DFH – ESCALA SISTO: UM ESTUDO EXPLORATÓRIO SOBRE A PRECISÃO DE JUÍZES

DIAS, Augusto Rodrigues[1]; BRITO, Beatriz Patrício de[2]; BOVENZO FILHO, Carlos Eduardo[3].

RESUMO

O objetivo foi estabelecer a precisão de avaliadores para os itens de pontuação do DFH – Escala Sisto, a partir de juízes sem experiência na avaliação de desenhos da figura humana. A amostra foi composta por 20 indivíduos, de ambos os sexos, com idades variando entre 19 a 51 anos de idade. Os resultados apontaram que os juízes, percentualmente, tenderam a concordar em suas avaliações sem, contudo, atingirem o valor mínio estabelecido. Foram consistentes nas avaliações realizadas, apresentando uma consistência interna que variou entre 0,646 a 0,913 e Alfa de Cronbach de 0,977. Em termos de fidedignidade dos avaliadores, foi apurado um valor kappa de 0,664 para o conjunto de trinta itens do instrumento. Conclui-se que o instrumento é fidedigno sob o ponto de vista da precisão de juízes para a população em questão, bem como existe a possibilidade de que a experiência prévia exerça alguma influência no processo avaliativo.

Palavras-chave: DFH Escala Sisto. Precisão de juízes. Parâmetros psicométricos.

ABSTRACT

The objective in this paper was to verify the accuracy of evaluators for the scoring items of the D.F.H - Escala Sisto, from judges without experience in the evaluation of drawings of the human figure. The sample consisted of 20 individuals, of both sexes, with ages varying from 19 to 51 years old. The results showed that the judges, in percentage, tended to agree on their assessments without, however, reaching the established minimum value. They were consistent in the evaluations performed, presenting an internal consistency that varied between 0.646 to 0.913 and Cronbach's alpha of 0.977. In terms of the reliability of the evaluators, a kappa value of 0.664 was found for the set of thirty items of the instrument. It was concluded that the instrument is reliable from the point of view of the precision of judges for the population in question, as well as the possibility that previous experience may have some influence on the evaluation process.

Key words: DFH Scale. Inter-rater reliability. Psychometric parameters.

[1] Psicólogo, mestre em Avaliação Psicológica pela Universidade São Francisco (2005). Docente nos cursos de graduação em Psicologia do Centro Universitário UNIFAAT - Atibaia/SP e Univeritas – UNG - Guarulhos/SP. E-mail: psicodias2@yahoo.com.br – fone contato (11) 998494471

[2] Discente do 9º semestre do Curso de Psicologia do Centro Universitário UNIFAAT – Atibaia/SP. E-mail: bbeatriz_patricio@yahoo.com.

[3] Psicólogo, graduado pela Univeritas – UNG (2019), Assistente de Psicologia, Produtos e Pesquisa da Vetor Editora. Email: carlos.filho@vetoreditora.com.br – fone (11) 3146-0337 - Ramal 131.

Introdução

De acordo com o Conselho Federal de Psicologia (2018), a avaliação psicológica pode ser conceituada como um processo estruturado de investigação de fenômenos psicológicos, com o objetivo de prover informações à tomada de decisão, no âmbito individual, grupal ou institucional, com base em demandas, condições e finalidades específicas. Além desse aspecto, o CFP, por meio da resolução 09/2018, evidencia que na realização da Avaliação Psicológica, o(a) psicólogo(a) deve basear sua decisão, obrigatoriamente, em métodos e/ou técnicas e/ou instrumentos psicológicos reconhecidos cientificamente para uso na prática profissional. Ao se considerar especificamente os testes psicológicos, o profissional da Psicologia deve estar atento aos parâmetros psicométricos fundamentais apresentados em seus manuais, em especial, aos relativos à validade e precisão ou fidedignidade.

No que se refere à precisão ou fidedignidade dos testes psicológicos, essa diz respeito ao grau de reprodutibilidade da medida. Segundo Pasquali (2009), é mais fácil entender este conceito quando se conjectura: se a operação de mensuração for repetida de diferentes maneiras, os resultados serão os mesmos de cada vez? Nesse sentido, pode-se depreender que, quanto mais consistente for a medida, maior será a precisão e haverá menos erro que possa interferir com a medida do que se deseja apurar. Por outro lado, quanto maior for a discrepância entre as pontuações obtidas, menor será a consistência e, por conseguinte, maior será o erro de medida. Assim, o conceito de precisão ou fidedignidade opõe-se ao erro de medida, de modo que, quanto mais um teste for considerado preciso, mais ele se encontra livre de erros (ALVES; SOUZA; BAPTISTA, 2011). Em síntese, enquanto a fidedignidade na mensuração implica em consistência e precisão, a sua ausência acarreta exatamente o contrário, sendo que, em comum, as duas situações implicam em erros de mensuração (URBINA, 2007).

O estabelecimento da precisão ou fidedignidade de um teste psicológico pode ser efetuado de diferentes maneiras, dentre as quais pode-se citar a fidedignidade do avaliador ou de juízes (ANASTASI; URBINA, 2000; URBINA, 2007; ALVES; SOUZA; BAPTISTA, 2011). Esse método constituiu-se na solicitação de que dois ou mais avaliadores diferentes pontuem o protocolo de teste do(s) mesmo(s) sujeito(s) de forma separada, correlacionando depois os resultados das correções. Para Marín Rueda et al. (2008), as fontes de erro que podem ser encontradas nesse tipo de estudo provêm ou da subjetividade do avaliador ou das normas de correção apresentadas pelos manuais. A subjetividade do avaliador pode induzir à confusão e erros de avaliação, de tal modo que, para minimizar sua interferência, as normas de correção

devem estar adequadamente delineadas e objetivas. Quando tais condições encontram-se presentes nas normas de correção, diferentes juízes adotam o mesmo critério/entendimento na hora de julgar um item. O Teste do Desenho da Figura Humana – Escala Sisto (2005) utilizou esse procedimento para o estabelecimento da precisão do sistema de pontuação.

O Teste do Desenho da Figura Humana – Escala Sisto é um instrumento indicado para avaliar a inteligência de crianças pequenas, ou com diminuição da capacidade auditiva ou deficiências neurológicas, não alfabetizadas ou que não falam a língua do examinador, na faixa etária de 05 a 10 anos de idade. É útil e recomendado para o profissional da Psicologia que necessita de uma medida rápida do desenvolvimento cognitivo da criança, por oferecer uma indicação bem razoável da localização do examinando em relação aos seus pares mais novos e mais velhos. Em termos de aplicação, solicita-se à criança que faça dois desenhos da figura humana (um masculino e outro feminino). A tarefa em si é muito atrativa para as crianças pelo fato de não ser ameaçadora, e tem a duração aproximada de 20 a 25 minutos.

O sistema de pontuação do DFH – Escala Sisto é composto por trinta itens, cada qual com sua definição e exemplos de desenhos, que devem ser consultados para definição de sua consideração ou não na hora de pontuar. Os trintas itens são apresentados numa escala em termos de habilidade da criança, no caso de seu desenvolvimento, e da dificuldade do detalhe desenhado em função da idade. A correção é realizada pelo total de pontos adquiridos pelo respondente em cada um dos desenhos realizados e envolve uma avaliação quantitativa e qualitativa. Os resultados são expressos por intermédio de uma escala de percentis ou medida Rasch, nas quais se identifica o posicionamento do respondente em relação aos demais indivíduos da amostra normativa.

Em relação aos estudos de precisão ou fidedignidade por avaliadores, Sisto (2005) desenvolveu uma pesquisa tendo três mestrandos como juízes. Esses possuíam boa experiência em correção da figura humana e avaliaram 15 protocolos de crianças com diferentes idades, mediante o fornecimento de definições e exemplos de detalhes que deveriam ser pontuados ou não. Em termos de resultados, foram apurados coeficientes de correlação entre os três juízes de 0,86, 0,90 e 0,91. Tais valores foram avaliados por Sisto (2005) como muito bons.

A partir desses resultados, o autor do teste considerou que as flutuações observadas nas avaliações realizadas pelos juízes não interferiram significativamente nos resultados, implicando considerar que a margem de acerto foi superior à margem de erro de medida. Em outras palavras, concluiu que as normas de correção elaboradas estavam claras e objetivas o suficiente para minimizar o efeito da subjetividade dos avaliadores, demonstrando que as

definições e exemplos dados como guia de avaliação do teste se encontravam adequadamente descritos e permitiam um bom entendimento dos critérios adotados, independentemente de quem os utilizasse para avaliar os desenhos realizados pelas crianças.

Ocorre que se for considerada a experiência na avaliação de desenhos da figura humana dos avaliadores escolhidos para o estabelecimento da precisão de juízes, essa pode, em princípio, ter influenciado os resultados obtidos. A partir dessa ótica, o entendimento de que as definições e exemplos dados como guia no sistema de correção do teste estariam adequadamente descritos, permitindo um bom entendimento dos critérios adotados, independentemente de quem os utilize para avaliar os desenhos realizados pelas crianças, poderia não refletir a realidade. Ao se considerar essa possibilidade, pode-se questionar: o que aconteceria com os resultados do estudo de precisão entre avaliadores se as avaliações fossem realizadas por juízes sem experiência na correção desses desenhos? Seria possível que as definições e exemplos dos 30 itens que compõem o sistema de correção se apresentariam adequadamente descritos e objetivos? Será que as medidas erro se manteriam nos patamares apresentados?

Para dirimir essas dúvidas, a presente pesquisa se propôs a desafiar os resultados obtidos pelo autor do instrumento no estabelecimento da precisão entre avaliadores para o sistema de correção do Teste da Figura Humana – Escala Sisto. Para tanto, realizou um estudo de precisão de avaliadores a partir de avaliações realizadas às cegas por juízes sem experiência na correção de desenhos da figura humana.

Método

Amostra:

A amostra foi composta por 20 estudantes de Psicologia, de ambos os sexos, regularmente matriculados no curso de Psicologia de uma Instituição de Ensino Superior localizada no interior do estado de São Paulo. As idades variaram de 19 a 51 anos de idade, com média de 26,4 anos e um desvio padrão de 10,34 anos. Todos, sem exceção, eram inexperientes na avaliação de desenhos de figura humana e realizaram suas avaliações às cegas.

Instrumentos:

Para a coleta de dados foi utilizado um protocolo de um desenho de figura humana realizado por uma criança do sexo feminino, com idade de 10 anos, cursando a 6ª série do

ensino fundamental. Além deste protocolo, foram utilizados 20 manuais do Teste do Desenho da Figura Humana – Escala Sisto.

Procedimento:

O projeto da pesquisa foi encaminhado ao Comitê de Ética em Pesquisa do Centro Universitário Unifaat que, após análise, forneceu parecer favorável. A coleta de dados ocorreu em um único dia e durante a realização das atividades práticas de uma disciplina relativa à área de avaliação psicológica. Os objetivos da pesquisa foram expostos e se fez o convite à participação. Os interessados em participar receberam o termo de Consentimento Livre e Esclarecido, solicitando-se a leitura, preenchimento e assinatura.

A coleta de dados propriamente dita ocorreu após a assinatura dos termos de consentimento. Foram distribuídos os cinco manuais do teste e o protocolo fictício, um para cada participante. Os participantes foram orientados a seguir as definições dos itens para pontuação descritas no manual e verificar se essas estavam presentes ou não no protocolo. De modo específico, foram orientados a toda vez que julgassem a presença de um item no desenho, indicar sua ocorrência e nada assinalar se julgassem o item ausente. Tomou-se o cuidado para que nenhum dos avaliadores tivesse conhecimento, em momento algum, da avaliação realizada pelos seus pares. O participante, ao terminar sua avaliação do protocolo, deveria entregá-la ao pesquisador responsável pela coleta de dados, encerrando a participação.

Plano de análise de dados:

Para a análise dos dados foram utilizados os recursos da estatística descritiva e inferencial constantes nos softwares XLSTAT e Minitab, versão 2018 para o Windows. Inicialmente, apuraram-se as estatísticas descritivas, por meio da frequência de assinalamento/concordância (presença ou ausência) feita pelos 20 juízes para os 30 itens do DFH – Escala Sisto. A concordância mede com que frequência dois ou mais avaliadores atribuem exatamente a mesma classificação (MATOS, 2014). Stemler (2004) indica o valor de 75% como o mínimo de concordância aceitável, e valores a partir de 90% são considerados altos. No presente estudo, para orientar a interpretação da concordância entre os juízes, foi estabelecido um índice de concordância mínimo aceitável entre avaliadores igual ou superior a 80%, ou 16 em 20 juízes.

Além da frequência de concordância, com o intuito de aprofundar o entendimento em relação aos dados apurados, buscou-se aferir a consistência interna dos 20 juízes na avaliação que fizeram. O objetivo foi verificar em que medida os juízes foram consistentes entre si no

julgamento dos 30 itens de classificação do DFH – Escala Sisto. A consistência interna é geralmente apurada por meio do Alfa de Cronbach (α), calculado ao se parear correlações entre os itens. Em geral, considera-se aceitável que um α de 0,6 a 0,7 indique fiabilidade aceitável e, acima de 0,8, boa fiabilidade. Valores iguais ou maiores a 0,95 apontam alta fiabilidade, entretanto essa geralmente não é desejada, já que aponta que os itens podem ser redundantes.

Assim, para aferir a consistência interna foram calculados a correlação item-total e o Alfa de Cronbach. Para tanto, considerou-se cada juiz, julgando a presença ou não de cada um dos 30 itens de avaliação do instrumento, como um item de um teste hipotético composto por 20 itens. O critério adotado para interpretar os valores apurados da correlação item-total respeitou a regra usual de que um item deve se correlacionar com a pontuação total acima de 0,30, sendo que itens com correlações mais baixas devem ser descartados (KLINE, 1986). Com relação ao Alfa de Cronbach, foi estabelecido um valor igual ou superior a 0,80, conforme indicado por Guilford, 1950 e Anastasi e Urbina (2000). De acordo com Dancey e Reidy (2006), valores dessa magnitude são considerados fortes.

Apurou-se também o Kappa de Fleiss como índice de fidedignidade ajustado para múltiplos avaliadores em dados nominais (SIEGEL; CASTELLAN, 1988). Como critério para essa estatística, adotou-se a classificação proposta por Landis e Koch (1977), a saber: < 0 = acordo deficiente; 0,01 - 0,20 = ligeiro acordo; 0,21- 0,40 = acordo justo; 0,41 - 0,60 = acordo moderado; 0,61- 0,80 = acordo substancial, e 0,81 - 1,00 = concordância quase perfeita. Para dar sentido aos valores de concordância expressos pela estatística kappa, considerou-se que os trinta itens de avaliação do instrumento possuem, em suas definições e exemplos/desenhos, clareza e objetividade suficientes para que um indivíduo, sem experiência e conhecimento na avaliação de desenhos da figura humana, indique a presença ou ausência dos mesmos na avaliação de um protocolo. A consideração desse pressuposto foi necessária para que o valor kappa apurado refletisse somente a influência ou não da falta de experiência no processo avaliativo e não outra variável interveniente. Além deste critério, buscou-se comparar os valores kappa apurados com relação ao valor mínimo (0,60) estabelecido pela Resolução 09/2018 do CFP para estudos de precisão.

Resultados

A tabela 1 apresenta a distribuição das indicações (presença ou ausência do item no desenho) e respectivas porcentagens por item do DFH – Escala Sisto, com base nas avaliações realizadas pelos 20 juízes sem experiência na avaliação de desenhos da figura humana.

Tabela 1. Distribuição das indicações e respectivas porcentagens relativas à avaliação realizada pelos juízes sem experiência para o DFH – Escala Sisto.

Itens	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Presença	20	20	20	20	19	14	20	20	15	03
%	100	100	100	100	95	70	100	100	75	15
Ausência	00	00	00	00	01	06	00	00	05	17
%	000	000	000	000	05	30	000	000	25	85
Nº total	20	20	20	20	20	20	20	20	20	20
% total	100	100	100	100	100	100	100	100	100	100

Itens	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Presença	17	14	20	12	20	18	18	10	06	19
%	85	70	100	60	100	90	90	50	30	95
Ausência	03	06	00	08	00	02	02	10	14	01
%	15	30	00	40	00	10	10	50	70	05
Nº total	20	20	20	20	20	20	20	20	20	20
% total	100	100	100	100	100	100	100	100	100	100

Itens	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Presença	06	18	20	07	00	00	01	00	20	00
%	30	90	100	35	00	00	05	00	100	00
Ausência	14	02	00	13	20	20	19	20	00	20
%	70	10	00	65	100	100	95	100	00	100
Nº total	20	20	20	20	20	20	20	20	20	20
% total	100	100	100	100	100	100	100	100	100	100

Ao se considerar os trinta itens do instrumento em estudo, observou-se que os vinte juízes tenderam a avaliar da mesma forma 73,33% (n=22) desses, divergindo em 26,66% (n=08). Os itens em que os juízes apresentaram concordância igual ou superior ao estabelecido (80% ou 16 em 20 juízes) no plano de análise de dados foram os itens 1, 2, 3, 4, 5, 7, 8, 10, 11, 13, 15, 16 17, 20, 22, 23, 25, 26, 27, 28, 29 e 30. Para os itens 6, 9, 12, 14, 18, 19, 21 e 24 o percentual de concordância foi inferior, inclusive abaixo do indicado pela literatura como o mínimo aceitável (75% ou 15 em 20 juízes).

Com a intenção de estabelecer uma comparação entre os percentuais apurados no presente estudo e os resultados das correlações apresentados por Sisto (2015), para o estudo da precisão entre avaliadores, calculou-se inicialmente a média aritmética simples dessas correlações. No manual do teste, o autor informa que os valores das correlações obtidos entre os três juízes foram 0,86, 0,90 e 0,91. Assim, o valor médio de correlação foi de 0,89, que, transformado em porcentagem, indicou percentuais de concordância e discordância nas avaliações dos protocolos feitas por indivíduos experientes na avaliação de desenhos da figura humana, na ordem de 79,2% e 20,8%, respectivamente. O percentual de concordância em

questão encontra-se muito próximo ao valor de porcentagem adotado (80%) e supera o indicado na literatura (75%).

Com base no exposto e considerando-se as definições e exemplos dados para que cada item do instrumento esteja adequadamente delineado, pode-se conjecturar que a falta de experiência na avaliação de desenhos da figura humana tenha exercido alguma influência no momento da avaliação. Assim, em princípio, a avaliação dos protocolos do teste realizada por avaliadores sem experiência tende a apresentar uma medida erro maior em relação às avaliações feitas por pessoas com experiência. Entretanto, esta conjectura deve ser vista com uma certa reserva, na medida em que carece de estudos estatísticos que confirmem a existência de uma diferença significativa entre as porcentagens consideradas.

Para apurar em que medida os vinte juízes foram consistentes entre si no julgamento dos trinta itens de avaliação do DFH – Escala Sisto, foram calculados a correlação item-total e o Alfa de Cronbach. A correlação entre o item e a pontuação total mostra a força que um determinado item possui no grupo de itens em que está inserido (DIAS; COUTO; PRIMI, 2009), bem como indica que se as respostas a esses itens forem internamente consistentes, existe evidência de que os itens medem a mesma construção. A tabela 2 apresenta essas estatísticas.

O critério adotado para interpretar os valores apurados da correlação item-total respeitou a regra usual de que um item deve se correlacionar com a pontuação total acima de 0,30, sendo que itens com correlações mais baixas devem ser descartados (KLINE, 1986). A partir desta condição, pode-se observar que os valores das correlações item-total variaram de 0,64 (juiz 14) a 0,91 (juízes 16 e 17), bem como a correlação mais fraca manteve-se acima do valor mínimo estabelecido pela literatura (0,30).

Tabela 2. Correlação item-total e Alfa de Cronbach para os 20 juízes, considerando-se os 30 itens de avaliação do DFH – Escala Sisto.

(continua)

Juiz	Média	Desvio Padrão	Correlação Item-total	Alfa de Cronbach
1	11,50	7,78	0,83	0,97
2	11,36	7,80	0,86	0,97
3	11,53	7,76	0,87	0,97
4	11,53	7,76	0,87	0,97
5	11,43	7,85	0,70	0,97
6	11,40	7,80	0,84	0,97
7	11,36	7,80	0,86	0,97
8	11,43	7,85	0,72	0,97
9	11,46	7,81	0,78	0,97
10	11,50	7,82	0,75	0,97

(conclusão)

Juiz	Média	Desvio Padrão	Correlação Item-total	Alfa de Cronbach
11	11,60	7,84	0,71	0,97
12	11,36	7,82	0,81	0,97
13	11,40	7,80	0,84	0,97
14	11,46	7,88	0,64	0,97
15	11,46	7,82	0,76	0,97
16	11,50	7,74	0,91	0,97
17	11,50	7,74	0,91	0,97
18	11,50	7,81	0,77	0,97
19	11,50	7,78	0,83	0,97
20	11,43	7,78	0,87	0,97

Alfa de Cronbach 0,97

Com relação ao Alfa de Cronbach, foi estabelecido um valor igual ou superior a 0,80, conforme indicado por Guilford (1950) e Anastasi e Urbina (2000). De acordo com Dancey e Reidy (2006), valores dessa magnitude são considerados fortes. O coeficiente Alfa de Cronbach apurado para os 20 juízes foi de 0,97, atestando a boa consistência interna, sendo que a exclusão de qualquer um dos 20 juízes, em especial o juiz 14, não elevaria significativamente este coeficiente.

Nesse sentido, bem como considerando-se que um nível alto de consistência interna indica que os itens projetados para avaliar a construção fornecem pontuações similares, de tal modo que se as respostas a esses itens forem internamente consistentes, existe evidência de que os itens medem a mesma construção, o resultado encontrado pode ser interpretado como um indicativo de que os 20 juízes tenderam a avaliar os 30 itens do instrumento no mesmo sentido. Em outros termos, tenderam a seguir o mesmo entendimento no julgamento da presença ou não dos itens do DFH – Escala Sisto na avaliação do desenho.

Dando sequência à análise de dados, apurou-se o coeficiente kappa, mais especificamente apurou-se o Kappa de Fleiss. Esse coeficiente constitui-se em um procedimento estatístico que em seu cálculo considera a probabilidade da ocorrência da concordância ao acaso (CROCKER; ALGINA, 2009). Explicando melhor, esse coeficiente pode ser entendido como a proporção de concordância resultante entre os juízes após a retirada da proporção de concordância devido ao acaso (FONSECA et al., 2007). Assim, permite avaliar o grau de consenso existente entre os avaliadores, expresso num índice que se assemelha ao da correlação (variando de 0 a 1). Valores acima de 0,60 são considerados aceitáveis, ou seja, um adequado índice de concordância (BISQUERRA; MARTÍNEZ; SARRIERA, 2004;

STEMLER, 2004). A tabela 3 apresenta o valor apurado e sua respectiva classificação, segundo Landis e Koch (1977).

Tabela 3. Coeficiente Kappa (Fleiss) apurado para o total de itens do DFH – Escala Sisto, faixa de concordância e respectiva classificação.

Nº de itens do DFH	Kappa (Fleiss)	Concordância	Classificação
30	0,664	$0,61 \leq K < 0,81$	Substancial

O coeficiente kappa foi significativo (p-valores < 0,001) para o conjunto de 30 itens, indicando concordância significativamente superior a 0 entre os avaliadores. A concordância entre os vinte avaliadores foi de 0,664, considerada substancial.

Para dar sentido ao valor de concordância expresso, considerou-se que os trinta itens de avaliação do instrumento possuem, em suas definições e exemplos/desenhos, clareza e objetividade suficientes para que um indivíduo, sem experiência e conhecimento na avaliação de desenhos da figura humana, indique a presença ou ausência dos mesmos na avaliação de um protocolo, conforme especificado no plano de análise de dados. A partir do valor apurado de k, é possível considerar que a falta de experiência prévia apresentou pouca influência na avaliação do desenho. Assim, os avaliadores, a partir das definições/exemplos dos itens de avaliação do DFH – Escala Sisto, mais concordaram do que divergiram em suas avaliações.

Ao se considerar o valor mínimo estabelecido pelo CFP para estudos de precisão, o instrumento como um todo (30 itens) pode ser considerado fidedigno sob o ponto de vista da fidedignidade do avaliador, em razão de ter apresentado um kappa de 0,66. Desse modo, mais uma vez as definições dos trinta itens de avaliação do DFH Escala Sisto sugerem ser suficientemente claras para que indivíduos que se encontram na mesma condição da amostra (sem experiência) realizem tal avaliação de modo relativamente seguro e fidedigno.

Considerações finais

O presente estudo teve como objetivo estabelecer a precisão do DFH Escala Sisto, a partir de avaliações realizadas às cegas por juízes sem experiência na correção de desenhos da figura humana, verificando a possível influência dessa variável sobre os resultados. De modo mais detalhado, buscou responder a seguinte questão: o que aconteceria com os resultados do estudo de precisão entre avaliadores se as avaliações fossem realizadas por juízes sem experiência na correção desses desenhos?

Ao se considerar os resultados encontrados, em termos percentuais, verificou-se que eles se situaram ligeiramente abaixo dos apontados por Sisto (2015) e dos valores definidos como critério mínimo estabelecido, mesmo apresentando uma excelente consistência interna. Assim, a falta de experiência na avaliação de desenhos de figura humana exerce alguma influência no processo de avaliação de um protocolo, apesar das diferenças observadas carecerem de significância estatística. No tocante à consistência interna, o resultado encontrado indicou que os 20 juízes tenderam a avaliar os 30 itens do instrumento no mesmo sentido. Em outras palavras, tenderam a seguir o mesmo entendimento no julgamento da presença ou ausência dos itens do DFH – Escala Sisto na avaliação do desenho.

Por fim, a análise da estatística kappa (fleiss) permitiu concluir que o instrumento se mostra fidedigno sob o ponto de vista da avaliação feita por juízes sem experiência na correção de desenhos da figura humana, na medida que o resultado apurado ficou acima do critério mínimo de precisão estabelecido na Resolução CFP 09/2018. Assim, o presente estudo corrobora os resultados apontados por Sisto (2015) para os estudos de precisão do instrumento na medida em que, caso fosse apresentado como um estudo de precisão à comissão consultiva do Sistema de Avaliação de Testes Psicológicos (SATEPSI), visando à aprovação do DFH – Escala Sisto, o instrumento seria considerado em condições de uso.

As contribuições do presente estudo são relevantes na medida em que, apesar dos resultados indicarem que, sob a perspectiva da precisão de juízes, o instrumento, considerando-se os trinta itens em conjunto, é fidedigno, apontou para questões que precisam ser melhor estudadas, tais como a efetiva influência ou não da experiência com desenhos da figura humana na avaliação de protocolos do teste, bem como a clareza e objetividade das descrições dos itens de avaliação do instrumento.

Com relação à influência ou não da experiência prévia na análise de desenhos de figura humana na avaliação do DFH – Escala Sisto, sugere-se a realização futura de um estudo em que se trabalhe com grupos contrastantes de avaliadores (um com e outro sem experiência), visando a correlacionar os resultados e apurar a ocorrência de possíveis diferenças significativas entre as avaliações realizadas. Com relação à clareza e objetividade das descrições dos itens de avaliação do instrumento, sugere-se a realização de uma pesquisa em que os participantes possam apresentar as possíveis dúvidas que tenham no entendimento dos itens de avaliação.

Apesar das contribuições, o estudo possui algumas limitações. Dentre elas, o fato de não realizar uma análise comparativa mais efetiva com os dados apresentados por Sisto (2015), principalmente devido à descrição muito sucinta do referido estudo. A sugestão dada de uma

pesquisa correlacional entre grupos contrastantes poderá elucidar a questão, conforme apontado.

Referências

ALVES, Gisele Aparecida da Silva; SOUZA, Maya Silva de.; BAPTISTA, Maklim Nunes. Validade e precisão de testes psicológicos. *In*: AMBIEL, R. A. M. et al. (org**.). Avaliação Psicológica: guia de consulta para estudantes e profissionais da psicologia**. São Paulo: Casa do Psicólogo, 2011. p.109-128.

ANASTASI, Anne.; URBINA, Susana. **Testagem psicológica**. 7 ed. Porto Alegre: Artes Médicas, 2000.

BISQUERRA, Rafael.; SARRIERA, Jorge Castellá; MARTÍNEZ, Francesc. **Introdução à estatística:** enfoque informático com o pacote estatístico SPSS. Porto Alegre: Artes Médicas. 2004.

CONSELHO FEDERAL DE PSICOLOGIA. **Resolução nº 009, de 25 de abril de 2018**. Estabelece diretrizes para a realização de Avaliação Psicológica no exercício profissional da psicóloga e do psicólogo, regulamenta o Sistema de Avaliação de Testes Psicológicos - SATEPSI e revoga as Resoluções n° 002/2003, nº 006/2004 e n° 005/2012 e Notas Técnicas n° 01/2017 e 02/2017. Disponível em: https://site.cfp.org.br/wp-content/uploads/2018/04/Resolução-CFP-nº-09-2018-com-anexo.pdf. Acesso em: 25 fev. 2020.

CROCKER, Linda; ALGINA, James. **Introduction to Classical and Modern Test Theory**. Belmont, CA: Wadsworth Group, 2009.

DANCEY, Christine. P., REIDY, John. **Estatística sem matemática para psicologia**. 3 ed. Tradução Lorí Viali. Porto Alegre: Artes Médicas, 2006.

DIAS, Augusto Rodrigues; COUTO, Gleiber; PRIMI, Ricardo. Avaliação da criatividade por meio da produção de metáforas. ***PSICO,*** v.40, n. 2, 2009. p. 210-219.

FONSECA, Ricardo; SILVA, Pedro; SILVA, Rita. Acordo inter-juízes: O caso do coeficiente kappa. **Laboratório de Psicologia**, v. *5*, n. 1, p.81-90, 2007. Disponível em: http://publicacoes.ispa.pt/index.php/lp/article/view/759. doi: https://doi.org/10.14417/lp.759. Acesso em: 15 fev. 2020.

GUILFORD, Joy Paul. **Fundamental Statistics in Psychology and Education**. 2 ed. Ed. New York: McGraw-Hill, 1950.

KLINE, Paul. **A handbook of test construction:** Introduction to psychometric design. New York, NY, US: Methuen. 1986.

LANDIS, J. Richard; KOCH, Gary G. The measurement of observer agreement for categorical data. **Biometrics**, v.*33, n.* 1, p. 159-174, 1977. Disponível em: http://www.jstor.org/stable/2529310. Acesso em: 20 fev. 2020.

MARÍN RUEDA, Fabian Javier; SUEHIRO, Adriana Cristina Boulhoça; SILVA, Marlene Alves da. Precisão entre avaliadores e pelo método teste-reteste no Bender-Sistema de Pontuação Gradual. **Psicologia: Teoria e Prática** [on line]1 v. 0, n. 1, p. 25-35, 2008. Disponível em: http://www.redalyc.org/articulo.oa?id=193818625003. Acesso em: 24 jan. 2020.

MATOS, Daniel Abud Seabra. Confiabilidade e Concordância entre juízes: aplicações na área educacional. **Est. Aval. Educ**., São Paulo, v. 25, n. 59, 2014. p. 298-324.

PASQUALI, Luiz. Psicometria. **Revista da Escola de Enfermagem da USP,** v. 43, n. spe. p. 992-999, 2009. Disponível em: http://sci-hub.tw/10.1590/S0080-62342009000500002. Acesso em: 24 jan. 2020.

SIEGEL, Sidney; CASTELLAN, N. John Junior. **Nonparametric statistics for the behavioral sciences**. 2. ed. New York: McGraw-Hill, 1988.

SISTO, Fermino Fernandes. **Desenho da Figura Humana – Escala Sisto**. São Paulo: Vetor Editora, 2005.

STEMLER, Steven. E. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. **Practical Assessment, Research & Evaluation,** v. 9, n. 4, 2004.

URBINA, Susana. **Fundamentos da testagem psicológica**. Porto Alegre: Artmed, 2007.