, A. P. P. (2012). Diretrizes para o ensino de avaliação psicológica. *Avaliação Psicológica, 11,* 309-316.

Pasquali, L. (2010). *Instrumentação psicológica. Fundamentos e práticas.* Porto Alegre: Artmed.

Primi, R., & Nunes, C. H. S. (2010). O Satepsi: desafios e propostas de aprimoramento, In *Conselho Federal de Psicologia. Avaliação psicológica: diretrizes na regulamentação da profissão.* Brasília, DF: Conselho Federal de Psicologia

Primi, R., Muniz, M., & Nunes, C. H. S. S. (2009). Definições contemporâneas de validade de testes psicológicos. In C. S. Hutz (Ed.), *Avanços e polêmicas em avaliação psicológica* (pp. 243-265). São Paulo: Casa do Psicólogo.

Raven, J. C. (2008). *Teste das matrizes progressivas – escala geral séries A, B, C, D e E. Manual técnico.* Rio de Janeiro: CEPA – Centro Editor de Psicologia Aplicada Ltda.

Richardson, R. J., & Cols. (1989). *Pesquisa social. Métodos e técnicas.* São Paulo: Atlas.

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika,77,* 4-20.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika,74,*107-120.

Sisto, F. F., Santos, A. A. A., & Noronha, A. P. P. (2004). *R-1 – Forma B – Teste não-verbal de Inteligência. Manual técnico.* São Paulo: Vetor.

Tellegen, P. J., & Laros, J. A. (2004). Cultural bias in the SON-R test: Comparative study of Brazilian and Dutch children. *Psicologia: Teoria e Pesquisa, 20,* 103-111.

Tellegen, P. J., & Laros J. A. (2014). SON-R 6-40. *Snijders-Oomen Non-verbal intelligence test. Volume I: Research report.* Hogrefe: Göttingen, Germany.

Ten Berge, J. M. F., & Kiers, H. A. L. (2003). *The minimum rank factor analysis program MRFA* (internal report). Department of Psychology, University of Groningen, The Netherlands.

Ten Berge, J. M. F., & Zegers, F. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43,*575-579.

Thompson, B. (2003). *Score reliability: Contemporary thinking on reliability issues.* Thousand Oaks, CA: Sage Publications.

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

Wechsler, D. (2011). *WAIS-III – Escala de inteligência Wechsler para adultos. Manual técnico.* São Paulo: Casa do Psicólogo.

Wechsler, D. (2013). *WISC-IV - Escala Wechsler de inteligência para crianças. Manual técnico.* São Paulo: Casa do Psicólogo.

Wechsler, S. M., & Schelini, P. W. (2006). Bateria de habilidades cognitivas Woodcock-Johnson-III: Validade de construto. *Psicologia: Teoria e Pesquisa, 22,* 287-296.

MANUSCRITO 2

Resultados preliminares da normatização e validação

do SON-R 6-40 para o Brasil

Título em inglês

Preliminary results of the normatization and validation

of the SON-R 6-40 for Brasil

Sugestão de Título Abreviado

Psychometric properties of the SON-R 6-40

RESUMO

Este artigo apresenta e discute os dados relativos às propriedades psicométricas dos itens e dos escores no SON-R 6-40, um teste não verbal de inteligência para pessoas de 6 a 40 anos de idade. Os dados do atual estudo fazem parte da pesquisa de normatização e validação do SON-R 6-40 para o Brasil que está em fase de andamento. Participaram 711 pessoas residentes nas regiões Nordeste, Centro-Oeste, Sudeste e Sul. A fidedignidade dos escores no SON-R 6-40 foi analisada por faixa estaria. Os parâmetros dos itens foram analisados usando a Teoria Clássica dos Testes e a Teoria de Resposta ao Item. Análise fatorial evidenciou um único fator e os itens apresentaram características psicométricas adequadas. De maneira geral, os resultados embasam o uso do SON-R 6-40 como uma escala de inteligência geral, com alta qualidade psicométrica e com índices de fidedignidade adequados para o uso nas diferentes faixas etárias contempladas.

Palavras-chave: pesquisa de normatização, estrutura fatorial, análise de itens; SON-R 6-40.

ABSTRACT

This article presents and discusses psychometric properties of test scores and of the items of the SON-R 6-40, a non-verbal test of intelligence for persons between 6 and 40 years of age. The data of the current study are a part of the normatization and validation research of SON-R 6-40 in Brazil which is in progress. So far, 711 persons from the Northeast, Center-west, Southeast and South of Brazil participated. The reliability of the scores on the SON-R 6-40 was analyzed per age group. The parameters of the items were analyzed using Classic Test Theory and Item Response Theory. Factor analysis evidenced a single factor and the items presented appropriate psychometric properties. Overall, the results justify the use of the SON-R 6-40 as a scale of general intelligence, with high psychometric quality and with appropriate reliability coefficients of the test scores for the different age groups.

***Keywords*:** normatization study, factor structure, item analysis, SON-R 6-40.

RESUMEN

Este artículo presenta y discute datos relativos a las propiedades psicométricas de los ítems y de los puntajes del SON-R 6-40, un test no verbal de inteligencia para personas de 6 a 40 años de edad. Los datos del actual estudio son parte de la investigación de normatización y validez del SON-R 6-40 para Brasil, que se realiza actualmente. Participaron 711 personas residentes de las regiones del Nordeste, Centro-Oeste, Sudeste y Sur. La confiabilidad de los puntajes del SON-R 6-40 fue analizada por grupo etario. Los parámetros de los ítems fueron analizados usando la Teoría Clásica de los Testes y la Teoría de Respuesta al ítem. El Análisis Factorial evidenció un único factor y que los ítems presentaban características psicométricas adecuadas. De manera general, los resultados sugieren el uso del SON-R 6-40 como una escala de inteligencia general, con una alta calidad psicométrica y con índices de confiabilidad adecuados para su uso en los diferentes grupos etarios que comprende.

Palabras clave: investigación de normatización ,estructura factorial, análisis de ítems, SON-R 6-40.

A história da psicologia indica que já existiam precursores da testagem psicológica em diferentes contextos antes mesmo do século XX. Porém, foi no início do século XX, mais precisamente em 1905, que foi publicado o primeiro teste de inteligência, mais conhecido como a Escala de Inteligência Binet-Simon. O psicólogo francês Alfred Binet foi chamado para criar um método para avaliar crianças que, devido a atrasos no desenvolvimento, não conseguiam se beneficiar das classes regulares do sistema educacional público francês e que necessitavam de educação especial. Assim, juntamente com Theodore Simon, Binet propôs um conjunto de testes que tinham o objetivo de avaliar o julgamento e a capacidade de raciocínio (Urbina, 2014). Após alguns anos, o teste foi revisado e traduzido para outros idiomas e, apesar da sua ampla divulgação e utilização, em 1940 Cattell criticou duramente os autores desta escala por haverem construído uma medida excessivamente verbal e dependente da escolaridade dos indivíduos. Ainda hoje, os testes tradicionais de inteligência tem sido alvo de críticas e, consequentemente, de revisões, por utilizarem habilidades de linguagem específicas, tanto nos conteúdos quanto nas instruções, o que colocariam os membros de minorias culturais ou pessoas com problemas de linguagem e auditivos em desvantagem (Laros, Jesus & Karino, 2013; Schelini, 2006).

Passados mais de cem anos da publicação do primeiro teste de inteligência, a avaliação deste construto continua a ser feita pelos psicólogos nos seus contextos profissionais por meio de testes. A avaliação cognitiva é um passo fundamental tanto para diagnosticar quanto para planejar uma intervenção e tomar decisões.

Segundo relatório do Fundo das Nações Unidas para a Infância (Unicef) publicado em 2013 sobre a situação mundial da infância, 93 milhões de crianças ou uma em cada 20 crianças com 14 anos de idade ou menos vivem com algum tipo de deficiência moderada ou grave (UNICEF, 2013). Segundo dados do censo de 2010

realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), no Brasil há 45.606.048 pessoas com pelos menos uma das deficiências investigadas pelo IBGE (deficiência visual, auditiva, motora e intelectual). Desse total, 2.611.536 pessoas possuem deficiência intelectual/mental. Além disso, ainda há a necessidade de avaliar indivíduos e suas aptidões intelectuais no momento da orientação profissional e seleção no mundo do trabalho (Pasquali, 2010). Segundo Baugartl e Primi (2006), estudos de revisão da literatura sobre os processos seletivos apontam as medidas cognitivas como as que apresentam maior capacidade preditiva do desempenho profissional seguidas de medidas de integridade e entrevistas estruturadas. Por meio desses dados, é possível observar a pertinência dos testes de inteligência e o oferecimento de maiores garantias ou evidências de validade às decisões quando eles são usados.

De modo semelhante ao surgimento do primeiro teste de inteligência publicado, o teste SON surgiu na Holanda a partir de uma necessidade sentida pela autora, Nan Snijders-Oomen, de mensurar o potencial de aprendizagem de crianças com problemas no desenvolvimento da linguagem (Tellegen & Laros, 2014). Naquele momento, na década de 1940, os testes disponíveis eram dependentes das habilidades verbais, tornando-os assim inadequados para a população de crianças surdas. Em vista disso, a autora consciente de que é impossível diagnosticar ou investigar de forma adequada um problema em questão sem bons instrumentos de medida, criou a primeira versão do teste SON para crianças surdas com idades compreendidas entre os 4 e 14 anos de idade.

Na primeira revisão do teste em 1958, o limite superior da faixa etária do teste foi expandido para 16 anos e normas foram estabelecidas tanto para crianças surdas como para crianças ouvintes (Snijders & Snijders-Oomen, 1958). Em 1975 a segunda revisão foi efetuada e duas versões do teste foram desenvolvidas para atender crianças e jovens

com idades distintas. Após outras versões que surgiram do teste, no final da década de 80, os autores Snijders, Tellegen e Laros, conhecedores das mudanças e dos constantes desenvolvimentos no campo da inteligência, propuseram o SON-R 5½-17 (Snijders, Tellegen & Laros, 1988). Nesta versão, o teste era destinado a crianças surdas e crianças ouvintes com idades entre 5½ à 17 anos e os esforços dos autores se concentraram em reunir as vantagens das versões anteriores dos testes SON.

Após duas décadas desde a publicação da versão SON-R 5½-17, várias razões levaram os autores do teste revisar esta versão, por exemplo: necessidade de atualização das normas, necessidade de modernizar o material do teste, necessidade de torná-lo mais adequado para avaliação de adultos e diminuição do tempo de administração, mas garantindo as características psicométricas e a qualidade dos subtestes (Tellegen & Laros, 2014). Assim, em 2011, foi publicado o SON-R 6-40 (Tellegen & Laros, 2011), sobre o qual versa o presente trabalho. O desenvolvimento do teste SON-R 6-40 seguiu diferentes fases de estudo para atender as novas demandas na avaliação da inteligência, tais como, realização de diversos estudos em diferentes países entre 2003 e 2009 com o objetivo de melhorar o conteúdo dos subtestes (Laros & Tellegen, 2004), estudos com crianças com necessidades especiais e pesquisa de normatização com quase duas mil pessoas na Holanda e Alemanha. Aqui no Brasil, a pesquisa de normatização está em andamento e este artigo apresentará análises parciais do estudo de normatização. A previsão é que no final do estudo de normatização o teste seja aplicado em uma amostra de 1.360 crianças, adolescentes e adultos, provenientes de todas as regiões brasileiras com idade entre 6 e 40 anos.

O teste SON-R 6-40 é um teste de inteligência destinado à avaliação de crianças e adultos com idade entre 6 a 40 anos. O teste é designado para avaliar um espectro das habilidades cognitivas sem a utilização da linguagem falada ou escrita. Os subtestes do

instrumento avaliam raciocínio abstrato e concreto, habilidade espacial e percepção visual. A presente pesquisa teve como objetivo avaliar as características psicométricas do instrumento SON-R 6-40. Mais especificamente, foi realizado um estudo exploratório da dimensionalidade do teste, assim como análise da consistência interna dos quatro subtestes e da escala geral, análise do crescimento dos escores brutos ao longo das faixas etárias estudadas e, por último, estimação dos parâmetros dos itens utilizando a Teoria de Resposta ao Item (TRI).

A inteligência é um fenômeno complexo e a sua conceituação e modelos sofreram alterações, refinamentos e evoluções com o passar dos anos. Para grande parte dos pesquisadores da área, a inteligência está associada à capacidade para aprender relações, utilizando conhecimentos prévios ou apenas o raciocínio (Almeida, 1994). Dentro do enfoque diferencial, de onde surgiram as teorias psicométricas, modelos teóricos como o de Spearman de um fator geral (Spearman, 1904), o de Thurstone das capacidades mentais primárias (Thurstone, 1938), o modelo de Cattell de inteligência fluida (*Gf*) e inteligência cristalizada (*Gc*) (Cattell, 1963) e o modelo de Carroll dos Três Estratos (Carroll, 1993) foram propostos tentando modelar a estrutura da inteligência. Hoje, o modelo que tem sido amplamente reconhecido e utilizado é o modelo Cattell-Horn-Carroll (CHC), que foi proposto por McGrew e Flanagan (1998). Esse modelo é organizado numa estrutura fatorial hierárquica de três níveis, assim como o modelo dos Três Estratos de Carroll, e segue uma ordem de especialização, do nível mais geral (Estrato III) até os fatores específicos do Estrato I (Flanagan & Harrison, 2012; Seabra, Laros, Macedo & Abreu, 2014).

O modelo CHC é uma síntese dos modelos psicométricos anteriormente propostos, reunindo as teorias de Cattell, Horn e Carroll, com algumas diferenças e aperfeiçoamentos, por exemplo, a importância e a interpretação do fator geral, que pode

ser compreendida como uma capacidade cognitiva geral ou a soma das habilidades específicas (Seabra, Laros, Macedo & Abreu, 2014; Flanagan & Harrison, 2012). No modelo CHC, o fator *g* influencia diretamente apenas os fatores do Estrato II e indiretamente as habilidades específicas localizadas no Estrato I (Seabra, Laros, Macedo & Abreu, 2014).

Cattell (1963) e Horn (Horn &Noll, 1997) diferenciaram a inteligência fluida (*Gf*) de inteligência cristalizada (*Gc*). A primeira envolve as habilidades de raciocínio, capacidade para resolução de problemas novos, para os quais a pessoa tem pouco conhecimento prévio, capacidade de perceber relações entre padrões de estímulo, compreender implicações e tirar conclusões das relações (Seabra, Laros, Macedo & Abreu, 2014; Carroll, 2005; McGrew, 2005). No modelo CHC, Schneider e McGrew (2012) descrevem a inteligência fluida como a habilidade que é utilizada quando os esquemas, os hábitos e os conhecimentos adquiridos falham na elaboração de uma solução para um problema novo. A *Gf* é composta pelas habilidades específicas (estrato I) de Indução, Raciocínio Sequencial Geral e Raciocínio Quantitativo (Seabra, Laros, Macedo & Abreu, 2014). Além disso, a *Gf* está associada a componentes não-verbais e é pouco dependente da influência de aspectos culturais (Schelini, 2006).

A inteligência cristalizada refere-se à aquisição e à solidificação de conhecimentos formais e informais, aprendidos por transmissão cultural ou pela escola (Seabra, Laros, Macedo & Abreu, 2014; Cattell, 1963). Esta habilidade cognitiva seria desenvolvida a partir das experiências culturais e educacionais, estando presente na grande parte das atividades escolares (Schelini, 2006). No modelo de Cattell , a *Gc* era uma dimensão mais ampla. Já no modelo CHC, a habilidade *Gc* foi subdividida em outras habilidades, sendo composta por: informação verbal geral, desenvolvimento da linguagem, conhecimento lexical, habilidade de escuta (compreensão de um discurso),

habilidade de comunicação e sensibilidade gramatical (Seabra, Laros, Macedo & Abreu, 2014; Schneider & McGrew, 2012). Estudos indicam que a *Gc* tende a alargar com o aumento da idade porque ela está relacionada às experiências culturais, ao contrário da inteligência fluida que parece declinar após os 21 anos de idade (Schelini, 2006; Horn & Noll, 1997).

No modelo CHC, a *Gf* está localizada no estrato II e é a dimensão mais próxima ao fator *g*, localizada no estrato III (McGrew, 2009). Em outras palavras isso significa que a inteligência fluida é a habilidade mais importante na previsão da capacidade geral de adaptação à situações novas, que demandam autonomia intelectual (Laros, Jesus & Karino, 2013).

Antonio, Mecca e Macedo (2012) afirmam que os instrumentos disponíveis atualmente no Brasil possuem características em sua padronização que limitam a avaliação de determinados grupos clínicos, por exemplo, pessoas com transtornos dentro do espectro do autismo, com deficiências sensoriais, distúrbios de linguagem, etc. Roid e Miller (1997) afirmam que os testes tradicionais de inteligência exigem habilidades e formas de responder que determinados grupos não desenvolveram de forma adequada e, consequentemente, não são indicados para avaliação, pois sua aplicação se torna inviável ou muito limitada. Além disso, alguns estudos apontam que indivíduos com transtornos no desenvolvimento tendem a apresentar maiores escores em testes não verbais (Duarte, Covre, Braga & Macedo, 2011; Flanagan et al., 2012; Decker, Euglund & Roberts, 2012). No que se refere aos instrumentos que mensuram inteligência fluida, estudos como o de Flanagan, McGrew e Ortiz (2000) apontaram que a terceira edição do WISC, medida que é utilizada no contexto brasileiro, não possui uma boa medida de inteligência fluida (Schelini, 2006). Já os testes SON tem como foco a mensuração da inteligência fluida e são citados como uma alternativa na

avaliação de grupos difíceis de testar (Mecca, Orsati & Macedo, 2014; Mecca et al, 2014).

Assim, entendendo que os testes não verbais permitem acessar habilidades a partir de instruções e respostas sem a utilização da fala, reconhecendo que há escassez de bons instrumentos de medida de inteligência não verbal no contexto brasileiro, levando em consideração a importância de estudos que apresentem a precisão e evidências de validade dos escores do instrumento e com o intuito de contribuir com o campo da avaliação cognitiva de crianças, adolescentes e jovens adultos, esta pesquisa foi desenvolvida, pois pretende apresentar dados sobre as características psicométricas dos itens dos subtestes e evidências de validade do teste SON-R 6-40.

Método

Participantes

Participaram deste estudo 711 pessoas, sendo 364 (51,4%) do sexo feminino. As idades tiveram média de 15,98 anos (*DP* = 8,38), com mínimo de 6 anos e 4 meses e máximo de 37 anos e 9 meses. Do total dos respondentes, 466 (65,54%) eram crianças e estavam cursando o ensino fundamental, 68 (9,56%) eram adolescentes e estavam cursando o ensino médio e 177 (24,89 %) eram adultos que foram contactados no ambiente de trabalho ou em faculdades. Os indivíduos que responderam o teste são provenientes de quatro regiões brasileiras: Nordeste (306), Centro-oeste (103), Sudeste (270) e Sul (32). Para a seleção dos municípios de cada região, além do índice de desenvolvimento humano (IDH), foram utilizados os seguintes critérios: (1) as cidades selecionadas deveriam contemplar os maiores estados da região e (2) estados com maiores e menores IDH do país deveriam ser inseridos na amostra. Para cada município onde ocorreu a coleta dos dados, trinta e quatro pessoas foram avaliadas (17 homens e 17 mulheres).

Instrumento

O SON-R 6-40 é um teste não verbal de inteligência focado na mensuração da inteligência fluida e que pode ser aplicado sem o uso da linguagem falada ou escrita. É composto por quatro subtestes: Analogias (36 itens), Mosaicos (26 itens), Categorias (36 itens) e Padrões (26 itens). Os subtestes Analogias e Categorias são subtestes de raciocínio e os subtestes Mosaicos e Categorias são de raciocínio espacial. Os examinandos não respondem todos os itens dos subtestes porque há critérios de interrupção da aplicação do subteste e procedimento adaptativo da testagem.

O subteste Analogias é composto por três séries de doze itens de múltipla escolha. Neste subteste, são apresentados três exemplos antes do início da testagem. O respondente deve descobrir a alteração ocorrida no primeiro par de figuras e utilizar a mesma alteração para identificar a resposta certa.

No subteste Mosaicos o respondente deve reproduzir uma figura modelo utilizando alguns quadrados coloridos que recebe. É composto por duas séries e dois exemplos são fornecidos.

No subteste Categorias o respondente deve descobrir o conceito subjacente aos três desenhos inicialmente apresentados e escolher dois desenhos que apresentam o mesmo conceito. É um subteste de múltipla escolha, composto por três séries de doze itens e três exemplos.

No subteste Padrões o respondente deve preencher com um lápis a parte omitida no desenho. É composto por duas séries e são fornecidos dois exemplos. Exemplos dos itens dos quatro subtestes podem ser encontrados no *website* dos testes SON (www.testresearch.nl).

Procedimento

Inicialmente, o projeto de pesquisa foi avaliado pelo Comitê de Ética em Pesquisa em Seres Humanos. Após sua aprovação, foi formada uma equipe para auxiliar nas aplicações do teste pelo território nacional. Um dos primeiros passos para assegurar a qualidade das aplicações foi selecionar aplicadores com experiência no uso de testes. A maior parte da equipe era formada por psicólogos licenciados que trabalhavam em instituições do setor privado ou público. E toda a equipe de aplicadores recebeu um treinamento pessoalmente sobre a aplicação do SON-R 6-40. O treinamento foi conduzido por um membro da equipe de desenvolvimento, sob a supervisão de um dos autores do teste SON-R 6-40. Após o treinamento, os aplicadores participaram de exercícios de simulação para verificar se a aplicação do teste estava sendo realizada de forma correta. Quando necessário, era fornecido *feedback* para os aplicadores com o objetivo de eliminar os poucos erros de aplicação.

Quando os respondentes eram crianças e adolescentes, um psicólogo da equipe entrava em contato com diretores de escolas públicas ou particulares a fim de obter permissão para a realização da pesquisa em seus estabelecimentos. Após permissão da escola, foram enviadas cartas aos pais descrevendo o objetivo da pesquisa e os procedimentos, além de termos de consentimento livre e esclarecido que deveriam ser assinados em caso de concordância de participação. Quando os respondentes eram adultos, o psicólogo da equipe entrava em contato com empresas, universidades, quartéis ou outra instituição a fim de alcançar esse público. O termo e a carta descrevendo os objetivos e procedimentos da pesquisa eram entregues diretamente ao respondente.

As aplicações do teste ocorreram em escolas públicas e particulares, durante o horário de aula, ou na casa ou local de trabalho do respondente em horário previamente

agendado. O tempo médio de aplicação do SON-R 6-40 foi de 50 minutos. Todas as aplicações ocorreram em sessões individuais e foram realizadas por psicólogos devidamente treinados para assegurar a padronização durante a testagem. Após a pesquisa, foi entregue ao participante ou ao responsável (quando o respondente era criança ou adolescente) um relatório descrevendo o desempenho do sujeito no teste.

Análise dos dados

As análises exploratórias e descritivas realizadas foram efetuadas no *software* SPSS *(Statistical Package for the Social Sciences)* versão 18. Para a investigação dos parâmetros psicométricos dos itens de cada subteste foi utilizada a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI). Para fazer as análises da TRI foi utilizado o *Bilog-MG* 3.0 (Zimowski, Muraki, Mislevy & Bock, 1996), que é um *software* que pode ser usado para a estimação dos modelos da TRI.

Para a verificação da estrutura fatorial do SON-R 6-40 utilizou-se o software FACTOR versão 9.2 (Lorenzo-Seva & Ferrando, 2013). Para determinar o número de fatores a extrair utilizou-se um tipo de análise Paralela - *Optimal Implementation of Parallel Analysis* (Timmerman & Lorenzo-Seva, 2011), em função da sua robustez para avaliar o número de fatores a ser retido (Damásio, 2012; Baglin, 2014; Timmerman & Lorenzo-Seva; 2011). A análise fatorial foi realizada usando *Minimum Rank Factor Analysis* (Ten Berge & Kiers, 1991), com base em correlações politômicas (Baglin, 2014).

Os itens muito fáceis ($p > 0,90$) e os itens muito difíceis ($p < 0,10$) não foram considerados na verificação da estrutura fatorial do SON-R 6-40. Assim, 46 dos 124 itens não foram incluídos na análise fatorial. Foram utilizadas parcelas de itens em vez de itens individuais para evitar o surgimento de fatores artificiais de dificuldade na análise fatorial. O uso de parcelas de itens diminui a possiblidade de surgimento de

fatores artificiais de dificuldade na análise fatorial com itens dicotômicos e gera soluções mais estáveis (Little, Cunningham, Shahar & Widaman, 2002; Rocha & Chelladurai, 2012). Assim, os 78 itens remanescentes foram distribuídos em 20 parcelas de itens. Cada parcela de itens consiste de três ou quatro itens: a distribuição dos itens nas parcelas de itens foi realizada na sequência dos itens. Assim, por exemplo, a primeira parcela de itens é composta pelos os três primeiros itens do subteste Analogias.

Em relação à análise da consistência interna, usou-se o coeficiente Lamba 2 de Guttman, uma vez que estudos apontam que esse coeficiente é uma estimativa melhor da fidedignidade e o Alfa de Cronbach tende a subestimar a fidedignidade da medida, estimando de forma conservadora a verdadeira fidedignidade (Maroco & Garcia-Marques, 2006; Sijtsma, 2009; Sijtsma, 2012). Para estimar o coeficiente Lambda 2 de Guttman foi utilizado o SPSS na sua versão 18.0.

Resultados e Discussão

O primeiro procedimento adotado na análise dos dados foi avaliar estrutura fatorial do instrumento. O pressuposto de unidimensionalidade, no caso dos modelos unidimensionais, deve ser assegurado para a TRI ser aplicada.

Andrade, Laros e Gouveia (2010) apresentam diferentes formas de avaliar a dimensionalidade de um instrumento. Nesta pesquisa foi utilizada uma Análise Paralela (AP) com 20 parcelas de itens do SON-R 6-40 para avaliar a quantidade de fatores a extrair. O programa FACTOR 9.2 foi usado e 500 matrizes de correlações aleatórias foram analisadas. Os *eigenvalues* empíricos e aleatórios (percentil 95) apresentaram os seguintes valores, em sequência por dimensão recomendada: (a) empíricos 68,6 e 6,8; (b) aleatórios 11,1 e 10,3. Estes resultados indicaram a presença de um único fator no instrumento.

Após a Análise Paralela, foi realizada uma análise fatorial com o método *Minimum Rank Factor Analysis – MRFA* (Ten Berge & Kiers, 1991), sugerida por Timmerman & Lorenzo-Seva (2011). Os resultados da análise fatorial podem ser observados na Tabela 1 e indicaram que o fator único explicou 69,2% da variância comum. Segundo Baglin (2014) a percentagem da variância comum explicada pode ser considerada uma medida de ajuste do modelo aos dados. Além disso, todas as cargas fatoriais para o modelo foram acima de 0,67, variando entre 0,67 a 0,81. Esses valores (cargas fatoriais acima de 0,60) torna adequada a análise com MRFA e os resultados também indicam uma estrutura unidimensional (Reise, Waller & Comrey, 2000).

Tabela 1. *Cargas fatoriais (CF) e comunalidades (h^2) da Análise Fatorial Exploratória do SON-R 6-40*

Parcela de itens	n itens	CF	h^2
Analogias 1	3	0,73	0,69
Analogias 2	4	0,69	0,72
Analogias 3	3	0,78	0,84
Analogias 4	4	0,71	0.76
Analogias 5	4	0,80	0,82
Analogias 6	4	0,71	0,69
Mosaicos 1	4	0,77	0,90
Mosaicos 2	4	0,77	0,85
Mosaicos 3	4	0,72	0,83
Mosaicos 4	4	0,81	0,80
Categorias 1	3	0,69	0,71
Categorias 2	3	0,67	0,71
Categorias 3	4	0,71	0,77
Categorias 4	5	0,70	0,81
Categorias 5	4	0,72	0,76
Categorias 6	4	0,73	0,81
Padrões 1	4	0,75	0,89
Padrões 2	4	0,78	0,82
Padrões 3	4	0,71	0,84
Padrões 4	5	0,81	0,85

Notas. Percentagem explicada da variância comum = 69,16%; RMSR (*Root Mean Square of Residuals*) = 0,0895.

É possível observar que a validade depende em certo grau da fidedignidade, porém a consistência interna fornece apenas evidências fracas referentes à validade de

um teste, sendo indicado pensar a fidedignidade como um pré-requisito para a validade, em vez da evidência da validade em si (Hogan, 2006). O índice de consistência interna dos quatro subtestes e da escala geral foram estimados para cada grupo etário utilizando o coeficiente Lambda 2 de Guttman. Para estimar os coeficientes de consistência interna não foram incluídos os itens muito fáceis ($p > 0,95$) e os itens muito difíceis ($p < 0,10$).

Na Tabela 2 é possível observar que os coeficientes de fidedignidade dos subtestes variaram entre 0,77 e 0,96. A escala geral apresentou coeficientes extremamente altos, variando entre 0,92 e 0,94. A Tabela 2 indica valores altos para os escores de cada subteste e em cada faixa etária. Para todos os grupos de idade os coeficientes de fidedignidade são superiores a 0,77 para os escores dos subtestes e superior a 0,95 para o SON-QI.

Tabela 2. *Fidedignidade dos escores por subtestes e escala geral*

		Fidedignidade (Lambda 2 de Guttman)				
Idade	N	Ana	Mos	Cat	Pad	QI-SON
6 anos	42	0,86	0,92	0,85	0,90	0,95
7 anos	41	0,91	0,88	0,89	0,90	0,96
8 anos	42	0,90	0,90	0,90	0,91	0,96
9 anos	41	0,83	0,86	0,88	0,90	0,95
10 anos	42	0,88	0,91	0,87	0,93	0,96
11 anos	43	0,88	0,87	0,92	0,77	0,96
12 anos	46	0,86	0,88	0,89	0,89	0,96
13 anos	46	0,88	0,85	0,84	0,91	0,96
14 anos	45	0,90	0,90	0,91	0,89	0,97
15 anos	46	0,86	0,87	0,91	0,89	0,96
16 anos	43	0,93	0,89	0,91	0,92	0,97
18 anos	39	0,92	0,91	0,93	0,91	0,97
20 anos	41	0,93	0,91	0,93	0,91	0,97
22 anos	39	0,92	0,91	0,93	0,93	0,98
27 anos	37	0,94	0,91	0,93	0,90	0,97
32 anos	40	0,91	0,84	0,94	0,89	0,97
37 anos	38	0,95	0,88	0,96	0,87	0,98
Todas	711	0,93	0,92	0,93	0,94	0,98

Nota. Ana = Analogias; Mos = Mosaicos; Cat = Categorias; Pad = Padrões.

Depois de estimar a fidedignidade dos escores, serão apresentados primeiramente os parâmetros de dificuldade e discriminação segundo a Teoria Clássica dos Testes (TCT) e depois os valores segundo a TRI . Segundo Hogan (2006), os procedimentos tradicionais de análise de itens são provenientes da TCT e dependem de dois conceitos: o índice de dificuldade e o índice de discriminação do item.

Apesar de receberem a mesma denominação da TCT, na TRI o parâmetro de dificuldade não é medido por uma proporção e o parâmetro de discriminação não é uma correlação. Os parâmetros na TRI são estimados a partir das respostas de um grupo de indivíduos submetidos a um conjunto de itens (Andrade, Tavares & Valle, 2000). No entanto, os parâmetros estimados pela TRI e TCT geralmente são comparáveis, mesmo sendo calculados de forma diferente (Fan, 1998).

É importante garantir a progressiva dificuldade dos itens no teste SON-R 6-40 porque os itens dos subtestes foram desenvolvidos a partir de uma teoria de dificuldade; isto é, uma revisão analítica dos fatores mais importantes que explicam os níveis sucessíveis de dificuldade dos itens (Snijders, Tellegen & Laros, 1989). Esse procedimento foi adotado para garantir uma ampla variedade de dificuldade dos itens, bem como facilitar uma aplicação adaptativa.

De acordo com Hogan (2006), os níveis de dificuldade dos itens geralmente são denominados valores p, sendo "p" uma referência ao percentual ou à proporção de acertos. A Tabela 3 apresenta o valor *p* de cada item nas diferentes séries dos quatro subtestes do SON-R 6-40. Observa-se que no subteste Analogias todos os itens estão colocados em ordem crescente de dificuldade. No subteste Mosaicos, apenas o item 3 da série A apresentou uma dificuldade maior que o item posterior. Já no subteste Categorias, dois itens (1 e 4) da série C apresentaram dificuldade maior que os itens seguintes. E por fim, apenas o item 1 da série B do subteste Padrões revelou maior dificuldade que o item posterior.

Tabela 3. *Valor p dos itens e a média dos valores p (N = 711)*

Analogias					Mosaicos			
Item	série a	série b	série c	média	Item	série a	série b	média
1	1,00	1,00	0,98	0,99	1	0,99	0,99	0,99
2	0,95	0,95	0,96	0,95	2	0,98	0,97	0,97
3	0,93	0,91	0,92	0,92	3	0,86*	0,90	0,88
4	0,85	0,86	0,86	0,86	4	0,88	0,87	0,87
5	0,76	0,72	0,76	0,75	5	0,74	0,79	0,76
6	0,40	0,49	0,56	0,48	6	0,53	0,71	0,62
7	0,37	0,42	0,47	0,42	7	0,41	0,61	0,51
8	0,27	0,28	0,37	0,31	8	0,38	0,48	0,43
9	0,17	0,19	0,33	0,23	9	0,21	0,26	0,23
10	0,10	0,10	0,16	0,12	10	0,12	0,18	0,15
11	0,06	0,07	0,10	0,08	11	0,03	0,09	0,06
12	0,03	0,04	0,05	0,04	12	0,03	0,05	0,04
					13	0,01	0,02	0,01
Média	0,49	0,50	0,54	0,51	Média	0,47	0,53	0,50
Categorias					Padrões			
Item	série a	série b	série c	média	Item	série a	série b	média
1	0,96	0,96	0,96*	0,96	1	0,95	0,92*	0,93
2	0,95	0,91	0,97	0,94	2	0,95	0,94	0,94
3	0,93	0,86	0,89	0,89	3	0,90	0,89	0,89
4	0,84	0,83	0,75*	0,80	4	0,84	0,89	0,86
5	0,63	0,71	0,80	0,71	5	0,70	0,76	0,73
6	0,35	0,42	0,52	0,43	6	0,67	0,71	0,69
7	0,33	0,37	0,39	0,36	7	0,53	0,62	0,57
8	0,17	0,35	0,32	0,28	8	0,27	0,43	0,35
9	0,14	0,19	0,21	0,18	9	0,24	0,29	0,26
10	0,09	0,13	0,12	0,11	10	0,12	0,21	0,16
11	0,05	0,11	0,08	0,08	11	0,08	0,13	0,10
12	0,03	0,05	0,04	0,04	12	0,03	0,06	0,04
					13	0,03	0,04	0,03
Média	0,45	0,49	0,50	0,48	Média	0,48	0,53	0,50

Nota. * item é mais difícil que o item seguinte.

O segundo conceito tradicional de análise de itens é o parâmetro de discriminação do item, que se refere ao poder ou potencial que o item tem de diferenciar sujeitos com magnitudes semelhantes no construto avaliado. Ou seja, o objetivo é que o item diferencie os indivíduos que apresentam mais da característica que está sendo mensurada daqueles que apresentam menos (Hogan, 2006).

Na literatura psicométrica, há diferentes formas de avaliar a discriminação do item: *r* bisserial, *r* ponto-bisserial, *r* tetracórica e coeficiente fi. Nesta pesquisa foi utilizada a *r* bisserial porque segundo Wilson, Wood e Gibbons (1991) esse índice é uma medida de associação entre o desempenho no item e o desempenho no teste, sendo menos influenciada pela dificuldade do item e tende a ser invariante quando o teste é aplicado em outros contextos.

A correlação bisserial pode apresentar valores entre -1 e +1, porém, é esperada uma correlação positiva, refletindo o fato de que as respostas corretas ao item são mais frequentes nos examinandos com escores totais altos (Valentini & Laros, 2011). Assim, itens que apresentam maior correlação são os que apresentam um maior poder de discriminação. Os valores negativos indicam que há algum problema com o item que precisa ser corrigido ou analisado, por exemplo, respostas erradas no gabarito (Urbina, 2014). A Tabela 4 apresenta os valores estimados dessa correlação.

Tabela 4. *A discriminação dos itens em cada subteste do SON-R 6-40 (N = 711).*

Analogias					Mosaicos			
Item	série a	série b	série c	média	Item	série a	série b	média
1	0,60	0,82	0,82	0,75	1	0,84	0,83	0,83
2	0,65	0,74	0,81	0,73	2	0,81	0,85	0,83
3	0,68	0,81	0,83	0,77	3	0,84	0,93	0,88
4	0,73	0,75	0,79	0,76	4	0,86	0,93	0,89
5	0,73	0,82	0,78	0,78	5	0,86	0,89	0,87
6	0,76	0,76	0,81	0,78	6	0,81	0,91	0,86
7	0,75	0,87	0,80	0,81	7	0,83	0,91	0,87
8	0,82	0,85	0,79	0,82	8	0,88	0,89	0,88
9	0,89	0,88	0,84	0,87	9	0,83	0,83	0,83
10	0,92	0,90	0,89	0,90	10	0,82	0,80	0,81
11	0,91	0,85	0,84	0,87	11	0,77	0,82	0,79
12	0,77	0,62	0,84	0,74	12	0,80	0,80	0,80
					13	0,72	0,79	0,75
Média	0,77	0,80	0,82	0,80	Média	0,82	0,86	0,84
Categorias					Padrões			
Item	série a	série b	série c	média	Item	série a	série b	média
1	0,45	0,86	0,80	0,70	1	0,81	0,96	0,88
2	0,79	0,72	0,73	0,75	2	0,95	0,97	0,96
3	0,69	0,74	0,79	0,74	3	0,96	0,97	0,96
4	0,74	0,72	0,71	0,72	4	0,92	1,00	0,96
5	0,63	0,72	0,76	0,70	5	0,87	0,93	0,90
6	0,72	0,86	0,80	0,79	6	0,94	0,95	0,94
7	0,73	0,77	0,81	0,77	7	0,90	0,94	0,92
8	0,87	0,92	0,86	0,88	8	0,75	0,88	0,81
9	0,92	0,94	0,86	0,91	9	0,83	0,83	0,83
10	0,90	0,94	0,83	0,89	10	0,79	0,82	0,80
11	0,91	0,94	0,87	0,91	11	0,83	0,83	0,83
12	0,84	0,88	0,79	0,84	12	0,86	0,84	0,85
					13	0,83	0,84	0,83
Média	0,76	0,83	0,80	0,80	Média	0,86	0,90	0,88

Antes de apresentar as estimativas dos parâmetros a partir da TRI, é necessário avaliar qual modelo (de um, dois ou três parâmetros) se adequa melhor aos dados empíricos. Embretson e Reise (2000) afirmam que existem vários testes estatísticos para indicar em que grau um dado modelo da TRI se ajusta aos dados. Essas estatísticas são chamadas de índices de bondade de ajuste (*Goodness of Fit*) (Andrade, Laros &

Gouveia, 2010). Um fraco ajuste do modelo não pode assegurar que os parâmetros dos itens e das habilidades são invariantes.

A escolha do modelo que foi utilizado baseou-se na orientação de De Ayala (2009). O autor sugere analisar os índices de ajustes de cada modelo e decidir qual modelo utilizar a partir do cálculo da diferença do ajuste do modelo 1 e 2, dividido pela diferença dos graus de liberdade dos dois modelos. Para ser significativo esse valor que é chamado razão critica deve ser maior que 1,96. Outra estimativa para avaliar o ajuste do modelo é o índice R^2_{Δ}, que indica o quanto o modelo melhorou.

A Tabela 5 apresenta esses valores. Para Mosaicos e Padrões não foi estimado o modelo de três parâmetros porque não se trata de subtestes de múltipla escolha. A partir dos dados de ajustes do modelo de 2 parâmetros foi escolhido para todos os subtestes.

Tabela 5. *Estatísticas de ajuste do modelo de TRI dos quatro subtestes do SON-R 6-40*

Analogias (36 itens)					
Modelo	Ajuste	Δ ajustes	Δ *df*	r. c.	R^2_{Δ}
1 PL	13.458,32	-	-	-	-
2 PL	13.325,39	132,93	36	3,69	0,98%
3 PL	13.513,76	-188,37	36	-5,23	-1,41%
Mosaicos (26 itens)					
1 PL	81.777, 22	-	-	-	-
2 PL	81.086,88	690,34	26	26,55	0,84%
Categorias (36 itens)					
1 PL	14.033,41	-	-	-	-
2 PL	13.743,39	290,02	36	8,05	2,06%
3 PL	13.813,83	-70,44	36	-1,96	-0,51%
Padrões (26 itens)					
1 PL	81.792,65	-	-	-	-
2 PL	80.529,06	1.263,59	26	48,60	1,54%
Escala Total (124 itens)					
1 PL	45.750,89	-	-	-	-
2 PL	44.901,44	849,45	124	6,85	1,85%
3 PL	46.853,34	-1.951,90	124	-15,74	-4,34%

Notas. O ajuste do modelo aos dados foi avaliado com o -2 *log likelihood*; Δ ajustes = diferença de ajustes; Δ *df* = diferença de graus de liberdade; r.c. = razão critica - para ser significativa a 5% a razão precisa ser > 1,96.

O parâmetro de dificuldade ou parâmetro *b* (também identificado como *location* ou *threshold*) na TRI é expresso em termos de escores padrões, variando de -3 (itens são muito fáceis) até +3 (itens muito difíceis). Esse parâmetro é medido na mesma escala de habilidade e corresponde ao valor do *teta* para o qual a probabilidade de acerto é de 0,50. Quanto maior o valor do parâmetro *b* do item, maior será a habilidade requerida para um indivíduo acertar o item (Andrade, Laros & Gouveia, 2010). Assim, quanto maior o valor de *b*, mais difícil é o item (Valentini & Laros, 2011).

A Figura 1 apresenta uma visão geral da ordenação dos itens dos quatro subtestes de acordo com o parâmetro *b*. Observa-se que a distribuição dos itens abrange certa extensão do construto avaliado e que de forma geral, os itens de todos os subtestes apresentam uma ordem crescente de dificuldade. Entretanto, com base nos índices de dificuldade, nota-se que alguns itens não apresentaram ordem de dificuldade progressiva, sugerindo então a necessidade de observá-los no momento do estudo de normatização. No estudo da versão original holandesa, os últimos itens do subteste Padrões apresentaram um índice de dificuldade menor do que o observado neste estudo (Tellegen & Laros, 2014).

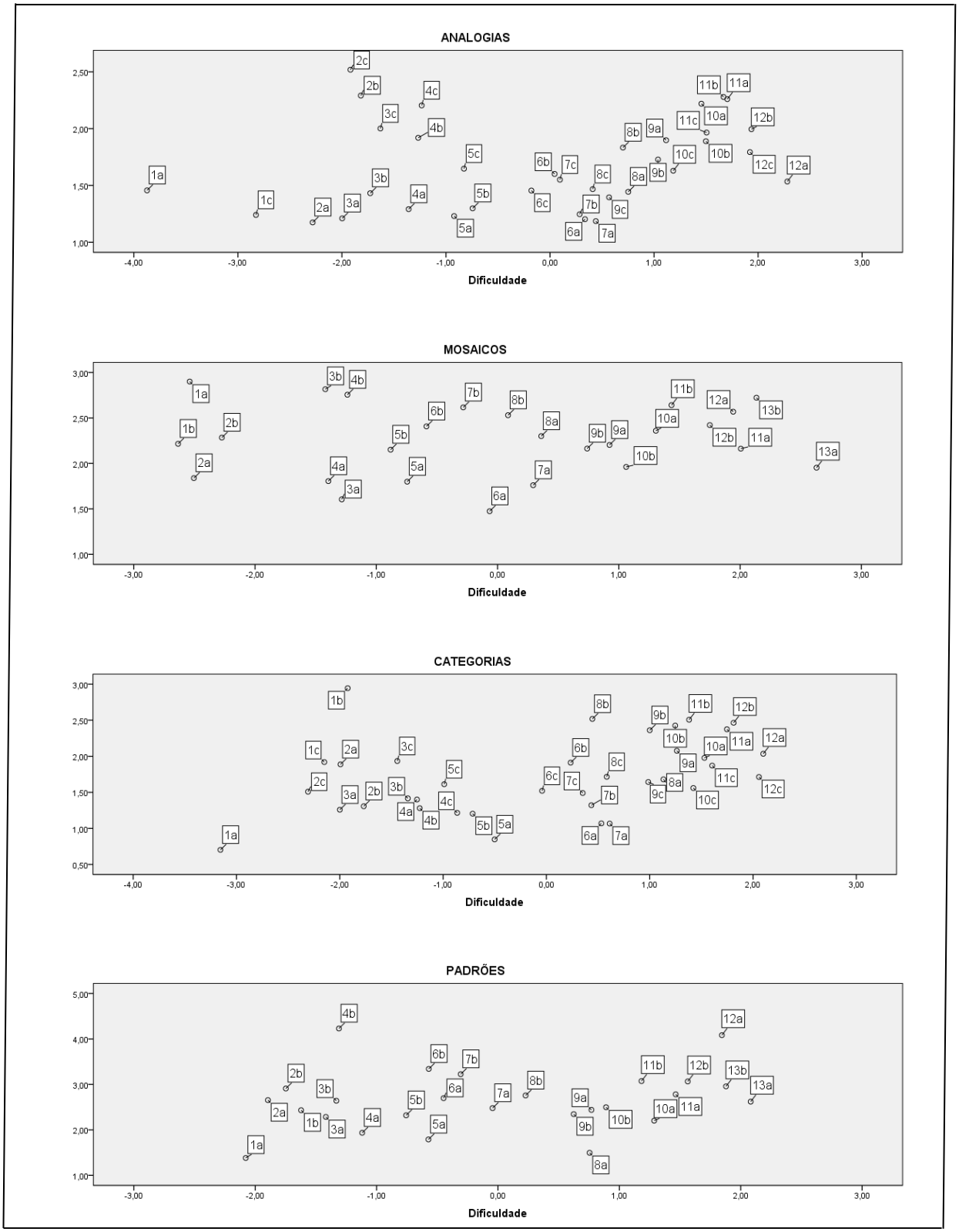

FIGURA 1

Dificuldade (paraâmetro *b)* dos itens dos quatro subtestes do SON-R 6-40.

Na TRI, o parâmetro de discriminação é representado pela letra *a* e também é identificado como valor do *slope* apresentado na fase 2 do programa *Bilog-MG*. Os valores do parâmetro *a* podem variar de mais infinito a menos infinito. Porém, geralmente eles apresentam valores entre 0,0 e 2,0, sendo que os valores apropriados de *a* seriam aqueles maiores que 1 (Andrade, Laros & Gouveira, 2010; Andrade, Tavares & Valle, 2000). Baker (2001) apresenta a seguinte classificação do parâmetro *a* por faixa de valores: 0,0 – nenhuma discriminação; 0,01 a 0,34 – discriminação muito baixa; 0,35 a 0,64 – discriminação baixa; 0,65 a 1, 34 – discriminação moderada; 1,35 a 1,69 – discriminação alta; e acima de 1,70 é considerado um item com discriminação muito alta (Andrade, Laros & Gouveia, 2010). É possível observar na Tabela 6 o valor estimado do parâmetro *a* de cada item. Com base nesses resultados, os itens dos subtestes são discriminativos.

Tabela 6. *Valores do parâmetro a dos itens dos subtestes (N = 711).*

Analogias					Mosaicos			
Item	série a	série b	série c	média	Item	série a	série b	média
1	1,45	0,89	1,24	1,19	1	2,90	2,21	2,55
2	1,17	2,29	2,52	1,99	2	1,83	2,28	2,05
3	1,20	1,43	1,99	1,54	3	1,60	2,81	2,20
4	1,29	1,91	2,19	1,79	4	1,80	2,75	2,27
5	1,23	1,30	1,64	1,39	5	1,79	2,15	1,97
6	1,20	1,60	1,45	1,41	6	1,47	2,40	1,93
7	1,19	1,26	1,54	1,33	7	1,76	2,61	2,18
8	1,45	1,84	1,46	1,58	8	2,30	2,53	2,41
9	1,93	1,72	1,39	1,68	9	2,20	2,16	2,18
10	2,22	1,91	1,62	1,91	10	2,35	1,96	2,15
11	2,29	2,37	1,95	2,20	11	2,16	2,64	2,4
12	1,56	2,07	1,51	1,71	12	2,56	2,42	2,49
					13	1,95	2,72	2,33
Média	1,51	1,71	1,70	1,65	Média	2,05	2,43	2,24
Categorias					Padrões			
Item	série a	série b	série c	média	Item	série a	série b	média
1	0,70	2,94	1,92	1,85	1	1,38	2,43	1,90
2	1,88	1,30	1,50	1,56	2	2,65	2,91	2,78
3	1,25	1,41	1,93	1,53	3	2,28	2,64	2,46
4	1,40	1,28	1,21	1,29	4	1,93	4,23	3,08
5	0,84	1,20	1,61	1,21	5	1,78	2,32	2,05
6	1,06	1,91	1,52	1,49	6	2,69	3,34	3,01
7	1,06	1,32	1,49	1,29	7	2,47	3,22	2,84
8	1,67	2,51	1,71	1,96	8	1,49	2,75	2,12
9	2,07	2,36	1,64	2,02	9	2,44	2,34	2,39
10	1,97	2,42	1,56	1,98	10	2,20	2,49	2,34
11	2,37	2,50	1,87	2,24	11	2,78	3,07	2,92
12	2,03	2,46	1,71	2,06	12	4,08	3,06	3,57
					13	2,62	2,95	2,78
Média	1,52	1,96	1,63	1,71	Média	2,36	2,90	2,63

Para os testes de múltipla escolha, a TRI também informa a probabilidade do examinando responder corretamente devido ao acaso (chute). Porém, nesta pesquisa não foi estimado o parâmetro *c* dos itens dos subtestes Analogias e Categorias porque o modelo de três parâmetros não apresentou um ajuste adequado.

O *Bilog-MG,* além de fornecer os valores dos parâmetros, também fornece a curva de informação para cada um dos itens dos subtestes e para o teste total. A curva

de informação do teste representa o somatório das informações de todos os itens. A Figura 2 apresenta a curva de informação total do teste SON-R 6-40. A curva de informação do teste é representada pela linha contínua e a linha pontilhada representa a curva do erro padrão da medida. Observa-se que nos extremos a curva do erro supera a curva de informação porque o teste produz mais erro de informação do que informação legítima.

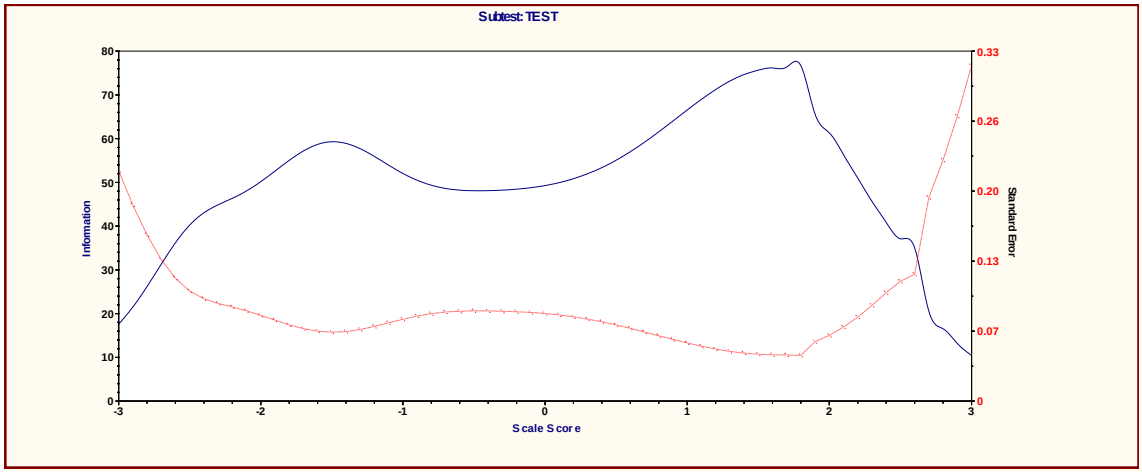

FIGURA 2
Curva de informação do teste SON-R 6-40.

Outra análise que foi realizada e que é reconhecida por Hogan (2006) como uma fonte potencial de informações a respeito da validade de construto, é verificar o aumento dos escores brutos em um teste a partir do aumento da idade. As mudanças desenvolvimentais são esperadas e há a expectativa que, por exemplo, o desenvolvimento em matemática aumente do terceiro ano para o quarto ano, do quarto ano para o quinto ano e assim por diante (Hogan, 2006).

A Tabela 7 apresenta o aumento da média dos escores brutos do teste nas diferentes faixas etárias contempladas. O valor apresentado na Tabela 6 é calculado a partir da soma dos escores brutos dos quatro subteste dividido pelo número de respondentes naquela faixa etária. Observa-se que há um aumento progressivo dos escores brutos com o aumento da idade, porém esse crescimento não é linear. Entre a

faixa etária de 6 a 7 anos há um aumento de 7,9 pontos. Já entre a faixa etária de 7 a 8 anos o aumentos do escore é de apenas 2,6 pontos. Entre a faixa etária de 27 e 32 anos há uma diminuição de 4,25 pontos. Apesar desse crescimento não linear da média dos escores brutos, apenas na faixa etária dos 14 anos, 32 e 37 anos há uma diminuição do escore. No estudo de normatização do teste na Alemanha e Holanda, os escores também não apresentaram um crescimento linear ao longo das faixas etárias e também houve um decréscimo nas seguintes faixas etárias: entre 22 e 27 anos; 32 e 37 anos (Tellegen & Laros, 2014). Além, disso, estudos indicam que o desenvolvimento cognitivo não se faz de maneira contínua, podendo haver picos ou estagnações (Wechsler & Schelini, 2006; Schrank & Flanagan, 2003).

Tabela 7. *Características do escore total bruto em cada faixa etária*

Idade	Média	DP	Assimetria	Curtose
6	30,98	13,00	0,21	-0,61
7	38,88	14,91	-0,57	-0,27
8	41,48	15,46	-0,26	0,19
9	48,05	13,27	-0,57	0,72
10	55,90	15,45	-0,35	-0,91
11	56,16	14,08	-0,44	0,07
12	57,76	15,60	-0,02	0,79
13	66,98	15,16	0,25	1,06
14	64,22*	18,28	-0,37	0,47
15	66,93	15,26	-0,18	0,09
16	67,93	19,29	-0,54	0,24
18	74,67	18,44	-0,04	-0,49
20	75,76	18,98	0,18	-0,48
22	80,95	20,04	-0,38	0,16
27	83,05	19,31	-0,54	-0,37
32	78,80*	18,00	-0,40	0,50
37	73,16*	20,41	-0,15	-0,50

DP = desvio padrão.

A Figura 3 apresenta o crescimento da média do escore bruto por grupo de idade. A figura indica que, de maneira geral, houve um crescimento das médias nas

pontuações das faixas etárias estudadas. Assim como os dados da Tabela 6, foi observada uma queda nas pontuações nas últimas faixas etárias pesquisadas.

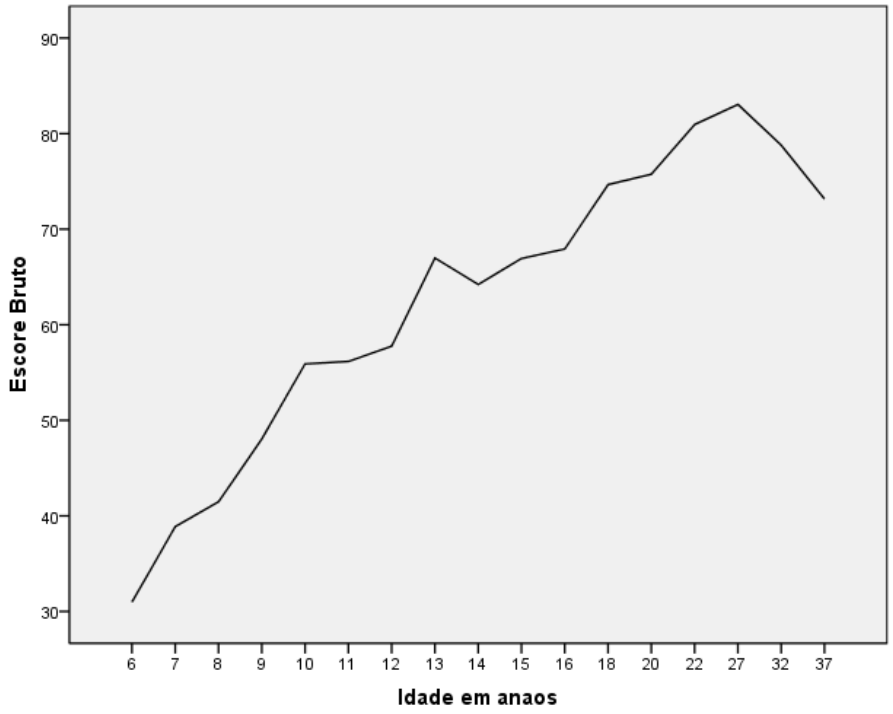

FIGURA 3
Média dos escores brutos por grupo de idade

A Figura 4 apresenta o escore bruto obtido em cada subteste nas faixas etárias estudadas e para os sexos feminino e masculino. É possível observar que os dados indicam que a relação entre escores brutos e aumento da idade também não é linear. Há um aumento considerável nas primeiras faixas etárias, entre 6 a 13 anos, porém esse crescimento do escore bruto diminui nas últimas faixas etárias estudadas. Esse dado também foi semelhante aos valores encontrados no estudo de normatização do SON-R 6-40 para Alemanha e Holanda (Tellegen & Laros, 2014).

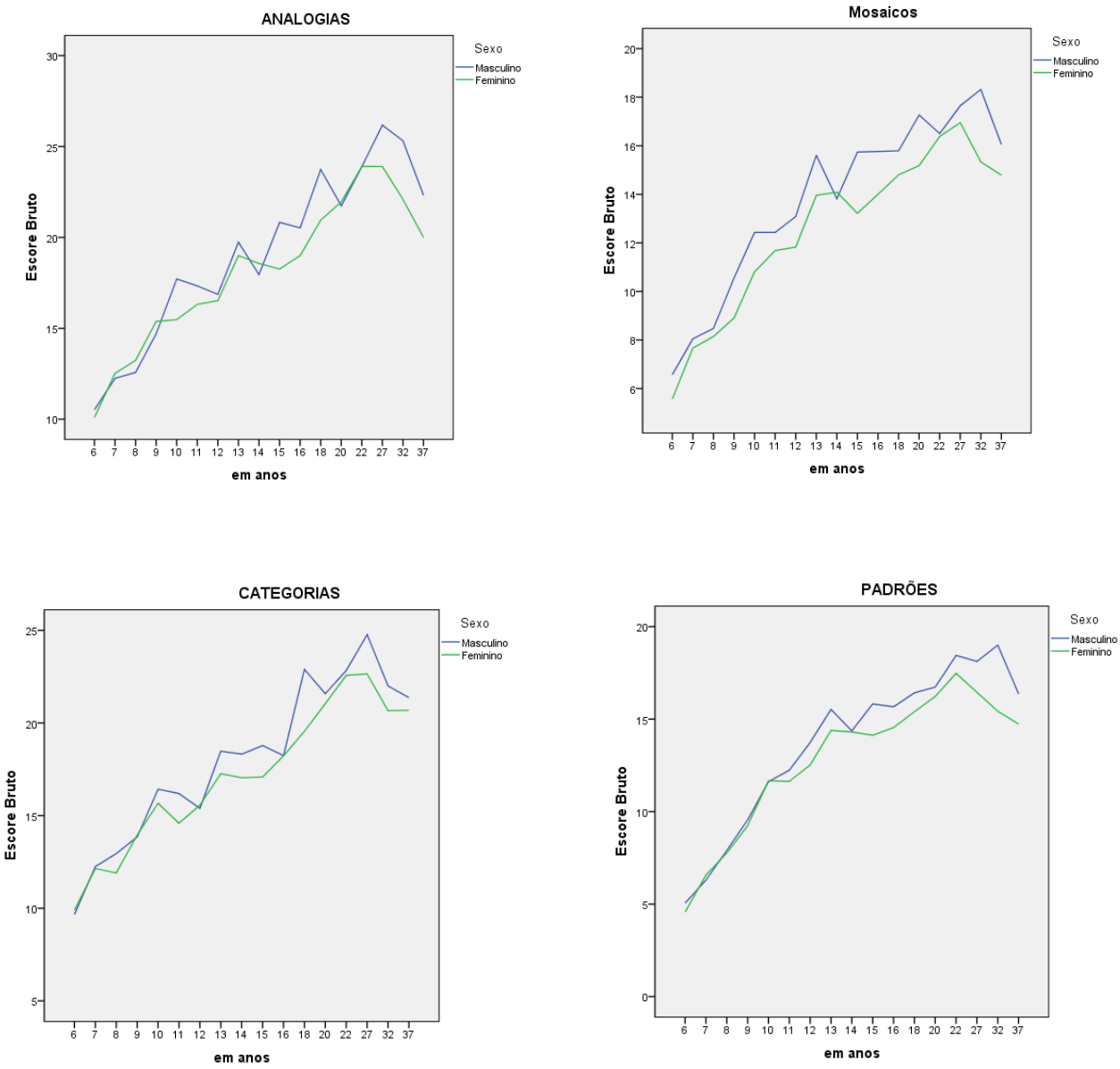

FIGURA 4
Média obtida nos subtestes por faixa etária para os sexos feminino e masculino.

Como pode ser observado, o sexo masculino apresentou um desempenho melhor nas médias dos subtestes. Entretanto, a pontuação dos dois grupos em algumas faixas etárias se intersecta e no subteste Padrões, por exemplo, os escores obtidos nas primeiras faixas etárias apresentam uma mesma tendência de crescimento.

Considerações finais

Este estudo teve como objetivo principal apresentar informações que revelam as caraterísticas psicométricas dos itens dos subtestes e evidências de validade de construto do teste SON-R 6-40. A Teoria de Resposta ao Item e a Teoria Clássica dos Testes foram utilizadas para estimar os parâmetros dos itens. As duas teorias foram usadas com o objetivo de tecer comparações e de estimular a utilização da TRI no contexto da avaliação psicológica.

Em relação às evidências de validade, foram realizadas duas análises: estudo exploratório da dimensionalidade do instrumento e análise do crescimento da pontuação ao longo das faixas etárias estudadas (mudanças desenvolvimentais). Pequenas quedas nas pontuações foram observadas, entretanto existe claramente uma tendência de crescimento da média dos escores brutos. Em relação às diferenças de gênero observadas, é necessário esperar a conclusão da coleta dos dados para a realização de análises mais específicas para identificar as trajetórias desenvolvimentais nos dois sexos e para verificar se realmente existe superioridade de um grupo. A consistência interna também foi estimada para cada faixa etária por meio do coeficiente Lambda 2 de Guttman.

De maneira geral, os resultados embasam o uso do SON-R 6-40 como uma escala geral, com alta qualidade psicométrica e com índices de fidedignidade adequados para o uso nas diferentes faixas etárias contempladas. Os parâmetros dos itens apresentam padrão semelhante aos parâmetros da versão original, já que foi possível tecer comparações entre os achados desse estudo e as características dos itens na versão holandesa porque o manual do teste discute de forma detalhada as análises empíricas dos itens.

Em relação às limitações desse estudo, ressalta-se que ainda são necessários estudos com a amostra completa para a faixa etária contemplada no teste SON-R 6-40, com a finalidade de investigar a adequação dos itens conforme a TRI. Também é importante a realização de mais pesquisas com grupos em diferentes contextos culturais para verificar a precisão da bateria e a invariância da estrutura fatorial. Além de aplicar em diferentes grupos, é importante utilizar diferentes estratégias para avaliar o instrumento psicológico destacado aqui, como por exemplo, validade de critério, teste-reteste, correlações com outros testes que avaliam o mesmo construto e diferentes técnicas de análises para obter evidências de validade do construto, como Análise Fatorial Confirmatória (CFA), Análise Fatorial de Informação Plena (FIFA), Análise Simultânea dos Componentes Principais (SCA), aplicação de técnicas para identificar itens que funcionam de forma diferente em relação a subgrupos específicos, ou seja, a presença de *Differential Item Functioning* (DIF), entre outros.

Por fim, considerando a pertinência e a contribuição dos testes de inteligência no diagnóstico precoce de deficiências intelectuais, por exemplo, e acreditando que intervenções bem elaboradas podem cooperar para melhorar as chances de alguém atingir sua capacidade plena, espera-se que esta pesquisa tenha contribuído para o campo da avaliação psicológica no Brasil, mais especificamente na área do desenvolvimento de instrumentos que buscam mensurar a inteligência fluida. Os resultados desta pesquisa permitem aos usuários do teste conhecer as características psicométricas dos itens que compõem os subtestes, possibilitando assim a avaliação da representatividade do traço latente. Espera-se que normatização do SON-R 6-40 seja concluída em breve e que o teste seja disponibilizado para a comercialização.

Referências

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria de resposta ao item: conceitos e aplicações.* São Paulo: ABE – 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,* 421-435.

Antonio, D. A. M., Mecca, T. P., & Macedo, E. C. (2012). O uso do teste não verbal Leiter-R na avaliação de inteligência em distúrbios do desenvolvimento. *Cadernos de Pós-Graduação em Distúrbios do Desenvolvimento, 12,* 9-15.

Baglin, J. (2014). Improving your exploratory factor analysis for ordinal data: A demonstration using FACTOR. *Practical Assessment, Research & Evaluation, 19*(5), 1-14.

Baumgartl, V. O., & Primi, R. (2006). *Contribuições da avaliação psicológica no contexto organizacional: um estudo com a BPR-5, o BFM -1 e o PMK.* São Paulo: Casa do Psicólogo.

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

Carroll, J. B. (2012). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues,* (pp. 69-76). New York, NY: The Guilford Press.

Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology, 31,* 161-179.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40,* 153-193.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1-22.

Damásio, B. F. (2012). Uso da análise fatorial exploratória em Psicologia. *Avaliação Psicológica, 11*(2), 213-228.

Decker, S. L., Englund, J. A., & Roberts, A. M. (2012). Intellectual and neuro-psychological assessment of individuals with sensory and physical disabilities and traumatic brain injury. In D. P. Flanagan & P. L. Harrison(Eds.), *Contemporary intellectual assessment: Theories, tests and issues,* (pp. 708-725). New York, NY: The Guilford Press.

De Ayala, R. J. (2009). *The theory and application of item response theory.* New York: Guilford Publishing.

Duarte, C. P., Covre, P., Braga, A. C., & Macedo, E. C. (2011). Visuospatial support for verbal short-term memory in individuals with Down syndrome. *Research in Developmental Disabilities, 32,* 1918-1923.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* New Jersey: IEA.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357-381.

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler Intelligence Scales and CHC theory: A contemporary approach to interpretation.* Boston: Allyn & Bacon.

Flanagan, D. P., & Harrison, P. L. (2012). *Contemporary intellectual assessment: Theories, tests and issues*. New York, NY: Guilford Press.

Flanagan, D. P., Alfonso, V. C., Mascolo, J. T., & Sotelo-Dynega, M. (2012). Use of ability tests in the identification of specific learning disabilities within the context of an operational definition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 643-669) New York, NY: The Guilford Press.

Fundo das Nações Unidas para a Infância - UNICEF. (2013). *Situação mundial da infância. Crianças com deficiência*. London: UNICEF Publishing.

Hogan, T. P. (2006). *Introdução à prática de testes psicológicos.* Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 53-91). New York: The Guilford Press.

Laros, J. A., & Tellegen, P. J. (2004). Cultural bias in the SON-R test: Comparative study of Brazilian and Dutch children. *Psicologia: Teoria e Pesquisa, 20,*103-111.

Laros, J. A., Jesus, G. R., & Karino, C. A. (2013). Validação brasileira do teste não-verbal de inteligência SON-R 2½-7[a]. *Avaliação Psicológica, 12,* 233-242.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weigthing the merits. *Structural Equation Modeling, 9,* 151-173.

Lorenzo-Seva, U., & Ferrando, P. J. (2013). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*(1), 88-91.

Maroco, J., & Garcia-Marques, T. (2006). Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? *Laboratório de Psicologia, 4,* 65-90.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 136-182). New York: Guilford Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities: Standing on the shoulders of the giants of psychometrics. *Intelligence, 37,* 1-10.

McGrew, K. S., & Flanagan, D. P. (1998). *The Intelligence Test Desk Reference (ITDR) – Gc-Gf Cross Battery Assessment.* Boston, MA: Allyn and Bacon.

Mecca, T. P., Orsati, F. T., & Macedo, E. C. (2014). Inteligência e transtornos do desenvolvimento. In A. G. Seabra, J. A. Laros, E. C. Macedo, & N. Abreu (Eds.), *Inteligência e funções executivas* (pp. 95-112). São Paulo: Memnon.

Mecca, T. P., Valentini, F., Laros, J. A., Lima, R. M. F., Schwartzman, J. S., & Macedo, E. C. (2013). Utilizando o teste não verbal de inteligência SON-R 2½-7[a] para avaliar crianças com Transtornos do Espectro do Autismo. *Revista Educação Especial, 26,* 603-618.

Pasquali, L (2010). *Instrumentação psicológica. Fundamentos e práticas.* Porto Alegre: Artmed.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*(3), 287-297.

Rocha, C. M., & Chelladurai, P. (2012). Item parcels in structural equation modeling: An applied study in sport management. *International Journal of Psychology and Behavioral Sciences, 2,* 46-53.

Roid, G. H., & Miller, L. J. (1997). *Leiter International Performance Scale-Revised.* Wood Dale, IL: Stoelting.

Schelini, P. W. (2006). Teoria das inteligências fluida e cristalizada: início e evolução. *Estudos de Psicologia, 11,* 323-332.

Schneider, J. W., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99-144). New York, NY: Guilford Press.

Schrank, F. A., & Flanagan, D. P. (2003). *WJ-III: Clinical use and interpretation.* Boston: Academic Press.

Seabra, A. G., Laros, J. A., Macedo, E. C., & Abreu, N. (2014). *Inteligência e funções executivas: Avanços e desafios para a avaliação neuropsicológica.* São Paulo: Memnon.

Sijtsma, K. (2009). On the use, misuse and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika, 77*, 4-20.

Snijders, J. Th., Tellegen, P. J., & Laros, J. A. (1988). *Snijders-Oomen niet-verbale intelligentie test SON-R 5½-17. [Snijders-Oomen non-verbal intelligence test SON-R 5½-17]*. Groningen: Wolters-Noordhoff.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Tellegen, P. J., & Laros, J. A. (2014). *SON-R 6-40. Non-verbal intelligence test: Research report.* Göttingen, Germany: Hogrefe Verlag.

Ten Berge, J. M. F., & Kiers, H. A. L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix. *Psychometrika, 56*, 309-315.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*, 209-220.

Urbina, S. (2014). *Essentials of Psychological Testing.* New Jersey: John Wiley & Sons.

Wechsler, S. M., & Schelini, P. W. (2006). Bateria de Habilidades Cognitivas Woodcock-Johnson III: Validade de Construto. *Psicologia: Teoria e Pesquisa, 22,* 287-295.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT. Scientific Software International.* Chicago: SSI.

MANUSCRITO 3

Evidências de validade convergente dos escores obtidos no teste SON-R 6-40

Título em inglês

Evidence of the convergent validity of obtained test scores on the SON-R 6-40

Sugestão de título abreviado

Validade convergente do SON-R 6-40

RESUMO

O objetivo deste estudo foi obter evidência de validade convergente dos escores do teste SON-R 6-40. O teste foi administrado junto com o WISC-IV em uma amostra de 120 crianças. O SON-R 6-40 é um teste não verbal de inteligência de origem holandesa, para o qual estão sendo elaboradas as normas brasileiras. Dez subtestes do WISC-IV e todos os subtestes do SON-R 6-40 foram administrados. A correlação, corrigida para atenuação, entre os escores totais do SON-R 6-40 e do WISC-IV foi de 0,73. Como esperado, a correlação mais alta foi obtida entre o SON-R 6-40 e o Índice de Organização Perceptual do WISC-IV (*r* = 0,84), índice composto por subtestes que avaliam a inteligência fluida. Os resultados obtidos são muito similares aos resultados encontrados em estudos realizados em outros países e indicam adequada validade convergente dos escores do SON-R 6-40 para a faixa etária investigada.

Palavras-chave: testes de inteligência; validade convergente; SON-R 6-40.

ABSTRACT

The purpose of this study was to acquire evidence of the convergent validity of the test scores on the SON-R 6-40. The test was administered together with the WISC-IV to a sample of 120 children. The SON-R 6-40 is a non-verbal test of intelligence of Dutch origin, for which Brazilian norms are being elaborated. Ten subtests of the WISC-IV and all four subtests of the SON-R 6-40 were administered. The correlation, corrected for attenuation, between the total scores on the SON-R 6-40 and the WISC-IV was .73. As expected, a higher correlation was obtained between the SON-R 6-40 and the Perceptual Organization Scale of the WISC-IV (r = .84), that is composed by subtests that assess fluid intelligence. The obtained results are very similar to those found in studies accomplished in other countries and indicate a satisfactory convergent validity of the test scores of the SON-R 6-40 for the investigated age group.

Keywords: intelligence tests; convergent validity; SON-R 6-40.

RESUMEN

El objetivo de este estudio fue obtener evidencias de validez convergente de los puntajes del test SON-R 6-40. El test fue administrado junto con el WISC-IV en una muestra de 120 niños. El SON-R 6-40 es un test no verbal de inteligencia de origen holandés, para el cual están siendo elaboradas las normas brasileñas. Diez subtests del WISC-IV y todos los subtests del SON-R 6-40 fueron administrados. La correlación, corregida para la atenuación, entre los puntajes totales del SON-R 6-40 y del WISC-IV, fue de 0,73. Como esperado, la correlación más alta fue obtenida entre el SON-R 6-40 y el Índice de Organización Perceptual del WISC-IV (r = 0,84), índice compuesto por subtests que evalúan la inteligencia fluida. Los resultados obtenidos son muy similares a los resultados encontrados en estudios realizados en otros países e indican una adecuada validez convergente de los puntajes del test SON-R 6-40 para el grupo etario investigado.

Palabras clave: tests de inteligencia; validez convergente; SON-R 6-40.

Primi (2003) afirma que a área da avaliação psicológica é responsável por operacionalizar as teorias psicológicas em eventos observáveis. Os instrumentos de avaliação apresentam atividades específicas aos respondentes como formas de se observar a manifestação do traço latente em questão, pois os traços latentes são características do indivíduo que não podem ser observadas diretamente (Andrade, Tavares & Valle, 2000). Assim, o traço latente deve ser inferido a partir da observação de variáveis secundárias que estão relacionadas a ele porque não é possível acessar diretamente o objeto (por exemplo, depressão), mas os atributos desse objeto (como perda de energia, alterações do sono, fadiga constante, baixa auto-estima). E os atributos só podem ser alcançados pelo comportamento manifesto (Pasquali, 2010).

A partir da forma como as pessoas respondem os instrumentos, as características psicológicas são deduzidas. Daí, então, a importância das pesquisas científicas para investigar as características e qualidade dos instrumentos. Como o comportamento humano e sua avaliação são complexos, é fundamental garantir a qualidade dos instrumentos utilizados no processo de avaliação para auxiliar o profissional no diagnóstico e no direcionamento da intervenção (Reppold, Gurgel & Hutz, 2014).

Fundamentados nos princípios da AERA, APA e NCME (1999), Reppold, Gurgel e Hutz (2014) afirmam que o ponto fundamental, primordial, no momento da construção e análise dos testes é as evidências de validade dos escores baseadas na estrutura interna e nas relações com variáveis externas convergentes. É necessário observar que qualquer medida não está deposta da possibilidade de erro. Todavia, a utilização de instrumentos que apresentam evidências de validade dos escores aumenta a confiança do usuário que os escores de um teste de fato indicam o construto de interesse e que as inferências baseadas nos escores de teste são adequadas (Hogan, 2006). Assim, diante da importância e necessidade de estudos que apresentam

evidências de validade de um instrumento novo com outros instrumentos já estabelecidos, o objetivo deste estudo foi obter evidências de validade convergente do teste não-verbal de inteligência SON-R 6-40 com o teste WISC-IV.

O teste SON-R 6-40 é a última versão dos testes SON, originalmente publicado na Holanda. No nome da bateria SON-R 6-40, a letra R indica que se trata de um teste revisado e os números fazem referência à faixa etária do público alvo do teste. Os testes SON (Snijders-Oomen Não-verbal) devem seu nome à primeira autora dos testes, Drª Nan Snijders-Oomen, que em 1943 desenvolveu uma bateria de testes que pretendia medir a inteligência fluida (Cattell, 1963). Como seu objetivo era a avaliação de crianças surdas, a bateria incluía diversas tarefas não-verbais relacionadas à habilidade espacial, raciocínio abstrato e concreto.

Os testes SON passaram por diversas revisões e refinamentos com o objetivo de preservar as características originais desses instrumentos. Atualmente existem três versões dos testes SON: o SON-R 2½-7, o SON-R 2½-7[a] e o SON-R 6-40. O SON-R 2½-7 avalia as habilidades cognitivas de crianças na faixa etária entre 2 anos e meio e 7 anos. O SON-R 2½-7[a] é a versão abreviada do SON-R 2½-7 que já foi normatizado para a população brasileira e que recebeu parecer favorável do Conselho Federal de Psicologia em 2012. A pesquisa de normatização e de validação no Brasil ocorreu em 2008 com uma amostra composta por 1.200 crianças de todas as regiões do país e de diferentes extratos socioeconômico (Laros, Jesus & Karino, 2013). Já a pesquisa de normatização e validação do SON-R 6-40 para o contexto brasileiro está em fase de andamento. O plano amostral engloba as cinco regiões do país e até o momento o teste foi aplicado em cerca da metade dos 1.360 participantes planejados.

Todas as versões dos testes SON têm a possiblidade de avaliar a inteligência geral sem envolver o uso da linguagem falada ou escrita. Essa característica dos testes

SON os torna muito adequados para a realidade brasileira onde ainda existe uma percentagem considerável de pessoas com dificuldades com a linguagem falada e escrita. A possibilidade de aplicar os testes SON sem uso de linguagem falada ou escrita também tornam os testes adequados para a avaliação de crianças com algum tipo de deficiência, por exemplo, crianças surdas, autistas (Mecca et al., 2013). Outras características dos testes SON que tornam estes testes uma alternativa atrativa na área de avaliação psicológica: o cuidado dos construtores dos testes no delineamento da amostra de normatização, a exclusão de itens com características psicométricas duvidosas ou com DIF (*Differential Item Functioning)* (Karino, Laros & Jesus, 2012), a atualização constante do material dos subtestes tornando-os atrativos (Tellegen & Laros, 2014), a diversificação das tarefas, o oferecimento de exemplos antes do início da testagem, a inclusão de um procedimento adaptativo e o fornecimento de *feedback* após a resolução de cada item.

O foco de mensuração do teste SON-R 6-40 é a inteligência fluida (Cattell, 1963), que pode ser definida como a capacidade para resolver problemas para os quais a pessoa tem pouco conhecimento prévio (Laros, Valentini, Gomes & Andrade, 2014). As operações mentais que exigem a formação e o reconhecimento de conceitos, a identificação de relações complexas, a compreensão de implicações e a realização de inferências representam a capacidade fluida (Carroll, 1993; Schelini, 2006). Estudos apontam que a carga fatorial da inteligência fluida (*Gf*) sobre o fator geral (*g*) poderia demonstrar uma unidade, o que implica em entender o fator *g* como equivalente à *Gf* (Schelini, 2006).

Os modelos teóricos sobre a inteligência foram aprimorados com a contribuição de Horn e Carroll, sendo que hoje a Teoria Cattell-Horn-Carroll (CHC), proposta por McGrew e Flanagan (1998), vem sendo utilizada na revisão de tradicionais testes de

inteligência e embasando a construção de novos instrumentos (Schelini, 2006). Um exemplo dessa mudança foi a revisão da terceira versão da Escala de Inteligência Wechsler para Crianças (WISC) que sofreu grandes mudanças para tentar se adequar mais ao modelo CHC. O WISC-IV passou por uma série de modificações, incluindo alterações no conteúdo dos subtestes, mudança de terminologia do QI Verbal e do QI de Execução e inclusão de três novos subtestes para medir a habilidade de raciocínio fluido, devido à ênfase que as teorias do funcionamento cognitivo atribuem à avaliação desse tipo de raciocínio. É importante ressaltar que uma adequação completa ao modelo CHC seria praticamente impossível uma vez que a teoria diferencia 16 dimensões amplas (Laros, Valentini, Gomes & Andrade, 2014). Considerando que cada dimensão deveria ser representada por pelo menos dois subtestes, resultaria numa bateria de pelo menos 32 subtestes.

Há estudos de evidências de validade convergente dos testes SON com as escalas Wechsler em países como Holanda (Faber, 2010) e Alemanha (Tellegen & Laros, 2014). No Brasil, foi realizado um estudo de validade convergente do SON-R 2½-7[a] com o WISC-III e o WPPSI-III (Karino, Laros & Jesus, 2011). As correlações do QI total do SON-R 2½-7[a] com as escalas as escalas completas do WISC-III e WPSSI-III foram respectivamente, 0,69 e 0,75. As correlações entre o QI total do SON-R 2½-7[a] com as escalas verbais WISC-III e WPSSI-III foram 0,52 e 0,66 e com as escalas de execução dos dois testes Wechsler 0,65 e 0,73. Os resultados encontrados nesse estudo revelam evidências de validade convergente do teste SON-R 2½-7[a] com o WISC-III e o WPPSI-III, corroborando com o que era esperado: menores correlações entre o SON-R 2½-7[a] com as escalas verbais dos dois testes Wechsler e maiores correlações entre o SON-R 2½-7[a] e as escalas de execução e escala geral dos testes Wechsler.

No Brasil, o WISC foi traduzido em 1964, sem nenhum estudo de aplicação em amostra brasileira. Hoje, são utilizadas duas versões para a avaliação da inteligência de crianças e adolescentes, o WISC-III e o WISC-IV. A última versão, que foi publicada em 2013, tem como objetivo avaliar a capacidade intelectual e o processo de resolução de problemas em crianças de 6 anos a 16 anos e 11 meses. Nesta versão, além do QI total, há quatro índices que podem ser mensurados: Índice de Compreensão Verbal (ICV), destinado para aferição das habilidades verbais; Índice de Organização Perceptual (IOP), que mede a organização perceptual; Índice de Memória Operacional (IMO), que analisa a atenção, concentração e a memória operacional e o Índice de Velocidade de Processamento (IVP), mede a agilidade mental e o processamento grafomotor (Wechsler, 2013).

Como as baterias Wechsler são largamente utilizadas para avaliação da inteligência de crianças e adultos em diferentes contextos (Figueiredo, Mattos, Pasquali & Freire, 2008; Mayes & Calhoun, 2008; Fiorello et al., 2007; Waber et al., 2006), selecionou-se a escala WISC-IV para realização de um estudo de validade convergente da bateria SON-R 6-40. Na Alemanha, os testes SON-R 6-40 e o WISC-IV foram aplicados em 35 participantes, com idade variando de 6 a 15 anos (Tellegen & Laros, 2014). A correlação encontrada entre o QI total do SON e o QI total do WISC-IV foi de 0,77. A correlação entre o Índice de Organização Perceptual (IOP) e o SON-R 6-40 foi de r = 0,74. O IOP é formado pelos subtestes Cubos, Conceitos Figurativos, Raciocínio Matricial e Completar Figuras, utilizados para a avaliação do raciocínio fluido. A correlação do SON-R 6-40 com o Índice de Compreensão Verbal foi de r = 0,67 e com o Índice de Memória Operacional foi de r = 0,65. Como esperado, a correlação encontrada entre o SON-R 6-40 e o Índice de Velocidade de Processamento foi relativamente baixa (r = 0,43). Nesse contexto, a presente pesquisa pretende obter

evidências de validade convergente dos escores obtidos no teste SON-R 6-40 com o WISC-IV.

Método

Participantes

Os participantes desta amostra foram 120 crianças (61 meninas), residentes de diferentes regiões administrativas do Distrito Federal, matriculadas em duas escolas públicas (N = 80) e em uma escola particular (N = 40). A idade das crianças variou de 10 a 14 anos. O critério de seleção adotado para a triagem das crianças foi ter entre 10 a 14 anos de idade.

Instrumentos

Foram utilizados dois instrumentos na pesquisa: o SON-R 6-40 e o WISC-IV. *SON-R 6-40* (Tellegen & Laros, 2014) é a última versão dos testes SON e seu público alvo são pessoas de 6 a 40 anos de idade. O teste avalia um espectro amplo das habilidades cognitivas sem envolver o uso da linguagem verbal ou escrita, pois as instruções podem ser dadas tanto de forma verbal quanto não verbal. Assim, a avaliação é realizada por meio de tarefas que não exigem qualquer tipo de explicação verbal ou nomeação de figuras. É um teste de aplicação individual, possui um procedimento adaptativo de aplicação e regras para interromper o teste limitam a aplicação de itens que são difíceis ou fáceis demais para a pessoa. É composto por quatro subtestes: Analogias (36 itens), Mosaicos (26 itens), Categorias (36 itens) e Padrões (26 itens). Os subtestes são aplicados nessa ordem, sendo que Analogias e Categorias avaliam raciocínio abstrato e os subtestes Mosaicos e Padrões avaliam o raciocínio espacial. Exemplos dos itens dos quatro subtestes podem ser encontrados no *website* dos testes SON (www.testresearch.nl).

WISC-IV - Escala Wechsler de Inteligência para Crianças - 4ª edição – (Wechsler, 2013) é um teste de aplicação individual e foi desenvolvido para avaliação de crianças e adolescentes a partir dos 6 anos de idade até os 16 anos e 11 meses. É composto por 15 subtestes, sendo que os subtestes são identificados como principais (10) e suplementares (5). É necessário aplicar os 10 subtestes principais para obter os cinco escores compostos do WISC-IV (QI Total e quatro índices).

Procedimentos

Após as unidades escolares concordarem em participar da pesquisa, cedendo seu espaço e tempo, foram enviadas cartas aos pais explicando os objetivos da pesquisa, os procedimentos e termos de consentimento. Aqueles que concordavam com a participação do seu filho ou tutelado deveriam devolver o termo assinado. No final da pesquisa, os pais receberam um relatório descrevendo o desempenho do filho na pesquisa e a escola recebeu relatório geral descrevendo como foi o desempenho das crianças e sugestões de atividades que podem ser desenvolvidas.

As aplicações foram realizadas com o apoio de uma equipe de alunas de graduação em psicologia que receberam um treinamento para padronizar o procedimento de aplicação dos testes. Foram observados os procedimentos de aplicação e avaliação dos subtestes descritos no manual de cada teste. Todas as aplicações foram individuais e os testes foram aplicados nas escolas, durante o horário das aulas. A aplicação dos testes ocorreu da seguinte forma: sessenta crianças responderam primeiro o teste SON-R 6-40 e depois o teste WISC-IV; a segunda metade da amostra (N = 60) respondeu primeiro o teste WISC-IV e depois o teste SON-R 6-40. Os testes foram aplicados em duas sessões e o intervalo de tempo entre a aplicação dos testes girou em torno de 3 a 4 semanas.

O tempo de aplicação do teste WISC-IV girou em torno de uma hora e 20 minutos e foram aplicados apenas os subtestes principais, a saber: Cubos, Semelhanças, Dígitos, Conceitos Figurativos, Código, Vocabulário, Sequência de Números e Letras, Raciocínio Matricial, Compreensão e Procurar Símbolos. O tempo de aplicação do teste SON-R 6-40 girou em torno de 50 minutos e foram aplicados todos os subtestes.

Análise dos dados

Inicialmente foram realizadas as transformações dos escores brutos em escores normatizados levando em consideração as informações presentes nos manuais de cada teste. É necessário utilizar escores normatizados em vez de escores brutos uma vez que nos escores brutos ainda existe variância compartilhada com a variável idade o que resultará em uma superestimação da correlação entre dois testes. É importante destacar que para transformar os escores brutos do SON-R 6-40 em escores normatizados as normas da Holanda /Alemanha (N = 1.933) foram utilizadas, uma vez que ainda não existem normas para o Brasil. Os escores normatizados dos subtestes do SON-R 6-40 estão numa escala com M = 10 e DP = 3. Para o escore total (o QI-SON) a escala é M = 100 e DP =15. Os escores normatizados dos subtestes do WISC-IV usam também a escala M =10 e DP = 3. Para os índices e escore total do WISC-IV a escala M =100 e DP = 15 é utilizada.

O coeficiente Lambda 2 de Guttman foi utilizado para estimar a fidedignidade dos escores dos testes. Esse coeficiente foi escolhido porque estudos apontam que esse índice é um dos índices mais adequados para estimar a fidedignidade dos escores, principalmente quando a amostra é pequena (Tellegen & Laros, 2004; Sijtsma, 2009; 2012).

Testes de normalidade, como o teste Shapiro-Wilk (Field, 2009), foram realizados e os valores da *skewness* (Miles & Shevlin, 2001) foram observados para

avaliar e assegurar os pressupostos de normalidade dos dados. Para a realização das análises de validade convergente do SON-R 6-40, utilizou-se a correlação bivariada de Pearson. Foram utilizados os escores normatizados para estimar a correlação entre os testes. Além das correlações brutas, foram calculadas as correlações corrigidas. Para tanto, utilizou-se a correção para falta de fidedignidade e para falta de variância (Hogan, 2006; Osborne, 2003; Thompson, 2003), cuja equação é: $r_u = [r_c (S_u/ S_c)] \div [\sqrt{} r_{xx}.r_{yy} - r_c^2 + r_c^2.(S_u^2/S_c^2)]$, na qual r_u é a correlação na amostra não-restrita, r_c é a correlação na amostra restrita, S_u é o desvio-padrão na amostra não-restrita, S_c é desvio-padrão na amostra restrita e r_{xx} e r_{yy} são os coeficientes de fidedignidade dos dois testes. Hogan (2006), Osborne (2003) e Thompson (2003) argumentam que as relações reais entre variáveis podem ser subestimadas caso uma parte da variância é variância de erro e quando a variância na amostra pesquisada é restrita. Segundo esses autores, pesquisadores precisam corrigir para atenuação no intuito de obter uma estimativa melhor da relação verdadeira entre as variáveis na população.

Resultados e Discussão

O primeiro passo foi calcular as estatísticas descritivas para os dois testes utilizados. A Tabela 1 apresenta os valores da média, desvio-padrão, erro padrão da média, o intervalo de confiança de 95% das médias e a amplitude dos escores normatizados de cada subteste e do QI do SON-R 6-40 e de cada subteste do WISC-IV, dos quatro índices e do QI Total do WISC-IV.

Tabela 1. *Estatísticas descritivas dos escores normatizados do SON-R 6-40 e WISC-IV.*

	SON-R 6-40						
Subtestes	M		DP	EP	IC 95%	Mín.	Máx.
Analogias	10,01	(6,68)	1,99	0,18	9,66 – 10,36	4	13
Mosaicos	9,98	(6,65)	2,22	0,20	9,59 – 10,37	2	12
Categorias	10,56	(7,23)	2,47	0,22	10,13 – 10,99	2	14
Padrões	10,18	(6,85)	1,99	0,18	9,83 – 10,53	2	15
QI-SON	98,30	(81,65)	8,56	0,78	96,77 – 99,82	65	111

	WISC-IV					
Subtestes	M	DP	EP	IC 95%	Mín.	Máx.
Cubos	9,71	2,63	0,24	9,24 – 10,18	3	16
Semelhanças	9,38	2,80	0,26	8,88 – 9,88	3	17
Dígitos	9,79	2,99	0,27	9,26 – 10,32	2	18
Conceitos Figurativos	9,27	2,71	0,25	8,79 – 9,75	1	14
Código B	9,90	2,55	0,23	9,45 – 10,35	4	19
Vocabulário	8,46	3,10	0,28	7,91 – 9,01	1	16
Seq. de Números e Letras	8,55	2,90	0,26	8,04 – 9,06	3	19
Raciocínio Matricial	9,37	2,64	0,24	8,90 – 9,84	4	17
Compreensão	8,41	3,07	0,28	7,87 – 8,95	1	15
Procurar Símbolos B	11,17	2,17	0,20	10,79 – 11,55	4	18
Índices	M	DP	EP	IC 95%	Mín.	Máx.
Índice de Compreensão Verbal	92,63	15,09	1,37	89,95 – 95,31	55	130
Índice de Organização Perceptual	96,60	12,46	1,13	94,39 – 98,81	69	126
Índice de Memória Operacional	95,18	14,65	1,33	92,58 – 97,78	65	138
Índice de Velocidade de Processamento	103,11	11,85	1,08	98,98 – 107,24	64	147
QI Total do WISC-IV	95,38	13,15	1,20	93,03 – 97,73	61	131

Notas. M = média; DP = desvio padrão; EP = erro padrão da média; IC 95% = intervalo de confiança de 95% da média. As médias dos escores normatizados do SON-R 6-40 entre parênteses são os escores observados em base das normas da Holanda / Alemanha. Ao lado encontram-se as médias corrigidas.

Os resultados da Tabela 1 mostram que as médias dos escores normatizados dos subtestes e do escore total do SON-R 6-40 (apresentadas entre parênteses) são bastante inferiores aos valores das médias (M = 10) na amostra normativa da Holanda /Alemanha. Esse resultado foi esperado uma vez que na normatização do SON-R 2½-7[a] no Brasil a diferença observada entre as crianças brasileiras e holandesas na Escala Geral foi 16,7 pontos o que equivale a 1,11 desvio-padrão. Os escores normatizados corrigidos foram obtidos através de um aumento de 1,11 desvio-padrão nos valores das

médias brutas. No caso das médias dos subtestes isso resultou em um aumento de 3,33 pontos (1,11 x 3) e no caso do escore total em um aumento de 16,65 pontos (1,11 x 15).

A Tabela 1 mostra também que a variância dos escores normatizados dos subtestes do SON-R 6-40 é bem inferior da variância nesses escores na amostra normativa da Holanda /Alemanha. A mesma observação aplica-se a variância do escore total do SON-R 6-40.

Verificando os valores das médias dos escores normatizados dos subtestes do WISC-IV podemos observar que, com exceção do subteste Procurar símbolos B, os valores são inferiores aos valores da amostra normativa do WISC-IV. Em geral, a variância dos escores normatizados dos subtestes também são inferiores aos valores da amostra normativa do WISC-IV.

Em relação aos escores compostos do WISC-IV a Tabela 1 mostra que três dos quatro índices (ICV, IOP e IMO) e o QI Total do WISC-IV tem valores abaixo do valor da média (100) da amostra normativa do WISC-IV. Os valores do desvio-padrão de três índices (IOP, IMO e IVP) e do QI Total também são menores do que o desvio-padrão (15) na amostra normativa do WISC-IV. Essas observações indicam que, em geral, na amostra do estudo atual existe menos variabilidade nas habilidades cognitivas avaliadas em comparação com a amostra normativa do WISC-IV e que em comparação com a amostra normativa as médias no QI Total e nos três dos quatro índices do WISC-IV são mais baixas.

A Tabela 2 mostra os valores dos coeficientes de fidedignidade e a correlação média entre os itens de todos os subtestes (exceto dos subtestes Códigos e Procurar Símbolos B) e da escala geral. A fidedignidade dos subtestes Códigos e Procurar Símbolos B não foi estimada pelo método da consistência interna porque não é correto medir a fidedignidade de testes de velocidade com esse método (Karino, Laros & Jesus,

2011). No manual do WISC-IV a fidedignidade desses dois subtestes também não foi apresentada, apenas a correlação média entre itens foi informada.

Além disso, o manual do WAIS-III, o teste de inteligência Wechsler para adultos, afirma que não é correto estimar a fidedignidade dos subtestes de rapidez com o coeficiente das duas metades (Wechsler, 2011). Na pesquisa com o WAIS-III, a fidedignidade dos subtestes Códigos e Procurar Símbolos foi estimada a partir do método teste-reteste, que é a correlação entre os resultados da primeira e da segunda aplicação. Ressalta-se que alguns itens não foram incluídos no momento do cálculo da fidedignidade por não apresentarem variância, o que pode dificultar a estimação da fidedignidade.

Tabela 2. *Índices de fidedignidade do SON-R 6-40 e do WISC-IV.*

Teste	Subteste	n° de itens	Lambda 2	*r* média entre itens
SON-R 6-40	Analogias	36	0,81	0,12
	Mosaicos	26	0,85	0,20
	Categorias	36	0,84	0,14
	Padrões	26	0,83	0,17
	Escala Total	124	0,92	0,10
WISC-IV	Cubos	14	0,83	0,27
	Semelhanças	23	0,89	0,22
	Dígitos	32	0,84	0,17
	Conceitos Figurativos	28	0,81	0,12
	Código B	119	-----	0,26
	Vocabulário	36	0,88	0,17
	Seq. de Números e Letras	30	0,87	0,19
	Raciocínio Matricial	35	0,88	0,18
	Compreensão	21	0,85	0,20
	Procurar Símbolos B	60	-----	0,17
	Índice de Compreensão Verbal	71	0,94	0,16
	Índice de Organização Perceptual	67	0,91	0,10
	Índice de Memória Operacional	43	0,90	0,14
	Índice de Velocidade de Processamento	96	-----	0,17
	Escala Total	219	0,95	0,08

Nota. Os subtestes Código B e Procurar Símbolos B do WISC-IV não foram incluídos no cálculo da fidedignidade da escala total.

A inspeção da Tabela 2 revela que os coeficientes de fidedignidade dos escores nos subtestes do SON-R 6-40 tem valores entre 0,81 e 0,85. Os escores na escala total do SON-R 6-40 mostra um valor de 0,92 na amostra pesquisada.

Os escores nos subtestes do WISC-IV têm coeficientes de fidedignidade entre 0,81 e 0,89. Os escores nos índices mostram coeficientes de fidedignidade entre 0,90 e 0,94, enquanto os escores na escala total tem um coeficiente de fidedignidade de 0,95.

Foram calculadas correlações de *Pearson* a fim de buscar associações entre o QI do SON-R 6-40 com os índices e com o escore total do WISC-IV. A Tabela 3 apresenta as correlações brutas e as correlações depois da aplicação da fórmula de correção para atenuação e falta de variância. A correlação entre o QI do SON e o QI total do WISC-IV

corrigida foi de 0,73. Como apresentado na Tabela 3, observa-se que a correlação corrigida entre o QI do SON e o Índice de Organização Perceptual foi a correlação mais alta (0,84). Esse resultado corrobora com o esperado e pode ser explicado pela última revisão do WISC-IV de focalizar mais na avaliação da inteligência fluida. O manual relata que foram incorporados três novos subtestes para medir a habilidade de raciocínio fluido: Raciocínio Matricial, Conceitos Figurativos e Raciocínio com Palavras, sendo que o último não foi aplicado no presente estudo.

Tabela 3. *Correlações entre o SON-R 6-40 com os Índices do WISC-IV.*

Índices do WISC-IV	QI SON-R	IC 95%
Índice de Compreensão Verbal	0,48 (0,45)	0,33 – 0,61
Índice de Organização Perceptual	**0,84 (0,71)**	0,79 – 0,89
Índice de Memória Operacional	0,44 (0,39)	0,28 – 0,57
Índice de Velocidade de Processamento	0,32 (0,26)	0,15 – 0,47
WISC-IV QI Total	0,73 (0,63)	0,63 – 0,80

Nota. A correlação bruta é apresentada entre parênteses. Os intervalos de confiança de 95% foram calculados para as correlações corrigidas.

Revisões recentes de outras escalas Wechsler de Inteligência, por exemplo, WAIS-III e WPPSI-III, também introduziram novos subtestes para aprimorar as medidas de raciocínio fluido. Um estudo realizado no Brasil com o SON-R 2½-7[a], também encontrou correlações mais altas entre as escalas de execução do que com as escalas verbais (Karino, Laros & Jesus, 2011). Diversas teorias do funcionamento cognitivo enfatizam a importância do raciocínio fluido (Carroll, 1993; Sternberg, 1995) e as tarefas que requerem esse tipo de habilidade estão ligadas à manipulação de abstrações, regras, generalizações e relacionamentos lógicos (Carroll, 1993).

Por fim, a Tabela 4 apresenta os valores encontrados das correlações entre os subtestes que compõem cada um dos testes. Os valores indicam maior correlação entre

os subtestes do SON-R 6-40 e os subtestes Cubos e Raciocínio Matricial do WISC-IV. Os subtestes Cubos e Raciocínio Matricial compõem o Índice de Organização Perceptual, índice que é destinado para avaliação da inteligência fluida no WISC-IV.

Como pode ser observada, a maior correlação foi entre o subtestes Mosaicos e Cubos. Tal resultado era esperado devido à similaridade das tarefas desses subtestes: a criança precisa reproduzir padrões que são apresentados utilizando peças ou cubos coloridos que lhe são oferecidas.

Tabela 4. *Correlações entre os subtestes do SON-R 6-40 com os subtestes do WISC-IV.*

	SON-R 6-40			
WISC-IV	Analogias	Mosaicos	Categorias	Padrões
Cubos	0,58 (0,44)	**0,73 (0,57)**	0,49 (0,37)	**0,62 (0,47)**
Semelhanças	0,35 (0,28)	0,38 (0,31)	0,44 (0,37)	0,19 (0,15)
Dígitos	0,35 (0,29)	0,33 (0,28)	0,27 (0,23)	0,24 (0,20)
Conceitos Figurativos	0,51 (0,39)	0,50 (0,38)	0,50 (0,38)	0,37 (0,28)
Código B	0,14 (0,11)	0,24 (0,21)	0,25 (0,21)	0,24 (0,20)
Vocabulário	0,50 (0,42)	0,25 (0,22)	0,36 (0,31)	0,22 (0,19)
Seq. Números e Letras	0,33 (0,27)	0,16 (0,13)	0,23 (0,19)	0,08 (0,07)
Raciocínio Matricial	**0,77 (0,54)**	0,60 (0,47)	**0,59 (0,46)**	0,41 (0,32)
Compreensão	0,35 (0,29)	0,20 (0,17)	0,24 (0,20)	0,18 (0,15)
Procurar Símbolos B	0,28 (0,21)	0,30 (0,22)	0,37 (0,27)	0,33 (0,24)

Nota. As correlações brutas são apresentadas entre parênteses: as correlações corrigidas para atenuação são apresentadas ao lado.

Considerações finais

O objetivo principal deste estudo foi adquirir evidências de validade convergente dos escores obtidos no teste SON-R 6-40 com o WISC-IV. Os resultados indicam evidências positivas de validade convergente do SON-R 6-40 para crianças entre 10 e 14 anos de idade. A fidedignidade dos escores nos subtestes do SON-R 6-40 foi satisfatória: o coeficiente variou de 0,81 a 085: para o escore total o coeficiente da fidedignidade foi 0,92.

Vários estudos foram realizados na Holanda e Alemanha com o SON-R 6-40 e outros testes de inteligência, tais como WISC-III, WISC-IV, WAIS-III, WNV (Wechsler & Naglieri, 2008) e NIO (van Dijk & Tellegen, 2004) (Tellegen & Laros, 2014). De forma geral, os resultados relatados aqui apresentam tendência semelhante aos resultados encontrados na pesquisa realizada na Alemanha com o SON-R 6-40 e o WISC-IV.

Em relação à média das crianças brasileiras ser inferior à média das crianças holandesas, os testes educacionais internacionais revelam também o padrão apresentado aqui. O Programa Internacional de Avaliação de Estudantes (PISA) é uma avaliação comparada aplicada a estudantes na faixa etária de 15 anos – idade que se pressupõe o término da educação básica na maioria dos países. A avaliação do PISA acontece a cada três anos e abrange três áreas do conhecimento: leitura, matemática e ciência – sendo que em cada edição da prova, há maior ênfase em cada uma dessas áreas. A última avaliação foi realizada em 2012, participaram 18.589 estudantes brasileiros, a ênfase da avaliação foi em matemática e a média brasileira ficou abaixo da média da OECD – Organização de Cooperação e de Desenvolvimento Econômico - e abaixo das médias de países como Argentina, Chile, México, Uruguai, Costa Rica, ficando acima de países como Colômbia e Peru (OECD, 2013).

No Brasil, já foram realizados estudos de evidências de validade convergente do SON-R 6-40 com outros testes de inteligência, tais como a BPR-5 e SON-R 2½-7[a] (Laros, Almeida, Lima, & Valentini, no prelo). Assim, é importante a realização de estudos futuros com outros instrumentos que são utilizados no Brasil ou no exterior, tais como a Escala de Inteligência Stanford-Binet 5 (Roid, 2003) e a Bateria de Avaliação Kaufman para Crianças – Segunda Edição (Kaufman & Kaufman, 2004).

Limitações deste estudo recaem sobre a faixa limitada de idade das crianças que responderam o teste e ao fato da amostra contar com um número restrito de participantes, estudantes de duas escolas públicas e uma escola particular do Distrito Federal. É importante enfatizar a necessidade da realização de diversos estudos para aferir a qualidade psicométrica dos escores dos instrumentos. A realização de mais pesquisas visando comparar o desempenho de crianças, jovens e adultos em diferentes contextos culturais, com amostras maiores para faixa etária que o teste SON-R 6-40 contempla e de diferentes regiões do país é desejável devido a grande dimensão geográfica no Brasil e para que se busque investigar os resultados aqui obtidos.

Além disso, não foram incluídas na amostra crianças com evidências de deficiências intelectuais, auditivas ou motoras graves. Embora o SON-R se apresente como instrumento relevante para pesquisas com sujeitos que apresentam algum tipo de deficiência, este estudo teve por objetivo avaliar uma amostra de crianças sem prejuízo no desenvolvimento para que estudos comparativos possam ser realizados posteriormente.

Considera-se, diante do exposto, que o presente estudo alcançou seus objetivos e demonstrou adequadas evidências de validade convergente do teste SON-R 6-40.

Referências

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria de resposta ao item: conceitos e aplicações.* São Paulo: ABE – Associação Brasileira de Estatística.

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

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1-22.

Faber, H. H. (2010). *Valideringsonderzoek SON-R 6-40: Samenhang met de WISC-III-NL.* Heymans Instituut, RUG: Intern Verslag.

Field, A. (2009). *Discovering statistics using SPSS.* London: Sage Publications.

Figueiredo, V. L. M., Mattos, V. L. D., Pasquali, L., & Freire, A. P. (2008). Propriedades psicométricas dos itens do teste WISC-III. *Psicologia em Estudo, 13,* 585-592.

Fiorello, C. A., Hale, J. B., Holdnack, J. A., Kavanagh, J. A., Terrell, J., & Long, L. (2007). Interpreting intelligence test results for children with disabilities: Is global intelligence relevant? *Applied Neuropsychology, 14,* 2-12.

Karino, C. A., Laros, J. A., & Jesus, G. R. (2011). Evidências de validade convergente do SON-R 2½-7[a] com o WPPSI-III e WISC-III. *Psicologia: Reflexão e Crítica, 24,* 621-629.

Karino, C. A., Laros, J. A., & Jesus, G. R. (2012). Funcionamento diferencial dos itens do teste não-verbal de inteligência SON-R 2½-7[a]. *Psicologia: Teoria e Pesquisa, 28,* 15-25.

Kaufman, A. S., & Kaufman, N. L. (2004). *Manual for the Kaufman Assessment Battery for Children – Second Edition (KABC-II).* Circles Pines, MN: American Guidance Service.

Laros, J. A., Jesus, G. R., & Karino, C. A. (2013). Validação brasileira do teste não verbal de inteligência SON-R 2½-7[a]. *Avaliação Psicológica,12* ,233-242.

Laros, J. A., Valentini, F., Gomes, C. M. A., & Andrade, J. M. (2014). Modelos de inteligência. In A. G. Seabra, J. A. Laros, E. C. Macedo & N. Abreu (Eds.), *Inteligência e funções cognitivas: avanços e desafios para a avaliação psicológica* (pp. 17-38). São Paulo: Memnon.

Laros, J. A., Almeida, G. O. N., Lima, R. M. F., & Valentini, F. (no prelo). Dimensionalidade e evidências de validade convergente do SON-R 6-40. *Temas em Psicologia, 23*(4).

Laros, J. A., Tellegen, P. J., Jesus, G. R., & Karino, C. A. (in press). *SON-R 2½-7[a], Teste não-verbal de inteligência. Manual de normatização e validação brasileira.*

Mayes, S. D., & Calhoun, S. L. (2008). WISC-IV and WIAT-II profiles in children with high-functioning austism. *Journal of Autism and Developmental Disorders, 38,* 428-439.

McGrew, K. S., & Flanagan, D. P. (1998). *The Intelligence Test Desk Reference (ITDR) – Gc-Gf Cross Battery Assessment.* Boston, MA: Allyn and Bacon.

Mecca, T. P., Valentini, F., Laros, J. A., Lima, R. M. F., Schwartzman, J. S., & Macedo, E. C. (2013). Utilizando o teste não verbal de inteligência SON-R 2½-7[a] para avaliar crianças com Transtornos do Espectro do Autismo. *Revista Educação Especial, 26,* 603-618.

Miles, J., & Shevlin, M. (2001). *Applying regression & correlation. A guide for students and researchers.* London: Sage Publications.

OECD (2013). *PISA 2012 results: What students know and can do - student performance in Mathematics, Reading and Science (Volume I)*, PISA, OECD Publishing.

Pasquali, L (2010). *Instrumentação psicológica. Fundamentos e práticas.* Porto Alegre: Artmed.

Primi, R. (2003). Inteligência: avanços nos modelos teóricos e nos instrumentos de medida. *Avaliação Psicológica, 1,* 67-77.

Reppold, C. T., Gurgel, L. G., & Hutz, C. S. (2014). O processo de construção de escalas psicométricas. *Avaliação Psicológica, 13,* 307-310.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition: Examiner's manual.* Austin, TX: Pro-Ed.

Schelini, P. W. (2006). Teoria das inteligências fluida e cristalizada: início e evolução. *Estudos de Psicologia, 11,* 323-332.

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika, 74,*107-120.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika,74,*107-120.

Tellegen, P. J., & Laros, J. A. (2004). Cultural bias in the SON-R test: Comparative study of Brazilian and Dutch children. *Psicologia: Teoria e Pesquisa, 20,* 103-111.

Tellegen, P. J., & Laros, J. A. (2014). *SON-R 6-40. Non-verbal intelligence test: Research report.* Göttingen, Germany: Hogrefe Verlag.

Waber, D. P., Gerber, E. B., Turcios, V. Y., Wagner, E. R., & Forbes, P. W. (2006). Executive functions and performance on high-stakes testing in children from urban schools. *Developmental Neuropsychology, 29,* 459-477.

Wechsler, D. (2011). *WAIS-III – Escala de inteligência Wechsler para adultos. Manual técnico.* São Paulo: Casa do Psicólogo.

Wechsler, D. (2013). *WISC-IV - Escala Wechsler de inteligência para crianças. Manual técnico.* São Paulo: Casa do Psicólogo.