Revista de Psicologia
ISSN 2179-1740

PRECISÃO ENTRE JUÍZES PARA UM NOVO SISTEMA DE CODIFICAÇÃO DO TESTE DE WARTEGG

PRECISION BETWEEN JUDGES FOR A NEW WARTEGG TEST CODIFICATION SYSTEM

Fernando Pessotto[1]
Ricardo Primi[2]

Resumo

Dentre os parâmetros psicométricos necessários para o uso de testes psicológicos está a precisão que, de forma geral, indica a consistência dos resultados. Uma das formas de verificação deste parâmetro é a concordância entre juízes, expressa num índice que informa o grau de concordância de protocolos avaliados por diferentes profissionais. Com o objetivo de verificar a concordância entre juízes para o Teste de Wartegg, o presente estudo contou com 4 juízes avaliando 8 protocolos de forma independente (às cegas). Os resultados indicaram coeficientes kappa variando de 0,23 à 0,74 para os protocolos, sendo que, para algumas variáveis os valores foram mais altos. De forma geral, embora tenham se encontrado bom grau de concordância entre as codificações dos juízes para alguns critérios, outros demonstraram-se não estar claros de forma suficiente. Neste sentido, se fazem necessárias reformulações dos critérios evidenciados como ruins a fim de favorecer o desenvolvimento do sistema proposto.

Palavras-chave: Consistência interna; técnicas de autoexpressão; avaliação psicológica; Wartegg.

Abstract

Among necessary psychometric parameters for psychological testing is precision, which, in general, indicates results consistency. Agreement between judges is a way to verify that parameter, expressed in an index that informs the degree of protocols evaluated by different professionals. Aiming to check the agreement between judges for Wartegg test, this study counted on four judges evaluating eight protocols independently (blind review). Results indicates kappa coefficients varying from 0,23 to 0,74 for the protocols, and, for some variables values where higher. In a general way, although a good degree of agreement might be found trough judges evaluations for some criteria, others showed not being clear enough. In this sense, it is needed to reformulate those criteria which were evidenced as bad ones in order to improve the proposed system.

Keywords: Internal consistency, selfexpression techniques, psychological evaluation; Wartegg.

1 Fernando Pessotto - Centro Universitário Salesiano (UNISAL) - Psicólogo, Mestre e Doutor em Psicologia pelo Programa de Pós-Graduação Stricto Sensu em Psicologia na Universidade São Francisco - área de concentração em Avaliação Psicológica. É coordenador do Laboratório de Psicodiagnóstico e Neurociências Cognitivas (LaPeNC) e editor associado da Revista Sul Americana de Psicologia. E-mail: fpessotto@gmail.com
2 Ricardo Primi - Universidade São Francisco - Psicólogo pela PUC-Campinas, Doutor em Psicologia Escolar e do Desenvolvimento Humano pela Universidade de São Paulo com parte desenvolvida na Yale University (EUA). Coordenador do Laboratório de Avaliação Psicológica e Educacional (LabAPE). Professor Associado do Programa de Pós Graduação em Psicologia da Universidade São Francisco (Mestrado e Doutorado em Avaliação Psicológica).

INTRODUÇÃO

Os testes psicológicos são procedimentos sistemáticos tendo como objetivo registrar e mensurar aspectos psicológicos relativos ao funcionamento, patológico ou não dos indivíduos, para auxiliar no processo de tomada de decisão. Para favorecer este cenário, o Conselho Federal de Psicologia estabeleceu parâmetros mínimos para que os testes possam ser utilizados em território nacional prevendo estudos de evidências de validade, fidedignidade e normatização com amostras brasileiras. Neste período, muitos instrumentos tiveram o parecer desfavorável por não apresentarem condições mínimas para o uso (CFP, 2003).

Dentre os tipos de testes encontram-se as escalas de inteligência e de autorrelato, que apresentam questões preestabelecidas ligadas a uma habilidade ou um traço latente e as técnicas de autoexpressão, compostas por estímulos pouco estruturados, não ligados a um construto a priori, a fim de favorecer a livre expressão individual na formulação da resposta (Fensterseifer & Werlang, 2008; Meyer & Kurtz, 2006). Devido esta diferença de perspectiva na compreensão do sujeito, as técnicas de autoexpressão são alvo de críticas frequentes, sendo as mais comuns referentes à subjetividade na interpretação dos resultados ligada à falta de parâmetros psicométricos como normas, evidências de validade e fidedignidade (Garb, Wood, Lilienfeld, & Nezworski, 2002; Lilienfeld, Wood, & Garb, 2000; Villemor-Amaral & Werlang, 2008).

Lilienfeld e cols (2000) sinalizam que, para algumas técnicas de autoexpressão, os critérios utilizados para correção e consequente interpretação, muitas vezes se baseiam apenas num julgamento clínico, subestimando o rigor técnico e metodológico de procedimentos psicométricos, o que inviabiliza generalizar as interpretações pretendidas. Villemor-Amaral (2008) acrescenta que a interpretação feita a partir de técnicas de autoexpressão deve utilizar parâmetros estruturais claros provenientes das teorias.

Villemor-Amaral (2006, 2009) salienta a necessidade da ampliação de evidências de validade para as técnicas de autoexpressão criando-se sistemas que favoreçam maior concordância entre os avaliadores por meio de critérios objetivos. Segundo Villemor-Amaral e Pasqualini-Casado (2006) existe um número reduzido de estudos destas técnicas, baseados em parâmetros psicométricos.

Neste sentido, por se tratar de instrumentos com respostas abertas, um dos focos de interesse no que diz respeito à sua precisão, é a subjetividade do avaliador. Isto pode ser constatado por meio da concordância entre juízes, ou seja, se dois examinadores corrigirem um mesmo protocolo de maneira independente, às cegas, os resultados devem ser semelhantes, indicando que os critérios de codificação são claros de maneira suficiente para embasar a tomada de decisão (Pasquali, 2001, 2003; Urbina, 2007).

Uma das formas de verificar esta concordância é por meio do coeficiente *kappa* que indica o grau de consenso entre os avaliadores expresso num índice semelhante ao da correlação variando de 0 à 1. Embora o valor 0 não indique ausência total de concordância, quando maior o valor, maior o grau de concordância entre os diferentes juízes, consequentemente, infere-se que os critérios para atribuição de classificações das respostas são suficientemente claros. Valores acima de 0,60 são considerados aceitáveis, ou seja, um bom índice de concordância (Bisquerra, Martínez, & Sarriera, 2004; Stemler, 2004), embora para situações de avaliação psicologia espera-se índices mais altos.

Dentre as técnicas de autoexpressão está o Teste de Wartegg que, desde 2003, encontra-se com parecer desfavorável no Sistema de Avaliação dos Testes Psicológicos (SATEPSI), em parte, devido à limitações em seu sistema de codificação. A este respeito, Roivainen (2009) salienta que a criação de critérios de caráter mais objetivos referente às frequências de respostas e à própria interpretação das variáveis poderiam favorecer melhores índices de precisão, minimizando o caráter subjetivo da correção e interpretação.

Este tipo de delineamento se mostrou útil para o Rorschach-SC, por exemplo, contribuindo para maior confiabilidade nas inferências realizadas a partir de seus resultados (Meyer & Archer, 2001; Pianowski & Villemor-Amaral, 2010). Kinget (1991) salienta que os estudos realizados com o Teste de Wartegg contribuíram para que a técnica avançasse a níveis satisfatórios para a ocasião, mesmo concebendo que outras pesquisas devam prosseguir com o desenvolvimento do sistema de correção, pontuação e interpretação.

Gronnerod e Gronnerod (2012) também escrevem que interpretações realizadas a partir de resultados do Teste de Wartegg podem chegar a níveis comparáveis com outros métodos de avaliação. Os autores chamam a atenção ao fato de que diferentes métodos têm usado de sistemas de correção e codificação distintos e ainda estudados a partir de vários referenciais teóricos de personalidade, não favorecendo o acúmulo de conhecimento acerca do Teste de Wartegg. Os autores concluem que não há nenhuma razão para rejeita-lo como um método para avaliação da personalidade, porém, é necessária a construção de um sistema sólido, tradição em pesquisas gerando conhecimento acumulado para sua utilização e indicam

a necessidade de novas pesquisas baseadas nos estudos já existentes a fim de fortalecer o método empregado.

Sobre estudos de precisão realizados com o Teste de Wartegg, alguns autores (Gronnerod & Gronnerod, 2012; Roivainen, 2009; Silva, 2008) verificaram primeiramente, o reduzido número de produções em comparação com outras técnicas de autoexpressão. De forma geral, observaram ainda que a maioria apresenta índices psicométricos não satisfatórios, e outros, por terem sido encontrados apenas os resumos de apresentações em congressos, não indicam a partir de que dados ou variáveis foram evidenciados, o que torna difícil a compreensão dos resultados apresentados. Gronnerod e Gronnerod (2012) complementam ainda que a inexistência de um sistema internacional de correção e interpretação dificulta estudos transculturais que poderiam auxiliar em melhores resultados.

No Brasil, Ramon (2006) realizou uma pesquisa com o Teste de Wartegg no sistema proposto por Freitas (1993) com o objetivo de verificar sua precisão contando com a participação de 18 psicólogos com idades variando entre 27 e 54 anos. O autor utilizou os resultados de 8 avaliadores agrupados em 4 pares, realizando inicialmente uma análise do coeficiente de contingência de 5 protocolos e 27 variáveis.

Para as 27 variáveis analisadas, sequência, seletividade, cobertura, tamanho, movimento 1 e conteúdos apresentaram alta precisão (acima de 0,70), detalhes 1 e sombreado 1 precisão satisfatória (acima de 0,60), composição 4 e movimento 2, mediana (entre 0,50 e 0,60), nível de forma, composição 2, composição 3, expansão, detalhes 2, pressão, sombreado 2, qualidade de linha 2, frequência de conteúdos, tipo de conteúdo, organização e composição 1 obtiveram precisão baixas (abaixo de 0,50) e afinidade, perseveração, qualidade de linha 1, qualidade das linhas 3 e composição 5 não apresentaram correlação. Estes dados demonstraram-se desfavoráveis à precisão do instrumento, visto que para mais da metade das variáveis foram observadas baixa ou nenhuma concordância entre a avaliação dos juízes.

Em seguida, Ramon (2006) agrupou as variáveis em três características baseadas nos resultados dos protocolos, sendo elas relacionamento interpessoal, afetividade e controle emocional e ambição. As análises de correlação foram realizadas considerando-se o resultado final dos protocolos comparando-se cada juiz com os outros 17 e ainda considerando cada uma das 3 características para as os 5 protocolos. Os resultados indicaram para 15 pares de juízes, 10 ficaram abaixo de 0,65 e 5 variando ente 0,66 e 0,72. O autor argumenta que estes índices também não foram considerados relevantes para demonstrar a precisão da interpretação, porém indicam que esta técnica pode obter dados satisfatórios desde que os avaliadores recebam critérios precisos para a interpretação de cada variável do teste, sendo esta a principal carência nos manuais dos testes, a falta de diretrizes objetivas e inequívocas para análise e interpretação (Ramon, 2006).

Posteriormente Alves, Dias, Sardinha e Conti (2010) avaliaram a adequação dos critérios de classificação propostos por Berlinck (2000). Participaram 191 sujeitos com idades variando entre 18 e 54 anos (M=28,6; DP=8,71) sendo 100 do sexo feminino. Para a correção dos protocolos os autores contaram com 2 juízes. Os juízes receberam treinamento de acordo com os critérios propostos por Berlinck (2000) utilizando 10 protocolos do Teste de Wartegg. Eles indicaram a presença ou ausência dos aspectos formais avaliados, a saber, pressão do lápis, tipo de linha, continuidade da linha, qualidade da linha, tamanho, sombreado, movimento e transparência, para os 8 quadros dos protocolos discutindo posteriormente as concordâncias e discordâncias com o objetivo de diminuir as dúvidas referentes aos critérios de classificação.

Os coeficientes *kappa* variaram de 0,79 a 1 sendo os maiores para movimento no Quadro 7 (1) e transparência no Quadro 1 (1) e os menores para transparência no Quadro 8 (0,66) e no Quadro 7 (0,70). As médias dos coeficientes para os oito campos variaram entre 0,84 (pressão do lápis) e 0,90 (linha descontínua). Os autores concluem que os resultados são importantes, pois os índices de concordância entre os juízes foram altos indicando que os avaliadores utilizaram os mesmos critérios para a correção. Até os índices mais baixos foram considerados significativos (Alves et al., 2010). Vale ressaltar, contudo, que o estudo contou com 2 juízes que corrigiram juntos 10 protocolos previamente, discutindo os critérios, antes da avaliação dos outros, o que pode ter favorecido os altos índices não caracterizando a independências das classificações, ou seja, correção às cegas.

Neste sentido, o presente estudo tem como objetivo avaliar a precisão para o conjunto de variáveis proposto por Pessotto (2015) para o Teste de Wartegg por meio da concordância entre juízes.

MÉTODO

Participantes

Participaram deste estudo quatro juízes, sendo dois com experiência em técnicas de autoexpressão, dos quais, um é doutor e o outro

doutorando e dois com experiência em avaliação psicológica de forma geral, ambos doutores.

Protocolos

Foram utilizados 8 protocolos do Teste de Wartegg aplicados de acordo com o conjunto de variáveis proposto por Pessotto (2015). Estes protocolos referem-se à quatro pacientes com diagnóstico de esquizofrenia anos? e os outros de sujeitos sem histórico de diagnóstico psiquiátrico, todos com idade variando entre 24 e 40 anos.

Procedimentos

Os juízes foram convidados a participar do estudo e, após o aceite e assinatura do Termo de Consentimento Livre e Esclarecido, receberam os 8 protocolos do Teste de Wartegg por meio digital, juntamente com os critérios de codificação do sistema desenvolvido por proposto por Pessotto (2015). Cada um teve como tarefa codificar os 8 protocolos de maneira independente.

Plano de análise de dados

Foram realizadas análises de correlação de Pearson entre as codificações dos juízes e para verificação da fidedignidade entre elas empregou-se o coeficiente *kappa.* As análises foram feitas considerando os protocolos de forma geral e à partir dos agrupamentos das variáveis nos critérios, à saber, localização, tamanho, conteúdo, qualidade do objeto, qualidade formal, características particulares, códigos especiais, pressão do traço, tipo de traço, repetição, sequência de execução e tipo de título. Para as análise foram utilizados o *software M-Plus* e o *Statistical Package for Social Sciences* (SPSS) na versão 21.

RESULTADOS E DISCUSSÃO

Na busca por evidências favoráveis de fidedignidade para o Teste de Wartegg, incialmente empregou-se a correlação de Pearson entre os protocolos avaliados pelos juízes. A análise foi empregada no nível do protocolo, ou seja, as classificações de cada um dos 8 quadros foram agrupadas. Os resultados podem ser observados na Tabela 1.

É possível observar altas correlações entre 3 juízes com magnitudes variando de 0,91 à 0,94. Apenas o juiz 3, que não possui experiência com técnicas de autoexpressão, apresentou correlação moderada, 0,53 com todos os outros juízes. Estes índices apontam inicialmente boa concordância entre as codificações realizadas, ao mesmo tempo que que, alguns critérios devem ser observados, visto que um dos juízes apresentou menor correlação com os demais. Em seguida empregou-se o coeficiente *kappa* para verificar a concordância entre os juízes. Os resultados podem ser observados na Tabela 2 para todos os protocolos agrupados e na Tabela 3 para os protocolos analisados separadamente.

De acordo com a Tabela 2 é possível observar que o coeficiente *kappa* para 3 pares de juízes demonstraram-se satisfatórios, conforme a Resolução 02/2003 do CFP tendo apresentado valores acima de 0,60, sendo alguns superiores à 0,90. O juiz 3, conforme já observado na análise de correlação, foi o que apresentou menor concordância entre as codificações apresentando índices não aceitáveis, sendo todos 0,23. Este mesmo padrão é observado nas concordâncias realizadas considerando os protocolos separadamente conforme observado na Tabela 3. Este fator pode se dar em razão deste ser um dos juízes sem experiência no uso das técnicas de autoexpressão. Embora a prática com estas técnicas sejam importantes para uma análise mais assertiva, a melhora na descrição dos critérios auxilia na concordância até mesmo para avaliadores com menos experiência, como pode ser visto em técnicas como Rorschach no Sistema Compreensivo ou R-PAS (Meyer & Archer, 2001; Pianowski & Villemor-Amaral, 2010).

Por fim, empregou-se a análise do coeficiente kappa para os agrupamentos de variáveis referentes às classificações, à saber localização, tamanho, conteúdo, qualidade dos objetos, qualidade formal, códigos particulares, características especiais, pressão do lápis, tipo de linha, repetição, sequência e título. Neste sentido, os agrupamentos das variáveis descritas anteriormente, oriundos dos 8 protocolos, foram agrupados para cada um dos juízes e comparados. Na Tabela 4 é possível verificar os resultados.

De acordo com a Tabela 4 é possível observar que de forma geral, os coeficientes não foram satisfatórios. Embora seja possível encontrar concordância total entre alguns juízes para as variáveis tamanho, conteúdo, qualidade formal códigos particulares, características especiais, repetição e sequência, os baixos níveis de concordância entre outros juízes indicam fragilidade nos critérios de classificação. Resultados semelhantes foram encontrados por Ramon (2006) utilizando os critérios de codificação conforme propostos por Kinget.

A classificação para qualidade de objetivos foi que apresentou pior índice não tendo sido encontrado concordância entre os juízes. A pressão do lápis foi a

segunda com menor concordância na interpretação geral tendo chegado ao máximo de 0,25 entre 3 juízes. Os códigos qualidade formal, códigos particulares, repetição e título, apesar de apresentarem concordância total para ao menos 1 par de juízes, quando verificada a concordância entre os outros pares, os valores se diferenciaram de forma acentuada, apresentando grande inconsistência para estas classificações.

Para localização, apesar de apresentar alguma concordância entre os juízes, não é significativa, ou seja, está abaixo do nível considerado aceitável. Por sua vez, o tipo de linha apresenta índices favoráveis entre 2 dos juízes, embora seja o inverso para os outros pares de juízes. Por fim, as que apresentaram maiores índices foram tamanho, conteúdo, características especiais e sequência tendo sido verificadas concordância total entre 3 pares de juízes, porém, mesmo assim, o quarto par não apresentou concordância para nenhuma delas. Um fato importante de se verificar é que o juiz 3 não apresentou concordância em nenhum dos pares a que foi submetido, podendo indicar falta de compreensão nos critérios de forma geral. Mesmo assim, ele foi mantido nas análises e buscar-se-á, a partir dos resultados, melhorar os critérios de classificação do sistema proposto.

Algumas das dificuldades na classificação, como apontados pelos juízes, pode ser em decorrência ao fato dos protocolos terem sido enviados digitalmente, o que dificulta a avaliação da pressão e tipo de linha, por exemplo. Outras variáveis como tamanho e localização podem não ter sido satisfatoriamente explicadas por se referirem ao "desenho realizado", não especificando se o estímulo original do quadro deve ser considerado para a classificação. No que diz respeito ao conteúdo, foram utilizados os códigos presentes no Rorschach (R-PAS) conforme indica Pessotto (2015) e muitos conteúdos que demonstraram-se frequentes, foram classificados como outros (NC), característica também apontada pelos juízes.

Embora estes resultados demonstrem-se desfavoráveis para o Teste de Wartegg, eles orientam futuros estudos ao sinalizarem os ajustes necessários em sua estrutura para se estabelecer um sistema coeso que favoreça a concordância entre avaliadores conforme indica Villemor-Amaral (2006, 2009) quando refere-se às técnicas de autoexpressão. Os critérios ora apresentados parecem não ser claros o suficiente para que sejam compreendidos de forma satisfatória na classificação dos desenhos, não favorecendo que a interpretação seja realizada por diferentes avaliadores, portanto, não se ajustam à proposta de Roivainen (2009) de serem objetivos a fim de minimizar o caráter subjetivo no momento da classificação das respostas.

Estes resultados podem ainda ser comparados àqueles encontrados por Gronnerod e Gronnerod (2012) em um estudo de meta-análise realizado com o Teste de Wartegg, em que verificaram índices baixos ou insignificantes relativos à precisão do instrumento. Os autores salientam que este não deve ser um motivo para que a técnica seja abandonada, mas sim um incentivo à novas pesquisas a fim de aperfeiçoar o sistema de codificação.

Por fim, embora Alves et al (2010) tenham encontrado resultados favoráveis para os critérios de avaliação propostos por Berlinck (2000), ressalta-se que os 2 juízes que participaram do estudo receberam treinamento em 10 protocolos tendo a possibilidade de sanar dúvidas que pudessem surgir no momento das codificações, o que desconfigura a avaliação às cegas. Este procedimento não foi adotado no presente estudo, procurando aproximar-se da situação real de avaliação, configurada pelo uso do instrumento por diferentes psicólogos, sem que, necessariamente, estes tenham possibilidade de tirar dúvidas acerca do sistema.

Embora estes resultados demonstrem-se desfavoráveis para o Teste de Wartegg, eles orientam futuros estudos ao sinalizarem os ajustes necessários em sua estrutura para se estabelecer um sistema coeso que favoreça a concordância entre avaliadores conforme indica Villemor-Amaral (2006, 2009). Os critérios ora apresentados parecem não ser claros o suficiente para que sejam compreendidos de forma satisfatória na classificação dos desenhos, não favorecendo que a interpretação seja realizada por diferentes avaliadores, portanto, não se ajustam à proposta de Roivainen (2009) de serem objetivos a fim de minimizar o caráter subjetivo no momento da classificação das respostas.

Por fim, embora Alves et al (2010) tenham encontrado resultados favoráveis para os critérios de avaliação propostos por Berlinck (2000), ressalta-se que os 2 juízes que participaram do estudo receberam treinamento em 10 protocolos tendo a possibilidade de sanar dúvidas que pudessem surgir no momento das codificações, o que desconfigura a avaliação às cegas. Este procedimento não foi adotado no presente estudo, procurando aproximar-se da situação real de avaliação, configurada pelo uso do instrumento por diferentes psicólogos, sem que, necessariamente, estes tenham possibilidade de tirar dúvidas acerca do sistema.

CONSIDERAÇÕES FINAIS

Esta pesquisa teve como finalidade verificar a precisão do Teste de Wartegg para o conjunto de variáveis proposto por Pessotto (2015) pela concordância entre juízes. De maneira geral, embora

tenha-se encontrado boas correlações entre as correções realizadas pelos juízes, os coeficientes kappa não mostraram-se satisfatórios. Embora tenham indicado bons índices de concordâncias para os protocolos de forma geral, as análises individuais das variáveis demonstraram-se, mesmo que parcialmente, insatisfatórias, indicando necessidade de reformulação do critérios de codificação. As adequações dos critérios e novos estudos já estão sendo realizados (Pessotto, 2015) com a finalidade de elaborar um sistema com critérios objetivos a fim de favorecer melhores índices de precisão, conforme indicado por Roivainen (2009) e Kinget (1991).

Referências

Alves, I. C. B., Dias, A. R., Sardinha, L. S., & Conti, F. D. (2010). Precisão entre juízes na avaliação dos aspectos formais do teste de Wartegg. *Aletheia*, (31), 54–65.

Berlinck, V. (2000). *O teste de completamento de desenos Wartegg em universitarios de São Paulo.* (Dissertação de Mestrado). Universidade de São Paulo, São Paulo.

Bisquerra, R., Martínez, F., & Sarriera, J. C. (2004). *Introdução à estatística: enfoque informático com o pacote estatístico SPSS*. Porto Alegre: Artes Médicas.

Conselho Federal de Psicologia - CFP. (2003). *Resolução nº* 002/2003 de 24 de março. Brasília, DF: CFP.

Fensterseifer, L., & Werlang, B. S. G. (2008). Apontamentos sobre o status científico das técnicas projetivas. In A. E. de Villemor-Amaral & B. S. G. Werlang (Eds.), *Atualizações em Métodos Projetivos para Avaliação Psicológica* (pp. 15–33). São Paulo: Casa do Psicólogo.

Garb, H. N., Wood, J. M., Lilienfeld, S. O., & Nezworski, M. T. (2002). Effective use of projective techniques in clinical practice: Let the data help with selection and interpretation. *Professional Psychology: Research and Practice*, *33*(5), 454–463. http://doi.org/10.1037/0735-7028.33.5.454

Gronnerod, J. S., & Gronnerod, C. (2012). The Wartegg Zeichen Test: a literature overview and a meta-analysis of reliability and validity. *Psychological Assessment*, *24*(2), 476–489. http://doi.org/10.1037/a0026100

Kinget, G. M. (1991). O teste de Completamento de Figuras. In E. F. Hammer (Ed.), *Aplicações Clínicas dos Desenhos Projetivos*. São Paulo: Casa do Psicólogo.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The Scientific Status of Projective Techniques. *Psychological Science in the Public Interest*, *1*(2), 27–66. http://doi.org/10.1111/1529-1006.002

Meyer, G. J., & Archer, R. P. (2001). The hard science of Rorschach research: What do we know and where do we go? *Psychological Assessment*, *13*(4), 486–502. http://doi.org/10.1037/1040-3590.13.4.486

Meyer, G. J., & Kurtz, J. E. (2006). Advancing Personality Assessment Terminology: Time to Retire "Objective" and "Projective" As Personality Test Descriptors. *Journal of Personality Assessment*, *87*(3), 223–225. http://doi.org/10.1207/s15327752jpa8703_01

Pasquali, L. (2001). Parâmetros psicométricos dos testes psicológicos. In L. Pasquali (Ed.), *Técnicas de Exame Psicológico – TEP* (pp. 11–136). São Paulo: Casa do Psicólogo.

Pasquali, L. (2003). *Psicometria: teoria dos testes na Psicologia e na Educação.* Petrópolis: Vozes.

Pessotto, F. (2015). *Teste de Wartegg (Sistema em Desenvolvimento)*. Itatiba: Laboratório de Avaliação Psicológica e Educacional (LabAPE) - Universidade São Francisco (USF).

Pianowski, G., & Villemor-Amaral, A. E. de. (2010). Location and formal quality of the Rorschach-SC in Brazil: validity with non-patient sample. *Psico-USF*, *15* (3), 333–343. http://doi.org/10.1590/S1413-82712010000300007

Ramon, R. R. (2006). *Wartegg: Precisão entre Avaliadores e Evidência de Validade com o Método de Rorschach* (Dissertação de Mestrado). Universidade São Francisco, Itatiba.

Roivainen, E. (2009). A Brief History of the Wartegg Drawing Test. *Gestalt Theory*, *31*(1), 55–71.

Silva, M. C. de V. (2008). O teste de completamento de desenhos de Wartegg (WZT). In Anna Elisa & B. S. G. Werlang (Eds.), *Atualizações em métodos projetivos para avaliação psicológica*. São Paulo: Casa do Psicólogo.

Stemler, S. E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation*, *9*(4).

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

Villemor-Amaral, A. E. de. (2006). Desafios para a cientificidade das técnicas projetivas. In A. P. P. Noronha, A. A. A. dos Santos, & F. F. Sisto (Eds.), *Facetas do fazer em avaliação psicológica* (pp. 163–171). São Paulo: Vetor.

Villemor-Amaral, A. E. de. (2008). A validade teórica em avaliação psicológica. *Psicologia: Ciência E Profissão*, *28*(1), 98–109. http://doi.org/10.1590/S1414-98932008000100008

Villemor-Amaral, A. E. de. (2009). Métodos Projetos em Avaliações Compulsórias: indicadores e perfis. In C. S. Hutz (Ed.), *Avanços e polêmicas em avaliação psicológica*. São Paulo: Casa do Psicólogo.

Villemor-Amaral, A. E. de, & Pasqualini-Casado, L. (2006). A cientificidade das técnicas projetivas em debate. *Psico-USF*, *11*(2), 185–193. http://doi.org/10.1590/S1413-82712006000200007

Villemor-Amaral, A. E. de, & Werlang, B. S. G. (2008). *Atualizações em métodos projetivos para avaliação psicológica*. Casa do Psicólogo.

Notas

[1] Os autores contaram com o apoio financeiro da CAPES

Lista de Tabelas

Tabela 1 - Correlação de Pearson entre as classificações realizadas pelos juízes

	Juiz 1	Juiz 2	Juiz 3
Juiz 2	0,94		
Juiz 3	0,53	0,53	
Juiz 4	0,91	0,92	0,53

Tabela 2 - Coeficiente kappa entre os juízes para todos os protocolos agrupados

	Juiz 1	Juiz 2	Juiz 3
Juiz 2	0,74		
Juiz 3	0,23	0,23	
Juiz 4	0,68	0,72	0,23

Tabela 3 - Coeficiente kappa entre os juízes para os 8 protocolos separadamente

Protocolo 1			
	Juiz 1	**Juiz 2**	**Juiz 3**
Juiz 2	0,94		
Juiz 3	0,33	0,37	
Juiz 4	0,94	0,94	0,37

Protocolo 2			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	0,94		
Juiz 3	0,37	0,41	
Juiz 4	0,94	0,88	0,38

Protocolo 3			
	Juiz 1	**Juiz 2**	**Juiz 3**
Juiz 2	0,85		
Juiz 3	0,24	0,21	
Juiz 4	0,81	0,92	0,20

Protocolo 4			
	Juiz 1	**Juiz 2**	**Juiz 3**
Juiz 2	0,92		
Juiz 3	0,49	0,49	
Juiz 4	0,89	0,87	0,46

Protocolo 5			
	Juiz 1	**Juiz 2**	**Juiz 3**
Juiz 2	0,81		
Juiz 3	0,38	0,39	
Juiz 4	0,73	0,73	0,32

Protocolo 6			
	Juiz 1	**Juiz 2**	**Juiz 3**
Juiz 2	0,94		
Juiz 3	0,31	0,34	
Juiz 4	0,79	0,79	0,36

Protocolo 7			
	Juiz 1	**Juiz 2**	**Juiz 3**
Juiz 2	0,85		
Juiz 3	0,35	0,35	
Juiz 4	0,82	0,86	0,40

Protocolo 8			
	Juiz 1	**Juiz 2**	**Juiz 3**
Juiz 2	0,84		
Juiz 3	0,30	0,31	
Juiz 4	0,76	0,70	0,29

Tabela 4 - Coeficiente kappa para as 12 classificações do novo sistema do Teste de Wartegg

Localização			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	0,29		
Juiz 3	0	0	
Juiz 4	0,09	0,11	0

Características Especiais			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	1		
Juiz 3	-0,08	-0,08	
Juiz 4	1	1	-0,08

Tamanho			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	1		
Juiz 3	-0,12	-0,12	
Juiz 4	1	1	-0,12

Pressão do Lápis			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	0,25		
Juiz 3	0	0	
Juiz 4	0,25	0,25	0

Conteúdo			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	1		
Juiz 3	0,45	0,45	
Juiz 4	1	1	0,45

Tipo de Linha			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	0,69		
Juiz 3	-0,14	-0,14	
Juiz 4	0,43	0,69	-0,06

Qualidade dos Objetos			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	0		
Juiz 3	0	0	
Juiz 4	0	0	0

Repetição			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	1		
Juiz 3	0	0	
Juiz 4	0,33	0,33	0

Qualidade Formal			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	0,25		
Juiz 3	0,25	0,25	
Juiz 4	1	0,25	0,25

Sequência			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	1		
Juiz 3	0,33	0,33	
Juiz 4	1	1	0,33

Códigos Particulares			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	0,50		
Juiz 3	0,05	0,05	

Título			
	Juiz 1	Juiz 2	Juiz 3
Juiz 2	1		
Juiz 3	0	0	

RECEBIDO EM: 18/01/2017
PRIMEIRA DECISÃO EDITORIAL : 18/05/2017
VERSÃO FINAL: 23/05/2017
APROVADO EM: 31/08/2017