Estudo de fidedignidade do avaliador em provas de compreensão leitora e oral[1]

Patrícia Silva Lúcio[2]
Universidade Estadual de Londrina, Londrina-PR, Brasil
Adriana de Souza Batista Kida, Carolina Alves Ferreira de Carvalho, Hugo Cogo-Moreira, Clara Regina Brandão de Ávila
Universidade Federal de São Paulo, São Paulo-SP

RESUMO

A fidedignidade do avaliador refere-se ao grau em que diferentes avaliadores empregam os mesmos critérios na correção de testes. Neste estudo, investiga-se a fidedignidade do avaliador de uma prova de compreensão leitora e outra de compreensão oral. A Prova de Compreensão Leitora (PCL) é composta por 15 textos, divididos em duas aplicações (Forma A e Forma B), e contendo questões abertas (respondidas oralmente). A Prova de Compreensão Oral (PCO) é formada por oito questões orais a uma narrativa ouvida (gravação). A amostra foi composta por 218 crianças do 2º ao 5º ano de escolas públicas de São Paulo (73 na Forma A; 56 na Forma B; 89 na PCO) e nove avaliadores (três para cada prova). A concordância foi calculada por meio do Fleiss de Kappa. A maioria dos itens apresentou níveis de concordância adequados, atestando para a homogeneidade na correção das duas Provas.

Palavras-chave: fidedignidade; compreensão da leitura; compreensão verbal; texto.

ABSTRACT – Inter-rater reliability investigation for reading and oral comprehension tasks

Inter-rater reliability is the degree to which different examiners employ the same criteria for evaluating test results. The present study aims to investigate the inter-rater reliability for two tasks, one of reading comprehension and another for oral comprehension. The Reading Comprehension Task consists of 15 texts divided into two blocks (Form A and Form B), followed by orally answered open-ended questions. The Oral Comprehension Task (OCT) contains eight open-ended questions for a narrative (orally answered and recorded). The sample consisted of 218 children from 2nd to 5th year of elementary school in Sao Paulo (73 in Form A; 56 in Form B; 89 in OCT) and nine examiners (three for each form). Fleiss' Kappa was used to obtain the reliability index. Most items had adequate levels of agreement, which evidenced the consistency of the correction system.

Keywords: reliability; reading comprehension; verbal comprehension; text.

RESUMEN – Investigación de la fiabilidad entre evaluadores en pruebas de comprensión lectora y oral

La fiabilidad del evaluador se refiere al grado en que los diferentes evaluadores emplean los mismos criterios en la corrección de los testes. En este estudio, se investiga la fiabilidad del evaluador en prueba de comprensión lectora y otra de comprensión oral. La prueba de comprensión lectora (PCL) consta de 15 textos, divididos en dos bloques de aplicación (Formulario A y Formulario B) y contienen preguntas abiertas relacionadas. La Prueba de Comprensión Oral (PCO) consta de ocho cuestiones orales sobre una narrativa oída (grabación). La muestra fue compuesta por 218 niños de segundo a quinto primaria de escuelas públicas de San Pablo (73 en el Formulario A; 56 en el Formulario B; 89 en la PCO) y nueve evaluadores (tres para cada prueba). Se calculó el acuerdo con Fleiss de Kappa. La mayoría de los ítems presentó niveles adecuados de concordancia, indicando homogeneidad en la corrección de las dos pruebas.

Palabras clave: fiabilidad; comprensión lectora; comprensión verbal; texto.

Todo processo de avaliação efetuado por meio de testes está sujeito a erros, os quais podem ser estimados a partir de investigações propostas pela psicometria. Essas estimativas fornecem índices ou valores psicométricos que nos indicam o grau em que escores dos indivíduos nos testes são consistentes ou isentos de flutuações. A investigação do erro de medida nos testes psicológicos é feita a partir dos índices de fidedignidade, que indicam o grau de confiança que se pode ter nas interpretações que se pretende extrair dos escores, atestando se estes estão livres de erros ou de fontes de vieses (Chadha, 2009; Urbina, 2014).

[1] Agradecemos o apoio financeiro da Fundação de Amparo à Pesquisa de São Paulo (FAPESP) para a realização da pesquisa, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão de bolsa de doutorado à primeira autora e à Universidade Estadual de Londrina pela concessão de bolsa de iniciação científica a alunos envolvidos no projeto.

[2] Endereço para correspondência: Universidade Estadual de Londrina, Departamento de Psicologia e Psicanálise, Rodovia Celso Garcia Cid, PR445, Km 380, Campus Universitário, Caixa Postal 10.011, 86057-970, Londrina-PR. Tel.: (43) 3371-9347. *E-mail:* pslucio@gmail.com

Dentro do paradigma da estatística clássica, existem basicamente duas fontes de erros sobre os testes: o erro sistemático (que afeta os escores dos testes de uma maneira consistente) e o erro aleatório (que pode distanciar o escore observado do escore verdadeiro do indivíduo em qualquer direção, para baixo ou para cima). Ambos os tipos de erro não estão relacionados ao construto considerado, mas apenas o primeiro é investigado pelas medidas de fidedignidade (isso porque as variações produzidas pelo erro aleatório tendem a zero, sendo controladas por amostras amplas, aleatórias e representativas). Desse modo, os índices de fidedignidade indicam o grau em que os escores observados (desempenho dos sujeitos no teste) se afastam do escore verdadeiro dos indivíduos (ou nível de habilidade), sendo, portanto, um índice de confiança ou confiabilidade (Raykov & Marcoulides, 2011).

Os erros sistemáticos podem ser devidos a diversas fontes: quando os mesmos indivíduos são avaliados em dois momentos distintos, pode haver diferenças entre a primeira e a segunda avaliação referentes à passagem do tempo (devido a mudanças no desenvolvimento, efeito de treino ou alterações nas condições de testagem, por exemplo). Outras fontes de erro estão dentro dos próprios testes. Os conteúdos (ou itens) dos testes, quando não são devidamente amostrados, podem gerar erros na avaliação dos sujeitos (por avaliar um construto diferente do pretendido, ou por simplesmente amostrar mais um conteúdo em relação ao outro, etc.). Como existem diferentes fontes de erros nos testes, estas devem ser avaliadas por meio de investigações distintas que fornecerão evidências a partir das quais é possível fazer um julgamento sobre a qualidade da consistência das informações produzidas pelo instrumento (AERA, APA, & NMCE, & 2014).

Em testes compostos por questões abertas, existe uma dificuldade na avaliação das produções dos indivíduos. Isso porque há uma grande variabilidade nas respostas individuais, as quais devem ser classificadas pelos avaliadores a partir de critérios preestabelecidos. A natureza do formato das questões abertas faz com que um elemento de subjetividade permeie a avaliação dos resultados dos testes. Esse tipo de fonte de erro é investigado, dentro da psicometria, pela fidedignidade do avaliador e se refere ao grau em que diferentes avaliadores utilizam os mesmos critérios para atribuir escores às respostas dos avaliandos (McHugh, 2012; Urbina, 2014).

Segundo Hallgren (2012), a concordância entre avaliadores é fundamental para pesquisas que utilizam algum tipo de classificação para conferir escores aos participantes, dela dependendo a confiança que se tem nos resultados produzidos. O autor ainda chama a atenção para problemas relacionados ao estudo da concordância entre avaliadores que frequentemente encontram-se em pesquisas, os quais são a escolha incorreta da análise estatística, interpretações equivocadas de seus resultados e a ausência de relatos dos intervalos de confiança. Por fim, Hallgren (2012) ressalta o problema de se desconsiderar as implicações dos índices de fidedignidade obtidos para o poder estatístico dos estudos que utilizam instrumentos que produzem escores dependentes da classificação realizada por avaliadores.

Desse modo, ao se estimar a fidedignidade do avaliador, o pesquisador deve estar atento para as características do seu objeto de estudo, devendo-se levar em conta a natureza do delineamento, o nível da medida (nominal, ordinal, escalar ou razão) e o número de observadores ou juízes utilizados para o cômputo. Nesse sentido, Hallgren (2012) destaca três aspectos a serem considerados: 1. se todos os participantes do estudo ou se apenas um subconjunto será avaliado pelos juízes; 2. se os participantes serão avaliados pelos mesmos avaliadores, no chamado delineamento cruzado total (*fully crossed design*) ou se diferentes sujeitos são categorizados por distintos avaliadores (*not fully crossed design*); e 3. as propriedades psicométricas do sistema de codificação utilizado (que afeta a variabilidade dos resultados e, portanto, pode gerar baixos índices de concordância mesmo com níveis de erro pequenos).

A obtenção de bons índices de fidedignidade do avaliador depende de fatores como a qualidade do treinamento dado aos juízes, o nível de detalhamento e clareza dos crivos de resposta, além da habilidade do aplicador da tarefa em extrair respostas significativas e completas dos avaliandos (é o caso, por exemplo, das respostas que exigem inquérito nas Escalas Wechsler de Inteligência). Mantida a qualidade de todos esses fatores, ainda assim, é essencial investigar os níveis de concordância em testes cuja subjetividade do avaliador possa interferir nos escores dos indivíduos (Urbina, 2014).

O presente trabalho constitui um estudo de investigação da fidedignidade do avaliador de duas tarefas de compreensão, uma leitora e outra oral, destinadas à avaliação de crianças em fases iniciais de escolarização. Ambas as provas estão em fase de construção e se caracterizam pela presença de questões abertas que devem ser respondidas a textos (no caso da tarefa de compreensão leitora, os textos são lidos e, na compreensão oral, é apresentada a gravação do texto). Nos dois casos, a investigação da fidedignidade do avaliador é essencial, pois a característica das provas implica que certo grau de subjetividade pode interferir na conferência dos escores.

A relevância do presente estudo se justifica principalmente por duas razões. Em primeiro lugar, porque, no contexto brasileiro, há um número limitado de instrumentos de investigação da compreensão leitora (p. ex., Cuetos, Rodrigues, & Ruano, 2010; Joly & Istome,

2008; Santos, Primi, Taxa, & Vendramini, 2002)[3] e oral (p. ex., Radanovic, Mansur, & Scaff, 2004; Ortiz, Ozborn, & Chiari, 1993). Nesse sentido, fica evidente a relevância de trabalhos que busquem construir ferramentas de avaliação nas áreas citadas, seguindo-se os parâmetros da psicometria. Em segundo lugar, para nenhum desses instrumentos cujos escores são passiveis de sofrer interferência da subjetividade do avaliador foram encontrados estudos de fidedignidade do avaliador[4]. Isso aponta para uma preocupação com a qualidade dos instrumentos de avaliação educacional e poderá servir de inspiração a outros pesquisadores que investigam instrumentos com características semelhantes aos que aqui são apresentados.

Método

Instrumentos

A Prova de Compreensão Leitora (PCL) é um instrumento de avaliação educacional destinado ao público infantil e que se encontra em fase de construção. O referido instrumento é composto por 15 textos narrativos e expositivos dispostos em ordem crescente de dificuldade (baseada no número de palavras, na quantidade e tipos de inferências a serem realizadas e na complexidade do conteúdo tratado). Os textos e as questões foram divididos em dois blocos de aplicação: a Forma A (sete textos) e a Forma B (oito textos). Possui ainda um texto comum que foi aplicado a toda a amostra, chamado de calibração, o qual servirá como base para a equalização dos escores da amostra para dados normativos e validação de construto da tarefa (Embretson & Reise, 2000), assunto que não será tratado aqui. A aplicação do instrumento é individual e, após a leitura pela criança (que opta por fazê-la oral ou silenciosamente), são fornecidas de nove a 12 questões abertas que se referem à passagem lida e que devem ser respondidas oralmente. Um crivo de correção foi criado contendo modelos de respostas corretas e incorretas, o qual deve ser usado pelo aplicador para nortear a correção. As respostas são pontuadas com um (1) ou zero (0) ponto, caso sejam consideradas corretas ou incorretas, respectivamente. Maiores detalhes sobre a tarefa podem ser encontrados em Lúcio, Kida, Carvalho, Cogo-Moreira, e Ávila (2015).

A título de exemplo, reproduz-se, com as respectivas questões, o Texto 1 da Forma A da PCL: 1. Texto. "A menina esperava à janela. / - Olha, lá vem o Téo! Nós podemos tomar o sorvete que você comprou, mamãe? / - Vocês querem de chocolate, Ana? / - Eu sim! Mas ele gosta de morango. / Ana e Téo tomaram sorvete e brincaram no jardim". 2. Questões: 1. Onde a menina esperava?; 2. Quem chegou à casa de Ana?; 3. Quem está conversando?; 4. Quem gosta de sorvete de morango?; 5. Por que Ana teve que pedir para tomar sorvete?; 6. Qual o sabor do sorvete que Ana prefere?; 7. Para que Ana esperava Téo?; 8. Onde Ana e Téo brincaram?; 9. Como estava o tempo nesse dia?; 10. Do que Ana precisa para brincar? Por quê? ".

A Prova de Compreensão Oral (PCO) busca avaliar, por meio de oito questões abertas, a compreensão que a criança obteve de uma passagem (texto narrativo) ouvida por meio de uma gravação (Carvalho, Kida, Lúcio, Cogo-Moreira, & Ávila, submetido). O texto escolhido foi uma adaptação de "O macaco e o coelho", de Monteiro Lobato. Após ouvir a história, a criança deve responder oralmente às oito questões feitas pelo avaliador (por exemplo, na Questão 1, pergunta-se: "No acordo que o macaco e o coelho fizeram quem deveria matar as cobras? "). De forma semelhante à PCL, as respostas são categorizadas como corretas (um ponto) ou incorretas (zero ponto), conforme crivo de correção estabelecido por uma banca de especialistas.

Procedimento e Amostragem

O presente estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de São Paulo (processo número 38406/12).

O estudo da fidedignidade foi feito a partir de um delineamento de cruzamento total (*fully crossed*), em que todos os juízes corrigem os protocolos dos mesmos participantes. Em relação à PCL, a amostra original foi composta por 739 crianças, sendo que 427 responderam à Forma A e 312 à Forma B. As crianças foram avaliadas por um total de 10 fonoaudiólogas treinadas. Desse modo, foram sorteadas duas fonoaudiólogas (uma que avaliou as crianças que responderam à Forma A e outra foi avaliadora da Forma B). As crianças avaliadas por essas profissionais compuseram a amostra para o estudo da fidedignidade.

Após essa primeira seleção, quatro avaliadores independentes (um do sexo masculino e três do sexo feminino) foram aleatoriamente alocados para correção dos protocolos originais: dois avaliadores corrigiram as respostas dadas pelas crianças para a Forma A e dois para a Forma B. Os avaliadores eram alunos de iniciação científica do quarto ano do curso de Psicologia. Foi fornecido um treinamento, em que foram apresentados os textos, as questões e o crivo de respostas. Além disso, foram também instruídos a não trocarem informações sobre a correção dos protocolos. As respostas julgadas corretas receberam o escore de um ponto, e as incorretas, zero ponto. A distribuição final dos protocolos ficou assim representada:

[3] Para esta revisão, não foi considerado o instrumento de Saraiva, Moojen, e Munarski (2006) porque ele não apresenta estudos psicométricos de validade ou fidedignidade, além de carecer de um sistema de interpretação dos escores.

[4] Foi encontrado apenas um estudo de fidedignidade entre diferentes sistemas de correção do Teste de Cloze, com amostra de estudantes universitários (Oliveira, Boruchovitch, & Santos, 2007).

FORMA A: foram avaliados os protocolos de 73 crianças, sendo 47% meninas e 65,8% de escolas estaduais (n=19, 15, 23 e 16 para crianças do 2º ao 5º ano, respectivamente).

FORMA B: foram avaliados os protocolos de 56 crianças, sendo 61% meninas e 23,2% de escolas estaduais (n=12, 15, 16 e 13 para crianças do 2º ao 5º ano, respectivamente).

Para a Prova de Compreensão Oral (PCO), procedimento semelhante ao anteriormente descrito foi adotado. Todas as 739 crianças participantes do estudo realizaram a PCO. Partiu-se, assim, para o sorteio de uma das fonoaudiólogas que coletou os dados, e dois avaliadores independentes corrigiram os protocolos de respostas das crianças. Novamente, alunos de iniciação científica (de ambos os sexos) foram os juízes (diferentes dos que participaram do estudo com a PCL). A amostra final foi composta por 89 crianças (48% meninas) do 2º ao 5º ano (n=24, 30, 20 e 15, respectivamente) de escolas municipais escolas e estaduais (83%).

Análise de Dados

Os dados foram analisados por meio do coeficiente Fleiss de Kappa, que constitui um índice de fidedignidade ajustado para múltiplos avaliadores em dados nominais (Siegel & Castellan, 1988). Landis e Koch (1977) apresentam a seguinte classificação para os índices k de Fleiss de kappa: $k \leq 0$ (concordância pobre); $0 < k \leq 0{,}20$ (concordância baixa); $0{,}20 < k \leq 0{,}40$ (concordância razoável); $0{,}40 < k \leq 0{,}60$ (concordância moderada); $0{,}60 < k \leq 0{,}80$ (concordância substancial); $k > 0{,}80$ (concordância quase perfeita). Para ser considerado satisfatório, o ponto de corte para o índice de fidedignidade foi de 0,60, conforme recomendações de McHugh (2012). Os dados foram analisados por meio do programa *StatsToDo* (disponível em https://www.statstodo.com/index.php).

Resultados

Estatísticas descritivas

As Tabelas 1 e 2 apresentam as proporções de concordância entre os pares e o total de avaliadores para a PCL e a PCO, respectivamente. Como se observa, as proporções de concordância são elevadas. Na PCL (Tabela 1), a concordância variou de 84,3% a 96,6% na Forma A (média 91%) e de 84,9% a 97,9% na Forma B (média 93%). Na PCO (Tabela 2), a concordância foi igualmente elevada, variando de 82% a 100% (média 93,8%). Apesar de úteis como uma forma de explorar os dados, as porcentagens não são a forma mais adequada de representar a fidedignidade entre avaliadores por desconsiderar a probabilidade de concordância por erro aleatório (Hallgren, 2012; McHugh, 2012). Nesse sentido, parte-se para a análise formal dos indicadores de concordância.

Tabela 1
PCL: Porcentagem de Concordância entre Pares de Avaliadores e Concordância Média entre os Três Avaliadores – Formas A e B

Texto	Avaliadores da Forma A				Avaliadores da Forma B			
	1 & 2	1 & 3	2 & 3	Total	1 & 2	1 & 3	2 & 3	Total
Calibração	91,3%	86,0%	85,2%	87,5%	84,9%	90,7%	86,1%	87,2%
Texto 1	96,5%	89,9%	90,3%	92,2%	97,7%	94,5%	95,3%	95,8%
Texto 2	91,0%	87,6%	84,3%	86,9%	96,8%	94,6%	94,3%	95,2%
Texto 3	92,2%	90,2%	92,7%	91,7%	93,0%	90,7%	90,3%	91,3%
Texto 4	94,8%	90,9%	91,0%	92,2%	95,8%	87,9%	86,7%	90,6%
Texto 5	95,3%	91,3%	91,8%	92,8%	94,3%	91,9%	91,9%	93,0%
Texto 6	95,9%	93,8%	94,4%	94,6%	93,7%	94,6%	92,7%	93,4%
Texto 7	96,6%	87,7%	87,3%	90,3%	94,9%	93,8%	94,0%	94,2%
Texto 8	-	-	-	-	98,2%	91,1%	89,3%	92,9%
Média	94,2%	89,7%	89,6%	91,0%	97,9%	95,9%	95,8%	96,5%

Tabela 2
PCO: Porcentagem de Concordância entre Pares de Avaliadores e Concordância Média entre os Três Avaliadores

Questão	Pares de Avaliadores			Concordância Total
	1 & 2	1 & 3	2 & 3	
1	98,9%	98,9%	100,0%	99,3%
2	100,0%	100,0%	100,0%	100,0%
3	96,6%	96,6%	95,5%	96,3%
4	94,4%	94,4%	96,6%	95,1%
5	85,4%	92,1%	91,0%	89,5%
6	93,3%	94,4%	95,5%	94,4%

Tabela 2 (continuação)
PCO: Porcentagem de Concordância entre Pares de Avaliadores e Concordância Média entre os Três Avaliadores

Questão	Pares de Avaliadores			Concordância Total
	1 & 2	1 & 3	2 & 3	
7	83,1%	88,8%	92,1%	88,0%
8	82,0%	82,0%	100,0%	88,0%
Média	91,7%	93,4%	96,3%	93,8%

Análise da fidedignidade

As Tabelas 3 e 4 apresentam os resultados da análise de fidedignidade a partir do coeficiente Fleiss de kappa. Em relação à PCL, observou-se que, em média, para todos os textos, os coeficientes foram adequados (k>0,60), com classificações variando de substancial a quase perfeita (Tabela 3). A exceção ficou por conta do Texto 7 na Forma A, que alcançou índices considerados inaceitáveis de concordância (média 0,30). Dos oito itens que compõem o Texto 7 da Forma A, apenas o primeiro apresentou concordância satisfatória (k=0,90), com os demais apresentando valores muito baixos, inclusive negativos. Alguns itens também contribuíram para a redução dos intervalos de confiança dos Textos 3 a 7 de ambas as formas, sinalizando a necessidade de revisão nesses itens. Para a PCO, todas as questões apresentaram concordâncias satisfatórias, com média elevada de coeficientes Fleiss de kappa e intervalos de confiança também adequados (Tabela 4).

Tabela 3
PCL: Fleiss de Kappa, Intervalos de Confiança e Classificação Conforme Landis & Koch (1977) – Forma A e Forma B

Texto	Forma A			Forma B		
	Fleiss	I.C.	Classificação	Fleiss	I.C.	Classificação
Calibração	0,71	0,15-1,08	Substancial	0,74	0,15-1,15	Substancial
Texto 1	0,78	0,45-1,11	Substancial	0,85	0,31-1,15	Perfeita
Texto 2	0,74	0,31-1,07	Substancial	0,86	0,35-1,15	Perfeita
Texto 3	0,71	0,04-1,09	Substancial	0,77	-0,16-1,15	Substancial
Texto 4	0,63	0,01-1,13	Substancial	0,70	-0,33-1,18	Substancial
Texto 5	0,73	-0,17-1,11	Substancial	0,72	0,19-1,10	Substancial
Texto 6	0,72	-0,14-1,11	Substancial	0,83	0,17-1,15	Perfeita
Texto 7	0,30	-0,21-1,11	Razoável	0,90	0,36-1,15	Perfeita
Texto 8	-	-	-	0,89	0,49-1,15	Perfeita
Média	0,68	-0,21-1,13	Substancial	0,80	-0,33-1,18	Perfeita

Tabela 4
PCO: Fleiss de Kappa, Intervalos de Confiança e Classificação Conforme Landis & Koch (1977) para as Oito Questões da Tarefa

Questão	Fleiss	I.C.	Classificação
1	0,97	0,85-1,09	Perfeita
2	1,00	0,88-1,12	Perfeita
3	0,88	0,76-1,00	Perfeita
4	0,91	0,79-1,03	Perfeita
5	0,75	0,63-0,87	Substancial
6	0,89	0,77-1,01	Perfeita
7	0,76	0,64-0,88	Substancial
8	0,76	0,64-0,88	Substancial
Total	0,97	0,63-1,09	Perfeita

Discussão

De uma maneira ampla, e segundo Urbina (2014), a fidedignidade representa o grau em que é possível confiar nas informações fornecidas pelos escores de testes, pois são relativamente livres de erros. A fidedignidade precisa ser empiricamente investigada por meio de procedimentos específicos, tornando possível prever a flutuação

dos escores como consequência do processo de medida, sendo, portanto, um indicador de sua utilidade. Existem diversas fontes de erros que podem interferir na confiança que se tem nos escores dos testes e esta investigação constitui em si uma evidência preliminar da própria validade das interpretações pretendidas a partir dos escores. Desse modo, a fidedignidade do avaliador, tema do presente trabalho, aplica-se a testes cujos escores podem depender, em algum nível, da subjetividade daquele que aplica/corrige o teste. Sendo assim, ao se demonstrar a concordância entre diferentes avaliadores, atesta-se para o fato destes terem utilizado os mesmos critérios para correção, indicando, dessa forma, que estão avaliando o mesmo construto. Essa questão mostra o quão crucial constitui a investigação da fidedignidade do avaliador para a validade das informações que se pode extrair de testes contendo questões abertas.

O estudo apresentou uma investigação da fidedignidade do avaliador da Prova de Compreensão Leitora (PCL) e da Prova de Compreensão Oral (PCO), ambos instrumentos de avaliação educacional que se encontram em fase de construção. Sendo tarefas compostas por questões abertas, é essencial demonstrar a aplicabilidade dos crivos de correção, de modo que todos os avaliadores treinados na aplicação possam chegar a resultados semelhantes, ou seja, que os instrumentos sejam relativamente livres de erros ocasionados por interferência de subjetividade na correção. Grosso modo, o trabalho demonstrou que a maioria dos textos da PCL e todas as questões da PCO apresentaram indicadores adequados de fidedignidade, de modo que é possível concluir que os avaliadores utilizaram os mesmos critérios para conferir os escores aos participantes.

Para o PCL, de modo geral, a Forma B apresentou indicadores mais precisos do que a Forma A. em média, a concordância na Forma A foi de 0,68 (substancial) e da Forma B 0,80 (quase perfeita). Essa diferença se deu principalmente devido aos altos índices de discordância no Texto 7 da Forma A, em que a média de concordância dos itens foi de 0,30. Retirado esse texto, a concordância na Forma A passa a ser de 0,72 (substancial). Sendo o Texto 7 criado para ser o mais difícil da Forma A, é possível que as produções (respostas) das crianças tenham sido insuficientes para uma avaliação mais precisa por parte dos juízes. Dos oito itens, apenas um alcançou valor $k>0,60$. Tal resultado indica que o texto não é adequado para a avaliação da compreensão de crianças nas fases iniciais de escolarização.

Resultados muito satisfatórios foram encontrados para a PCO. Todas as questões obtiveram índices de confiabilidade elevada (média de 0,97) e com limites inferiores de intervalo de confiança pelo menos substancial (menor valor 0,63). Isso sinaliza que todas as questões utilizadas na prova são relativamente isentas de subjetividade na avaliação e não necessitam de revisões mais profundas.

Apesar de não terem sido encontrados estudos de fidedignidade do avaliador com tarefas de compreensão oral e leitora, foram encontradas algumas investigações de testes psicológicos que apresentam relatos desse tipo de fidedignidade e cujos resultados vão ao encontro dos índices obtidos no presente estudo. Por exemplo, Marques et al. (2002) não encontraram diferenças entre avaliadores na correção do Teste de Goodenough utilizando o método de Kendall ($p=0,07$). Pawlowski, Parente, e Bandeira (2013) utilizaram coeficientes de correlação intraclasse para análise de concordância das tarefas de praxias construtivas do Neupsilin e encontraram valores entre 0,73 e 0,91. Resultados semelhantes ao do presente estudo foram obtidos por Fensterseifer, Lima, Paranhos, e Werlang (2009) utilizando o coeficiente Kappa para cálculo da fidedignidade entre avaliadores para o Teste de Apercepção Familiar. A maioria das categorias analisadas pelas autoras apresentaram índices de concordância de substancial a quase perfeita, com valores acima de 0,79, mas sem relatos dos intervalos de confiança.

A partir do que foi exposto, ressalta-se a relevância da presente pesquisa para área da avaliação psicoeducacional, que carece de estudos que forneçam dados a respeito da fidedignidade de instrumentos cujas respostas dos avaliandos são susceptíveis à interferência da subjetividade do avaliador. De uma forma mais específica, os resultados obtidos no estudo irão proporcionar bases empíricas para a confiança nos crivos de correção da PCO e para o aprimoramento (revisão) de alguns itens, instruções e/ou do crivo da PCL. Em relação a essa última prova, a recomendação imediata é a retirada do Texto 7 da Forma A, pois poucos itens mostraram-se adequados. Para os outros textos, a retirada de itens com baixa concordância não causará maiores impactos, por constituírem uma minoria. A análise aqui conduzida segue os parâmetros da psicometria para avaliação da qualidade dos instrumentos de avaliação psicológica (AERA et al. 2014; Urbina, 2014) e ajudará a disponibilizar ao usuário final indicadores confiáveis das habilidades de compreensão leitora de crianças.

Finalmente, salienta-se, como limitação do trabalho, a amostra reduzida de crianças em relação à amostra geral do estudo. Por ter-se optado pelo método *fully crossed* (em que todos os participantes são avaliados pelos mesmos juízes), a amostra final ficou composta pelos protocolos das crianças avaliadas por uma aplicadora em cada uma das formas. O uso de delineamentos not *fully crossed* poderia proporcionar um aumento da amostra (e, possivelmente, uma melhoria dos índices de fidedignidade). Entretanto, a opção pelo método e pelo uso do Fleiss de Kappa se justificou pela possibilidade de relato de intervalos de confiança, os quais são raramente reportados em estudos que tratam da fidedignidade do avaliador. Certamente, esse é um diferencial em relação a outros estudos, muitos dos quais reportam apenas

porcentagens de concordância ou mesmo estatísticas consideram apenas pares de avaliadores, práticas essas que têm sido questionadas por pesquisadores na área (Hallgren, 2012).

Referências

American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NMCE]. (2014). *Standards for educational and psychological testing.* Washington, DC: Amer Educational Research Assn.

Carvalho, C. A. F., Kida, A. S. B., Lúcio, P. S., Cogo-Moreira, H., & Ávila, C. R. B. *Listening comprehension task: Psychometric parameter analysis.* (Manuscrito submetido para publicação).

Chadha, N. K. (2009). *Applied psychometry.* New Delhi, India: SAGE Publications India.

Cuetos F., Rodrigues B., & Ruano E. (2010). *PROLEC: Provas de Avaliação dos Processos de Leitura* (2ª. ed, Adaptado por A. S. Capellini, A. M. Oliveira & F. Cuetos). São Paulo: Casa do Psicólogo.

Embretson, S., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* New Jersey: Lawrence Erlbaum Associates.

Fensterseifer, L., Lima, G. Q., Paranhos, M. E., & Werlang, B. S. G. (2009). Fidedignidade entre avaliadores no Teste de Apercepção Familiar (FAT). *Psico (PUCRS), 40*(3), 287-293.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Joly, M. C. R. A., & Istome, A. C. (2008). Compreensão em leitura e capacidade cognitiva: estudo de validade do teste Cloze Básico – MAR. Psic: *Revista da Vetor Editora, 9*(2), 219-228.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. doi:10.2307/2529310

Lúcio, P. S., Kida, A. S. B., Carvalho, C. A. F., Cogo-Moreira, H., & Ávila, C. R. B. (2015). Construção de uma prova para avaliação da compreensão leitora no ensino fundamental: estudo piloto. *Temas em Psicologia, 23*(4), 1035-1050. doi: 10.9788/TP2015.4-17.

Marques, S., Pasian, S. R., Franco, M. A. P., Panosso, I. R., Viana, A. B., & Oliveira, D. A. (2002). Avaliação cognitiva de crianças com dificuldades de aprendizagem: Precisão do Teste de Goodenough (1926) e da EMMC (1993). *Paidéia, 12*(23), 105-112. doi: 10.1590/S0103-863X2002000200008

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282.

Oliveira, K. L., Boruchovitch, E., & Santos, A. A. A. (2007). Análise da fidedignidade entre dois tipos de pontuação do Teste de Cloze. *Psicologia em Pesquisa, 1*(1) 41-51.

Ortiz, K. Z., Osborn, E., & Chiari, B. M. (1993). O teste M1-Alpha como instrumento de avaliação da afasia. *Pró-Fono, 5(*1), 23-9. doi: 10.1590/S2179-64912011000300007

Pawlowski, J. Parente, M. A. M. P., & Bandeira, D. R. (2013). Fiabilidad del instrumento de evaluación neuropsicológica breve NEUPSILIN. *Avances en Psicologia Latinoamericana, 31*(1), 62-70.

Radanovic, M., Mansur, L. L., & Scaff, M. (2004). Normative data for the Brazilian population in the Boston diagnostic aphasia examination: Influence of schooling. *Brazilian Journal of Medical and Biological Research, 37*(11), 1731-1738. doi: 10.1590/S0100-879X2004001100019

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York, NY: Routledge.

Santos, A. A. A., Primi, R., Taxa, F., & Vendramini, C. M. M. (2002). O Teste de Cloze na avaliação da compreensão em leitura. *Psicologia: Reflexão e Crítica, 15*(3), 549-560.

Saraiva, R. A., Moojen, S. M. P., & Munarski, R. (2006). *Avaliação da compreensão leitora de textos expositivos para fonoaudiólogos e psicopedagogos* (2ª ed.). São Paulo: Casa do Psicólogo.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2ª ed.). New York: McGraw-Hill.

Urbina, S. (2014). *Essentials of psychological testing* (2ª ed.). Hoboken, New Jersey: John Wiley & Sons.

recebido em setembro de 2015
reformulado em janeiro de 2016
aprovado em janeiro de 2016

Sobre os autores

Patrícia Silva Lúcio é professora assistente da Universidade Estadual de Londrina e doutoranda pelo programa de Pós-Graduação em Psiquiatria e Psicologia Médica da Universidade Federal de São Paulo.

Adriana de Souza Batista Kida é fonoaudióloga e pós-doutora em Fonoaudiologia pela Universidade Estadual Paulista de Marília.

Carolina Alves Ferreira de Carvalho é fonoaudióloga e doutora em Ciências pela Universidade Federal de São Paulo.

Hugo Cogo-Moreira é pós-doutorando pelo Programa de Pós-Graduação em Psiquiatria e Psicologia Médica da Universidade Federal de São Paulo e professor orientador do mesmo programa.

Clara Regina Brandão de Ávila é livre-docente pela Universidade Federal de São Paulo, professora do Programa de Pós-Graduação em Distúrbios da Comunicação Humana da Universidade Federal de São Paulo e bolsista de produtividade em Pesquisa do CNPq – Nível 2 – CA MS.