PSICOLOGIA ARGUMENTO

ISSN 0103-7013

doi: http://dx.doi.org/10.7213/psicolargum.43.122.AO07

Concordância entre juízes para três variáveis do Crisi Wartegg System (CWS) no contexto brasileiro

Agreement between njudges for three variables of the Crisi Wartegg System (CWS) in the brazilian context

Ligiana Koller Gorgonio
FIG Unimesp
https://orcid.org/0000-0002-7204-9984
ligianakoller@yahoo.com.br

Augusto Rodrigues Dias
FIG Unimesp
https://orcid.org/0000-0003-4143-6890

Resumo

Os testes psicológicos, para que possam ser utilizados no contexto brasileiro, devem apresentar em seus manuais, estudos de fidedignidade e evidências de validade aprovados pelo Sistema de Avaliação dos Testes Psicológicos (SATEPSI) do Conselho Federal de Psicologia. Em razão dessa condição, o Teste de Wartegg encontra-se com o uso suspenso desde o ano de 2003, devido à ausência de estudos dessa natureza. Diante deste cenário, o presente estudo objetivou avaliar fidedignidade por meio da concordância entre juízes do Crisi Wartegg System (CWS), que se constitui em um sistema de interpretação para o Teste de Wartegg amplamente utilizado na Itália e Estados Unidos. Para tanto, dois juízes avaliaram 90 protocolos, seguindo as orientações de avaliação propostas pelo CWS, para três variáveis desse sistema, a saber: Qualidade Afetiva, Qualidade Formal e Caráter Evocativo. Foram calculados o r de Pearson, a estatística Kappa de Cohen e o Coeficiente de Correlação Interclasse (CCI). Os resultados indicaram, independentemente do recurso estatístico utilizado, alta confiabilidade entre os juízes ao mensurarem as variáveis em foco, Além do exposto, ao utilizar três estatísticas diferentes, foi possível verificar a força, direção e consistência das avaliações, o que confere robustez significativa ao estudo. Sugere-se o desenvolvimento de outros estudos de fidedignidade e evidências validade objetivando-se o uso do CWS no contexto brasileiro.

Palavras-chave: técnicas projetivas, Teste de Wartegg, medidas de Fidedignidade, concordância entre juízes

Resumen

For psychological tests to be used in the Brazilian context, their manuals must include reliability studies and validity evidence approved by the Psychological Testing Assessment System (SATEPSI) of the Federal Council of Psychology. Due to this requirement, the Wartegg Test has been suspended from use since 2003 due to a lack of such studies. Given this scenario, the present study aimed to assess reliability through inter-rater agreement using the Crisi Wartegg System (CWS), an interpretation system for the Wartegg Test widely used in Italy and the United States. To this end, two raters evaluated 90 protocols, following the CWS assessment guidelines for three variables in this system: Affective Quality, Formal Quality, and Evocative Character. Pearson's r, Cohen's Kappa statistic, and the Intraclass Correlation Coefficient (ICC) were calculated. The results indicated high reliability between raters in measuring the variables in focus, regardless of the statistical method used. Furthermore, using three different statistics allowed for an assessment of the strength, direction, and consistency of the evaluations, which provides significant robustness to the study. Additional reliability studies and validity evidence are suggested to support the use of the CWS in the Brazilian context.

Keywords: *projective techiniques, Wartegg Test, reliability measures, Inter-rater agrément.*

Abstract

Para que las pruebas psicológicas puedan ser utilizadas en Brasil, sus manuales deben incluir estudios de fiabilidad y evidencias de validez aprobadas por el Sistema de Evaluación de Pruebas Psicológicas (SATEPSI) del Consejo Federal de Psicología. Debido a esta condición, el Test de Wartegg está suspendido desde 2003 por falta de estudios de esta naturaleza. Ante este escenario, el objetivo de este estudiofue evaluar la fiabilidade del Crisi Wartegg System (CWS), que es un sistema de interpretación del Test de Wartegg ampliamente utilizado em Italia y Estados Unidos. Para ello, dos jueces evaluaron 90 protocolos, siguiendo las pautas de evaluación propuestas por el CWS, para três variables de este sistema, a saber: Calidad Afectiva, Calidad Formal y Carácter Evocador. Se calcularon la r de Pearson, el Kappa de Cohen y el Coeficiente de Correlación Interclase (ICC). Además, al utilizar tres estadísticos diferentes, fue posible verificar la fuerza, la dirección y la coherencia de las evaluaciones, lo que confiere una solidez significativa al estudio. Se sugere la realización de nuevo sestudios de fiabilidad y validez con vistas a la utilización del CWS em el contexto brasileño.

Palabras clave: *técnicas proyectivas, Test de Wartegg, medidas de fiabilidad, concordancia entre jueces*

Introdução

A medição de variáveis subjetivas, notadamente na área da avaliação psicológica, tem relação direta com as técnicas/testes de autoexpressão ou projetivos. Esses, se caracterizam pelo fato de possuírem estímulos vagos ou ambíguos, bem como instruções relativamente amplas. Como consequência, leva o indivíduo avaliado a fazer uso de referenciais próprios para organizar as informações e produzir uma resposta que permita o acesso a características sobre seu funcionamento psíquico (Villemor-Amaral & Cardoso, 2019). Assim, pode-se considerar que os testes projetivos são procedimentos psicológicos especialmente projetados para evocar do sujeito uma resposta que envolva a expressão de seus pensamentos internos, fantasias, desejos e percepções de si mesmo e do mundo ao seu redor (Ogbodo-Adoga, 2020).

Outro aspecto marcante dessas técnicas, refere-se ao processo avaliativo ou interpretativo que possuem. Geralmente, esse processo carrega uma boa dose de subjetividade do avaliador (Tarigan & Fadillah, 2022), que pode influenciar a avaliação feita. Em função dessainfluência,o processo avaliativo/interpretativo tem recebido constantes críticas, especialmente em relação aos critérios utilizados para correção. Tais críticas ocorrem pelo fato desses critérios estarem baseados apenas em um julgamento clínico que coloca em segundo plano o rigor técnico e metodológico de procedimentos psicométricos, o que acarreta a impossibilidade de se generalizar as interpretações pretendidas (Pessotto & Primi, 2017). Assim, a forma existente para contornar a problemática exposta indica que tais técnicas precisam apresentar critérios de correção bem claros e definidos, com o intuito de possibilitar a diferentes avaliadores coincidir nas avaliações que realizam de um mesmo protocolo de respostas.

Para garantir que os critérios de correção/interpretação tenham clareza e objetividade, bem como diminuam ao máximo possível a influência da subjetividade na avaliação, é comum recorrer-se a estratégia denominada concordância entre avaliadores. Essa estratégia busca determinar em que medida dois avaliadores coincidem nos seus diagnósticos relativamente a um conjunto de observações (Konstantinidis, Lisa & Xin Gao, 2022), a partir de critérios previamente definidos e utilizados para a avaliação. Em outras palavras, essa concordância quantifica a consistência das pontuações obtidas dos mesmos protocolos em classificações independentes repetidas por diferentes codificadores (Schneider, Bandeira & Meyer, 2020). Desse modo, a concordância entre

as classificações pode então ser usada como uma indicação da confiabilidade das classificações feitas pelos avaliadores (de Raadt, Warrens, Bosker, & Kiers, 2021).

Em termos estatísticos, o cálculo da concordância entre avaliadores pode ser efetuado por intermédio de diferentes recursos, como por exemplo a Correlação de Pearson *(r)* ou Spearman *(ρ)*, entendidos como uma medida de uma relação entre duas variáveis quantitativas e categóricas (Alsaqr, 2021); a estatística Kappa, que mede a concordância entre observadores corrigida pelo acaso (Bakeman, 2023), podendo essa apresentar diferentes versões (Kappa de Cohen (*κ*), o Kappa de Fleiss *(K)*, o Kappa ponderado (*κw*)). Além desses, também pode ser utilizado o Coeficiente de Correlação Intraclasse (ICC). Esse recurso permite medir a concordância geral entre duas ou mais medidas envolvendo variáveis quantitativas, obtidas com diferentes instrumentos de medida ou avaliadores (Han, 2020; Pérez & Martin, 2023).

Geralmente, a interpretação para essas estatísticas pode ser expressa da seguinte forma: as Correlações de Pearson (*r*) ou Spearman (*ρ)* são interpretadas a partir de uma classificação como exemplo, + 0,1 a + 0,3 ou - 0,1 a - 0,3, como uma correlação fraca; + 0,4 a + 0,6 ou - 0,4 a - 0,6, como uma correlação moderada; + 0,7 a + 0,9 ou - 0,7 a - 0,6, como uma correlação forte e, + 1,0 ou - 1,0, como sendo uma correlação perfeita (Akoglu, 2018). As estatísticas kappa são mensuradas por $\kappa < 0$ indicando concordância menor que o acaso; de $0{,}01 \leq \kappa \leq 0{,}20$, como uma concordância leve; de $0{,}21 \leq \kappa \leq 0{,}40$, uma concordância razoável; $0{,}41 \leq \kappa \leq 0{,}60$, concordância moderada; $0.61 \leq \kappa \leq 0.80$, concordância substancial, e; de $0{,}81 \leq \kappa \leq 1$: como uma concordância quase perfeita (Dettori & Norvell, 2020).O ICC, tendo como base o intervalo de confiança de 95%, é interpretado, quando < 0,5 como uma confiabilidade ruim; de $> 0{,}5 \leq 0{,}75$, confiabilidade moderada; de $> 0{,}75 \leq 0{,}9$, uma confiabilidade boa, e; > 0,9 ou mais, como uma confiabilidade excelente (Koo & Li, 2016).

Retornando ao universo de técnicas/testes projetivos ou de autoexpressão, destaca-se nesse trabalho o Teste de Completamento de Desenhos de Wartegg (WDCT) ou, como é mais conhecido no Brasil, Teste de Wartegg (WZT). No ano de 2003, esse instrumento teve seu uso suspenso pelo Conselho Federal de Psicologia (CFP), em razão dos manuais existentes à época não apresentarem estudos que atestassem a fidedignidade e evidências de validade das principais interpretações propostas (Alves et. al., 2010; Pereira, 2006). Salienta-se que o WZT era um instrumento amplamente utilizado

nacionalmente, em especial com um uso efetivo no contexto organizacional, mais especificamente em processos de seleção de pessoal (Pereira, 2006).

Desde antes e, principalmente após sua suspensão, alguns pesquisadores dedicaram-se a desenvolver estudos de fidedignidade e evidências de validade para alguns dos sistemas de avaliação existentes e/ou a criação de novos sistemas. Como exemplos, podem ser citados os trabalhos de Berlinck (2000, 2006), Ramon (2006), Pereira (2006), Souza, Primi & Miguel (2007), Alves et. al. (2010), Pessotto (2015), Pessotto & Primi (2017, 2018). Entretanto, apesar dos esforços, os estudos citados ainda não foram suficientes para retirar o WZT da lista de testes com parecer desfavorável emitida pelo CFP. Essa condição se deve, em parte, aos resultados não satisfatórios encontrados para os sistemas já existentes e, por outro lado, no caso de novos sistemas, ainda não terem seus estudos concluídos.

Assim, dada a condição atual do WZT no Brasil, buscou-se encontrar no cenário internacional um sistema de interpretação mais bem desenvolvido e com parâmetros psicométricos estruturados. Como resultado, encontrou-se o Crisi Wartegg System (CWS). Esse sistema foi desenvolvido inicialmente na Itália a partir do uso concomitantemente do Rorschach e Wartegg, tendo como base categorias de pontuação comumente usadas e previamente validadas no Rorschach (Crisi, 2018). No tocante aos estudos de fidedignidade, diversos são os que recorreram à concordância entre avaliadores.

Em termos desses estudos, podem se citar o primeiro estudo (publicado tanto na primeira (1998) quanto na segunda (2007) edição do manual do CWS em italiano), que avaliou a confiabilidade entre três pares de juízes na avaliação que realizaram independentemente de 18 protocolos (Crisi, 2018). O objetivo foi avaliar a eficácia e o impacto das orientações para a pontuação descritas no manual, tendo como base duas condições: a) realizar a pontuação dos protocolos sem fazer referência às diretrizes de pontuação que constam no manual do CWS, contando somente com a experiências anteriores na correção do WZT, e; b) pontuar os protocolos seguindo as orientações do manual do CWS. Os avaliadores foram divididos em três grupos de acordo com a experiência que possuíam em relação ao modo de correção do CWS, a saber: especialistas (profissionais com pelo menos cinco anos de experiência com o CWS), práticos (profissionais com dois anos de experiência), e iniciantes (profissionais que recentemente

concluíram o treinamento no CWS). As comparações foram feitas entre os pares Especialista-Especialista, Especialista-prático e Especialista-Iniciante, avaliando-se as variáveis Caráter Evocativo (CE+%), Qualidade Afetiva (QA+%), Qualidade Formal (QF+%) e Códigos Especiais. Para calcular o grau de concordância entre os juízes foi utilizado o coeficiente Kappa de Cohen corrigido para concordância casual *(k)*. Os resultados encontrados para a primeira condição foram: o maior grau de confiabilidade entre juízes na comparação Especialista-Especialista (k = 0,84 a 0,88, classificado como "quase perfeito"); menor concordância entre Especialista-Prático (k = 0,68 a 0,78, "substancial") e entre Especialista-Novato (k = 0,55 a 0,59, "moderado"). Para a segunda condição a concordância entre os juízes, em média, foi alta entre os pares, ou seja: Especialista-Especialista (k = 0,91, "quase perfeito) e Especialista-Prático (k = 0,84, "quase perfeito") e substancial entre Especialista-Novato (k = 0,69). Crisi (2018) ressaltou que, embora o estudo tenha revelado um alto nível de concordância entre os examinadores especialistas, mesmo os níveis mais baixos de confiabilidade encontrados não afetariam significativamente a interpretação dos resultados do teste, haja vista que a intepretação é feita com base em intervalos de valores normativos.

Daini et al. (2006) investigaram estados afetivos e impulsividade em pacientes em tratamento para transtornos alimentares. Para tanto, realizaram análises de confiabilidade entre dois pares de juízes independentes. Esses avaliaram 40 protocolos randomicamente selecionados. Cinco categorias específicas de pontuação foram consideradas: Caráter Evocativo (CE+%), Qualidade Afetiva (QA+%), Qualidade Formal (QF+%), Índice de ansiedade e Índice de Impulsividade. Em termos de resultados, a confiabilidade média entre os avaliadores para as categorias CE+%, QA+% e QF+% foi considerada excelente (K_{min}= 0,79; K_{max} = 0,82; e K_{med} = 0,80). Tanto para o Índice de Ansiedade (r = 0,80, p < 0,0001) quanto para o Índice de Impulsividade (r = 0,88, p < 0,0001) os coeficientes de confiabilidade foram considerados igualmente adequados. Concluíram que houve diferença significativa entre os pacientes com transtornos alimentares e os integrantes do grupo controle no que diz respeito à depressão, à impulsividade, ao teste de realidade e à agressividade, tendo como base a fidedignidade das avaliações realizadas.

Crisi (2011) avaliou a confiabilidade entre três juízes independentes, certificados no CWS, de uma amostra clínica de 30 protocolos selecionados aleatoriamente. Os coeficientes de correlação intraclasse (CCI) foram calculados para todos os índices

formais, exceto as categorias específicas de Conteúdo, Movimento e Códigos Especiais, cujas frequências eram muito baixas para comparações significativas. Os resultados apontaram, para a maioria dos índices avaliados, níveis excelentes de CCI, variando de 0,77 a 0,97 ($p < 0,01$) (EC+% = 0,77; QA+% = 0,90) e várias categorias de pontuação demonstraram concordância substancial entre os juízes, incluindo QF+% (QF+% = 0,60). Análises de confiabilidade entre juízes foram realizadas especificamente relacionadas as categorias Caráter Evocativo (CE+%) e Qualidade Afetiva (QA+%) para garantir a consistência das pontuações. Tanto o Kappa de Cohen *(k)*, quanto o Kappa de Fleiss*(K)* foram calculados, sendo que o primeiro examinou cada correlação possível entre pares de três avaliadores independentes (cegos) e o último examinou as correlações entre três avaliadores independentes simultaneamente. A análise preliminar dos dados brutos demonstrou níveis significativos de concordância total. Para o CE+% 76,6% dos campos demonstraram concordância total entre os avaliadores, 17,5% apresentaram concordância parcial e apenas 2,9% indicaram discordância total. Para a variável QA+%, 85,8% dos campos demonstraram concordância total, 14,2% indicaram concordância parcial e 0% demonstraram discordância total. Os resultados médios do kappa de Cohen *(k)* e do kappa de Fleiss *(K)* para o CE+% foram, respectivamente, $k = 0,64$ e $K = 0,64$. Já para a QA+% $k = 0,83$ e $K = 0,81$, ou seja, a confiabilidade geral entre avaliadores situou-se na faixa de classificação substancial (0,61 a 0,80).

Daini, Petrongolo, Manzo, & Bernardini (2012) conduziram um estudo que investigou os fatores de personalidade afetados pelo trabalho em um grupo de enfermeiros profissionais. Dois juízes independentes realizaram uma avaliação às cegas de 321 protocolos (111 do grupo experimental e 210 do grupo controle). Os coeficientes de confiabilidade foram calculados para categorias relevantes incluindo CE+%, QA+% e QF+%. Em termos de resultados, a confiabilidade geral média entre os avaliadores encontrada situou-se na faixa excelente (0,95).

Crisi, Vari, Carlesiano, Guzzi, & Zavattini (2014) conduziram um estudo sobre depressão e estados afetivos negativos em pacientes dermatológicos. Dois juízes independentes avaliaram às cegas 84 protocolos (42 do grupo experimental e 42 do grupo controle). Os coeficientes de correlação intraclasse (CCI) foram calculados para as categorias Caráter Evocativo (CE+%), Qualidade Afetiva (QA+%) e Qualidade Formal (QF+%). O resultado médio do CCI ficou na faixa quase perfeita (CCI = 0,83).

Especificamente, CE+% = 0,89, QA+% = 0,79 e QF+% = 0,80. Concluíram que a confiabilidade entre avaliadores (variando de substancial a quase perfeita) foi demonstrada em três categorias cruciais do CWS, incluindo dois índices originais desse sistema, a saber, o CE+% e a QA+%.

Crisi, Palm, & Lops (2016) realizaram a avaliação da confiabilidade entre juízes em uma amostra americana. Para tanto, 30 protocolos randomicamente selecionados, foram avaliados por cinco juízes americanos treinados e certificados no CWS e um sexto juiz, considerado especialista na Itália, também fez parte do estudo. Foi utilizado o kappa de Fleiss *(K)*. Os valores apresentados para as porcentagens gerais de concordância e K_{free} (kappa marginal livre) foram respectivamente: para o CE+% 82,2, K_{free} = 0,73 (substancial); para a QA+% 88,7, K_{free} = 0,83 (quase perfeita); e para a QF+% 87,1, K_{free} = 0,81 (substancial). Os resultados forneceram forte evidência de consistência na pontuação entre profissionais treinados no CWS. Na sequência, o coeficiente de correlação intraclasse (CCI) foi calculado com o intuito de analisar as decisões individuais de pontuação feitas pelos seis juízes. O CCI das pontuações médias encontradas para as variáveis Caráter Evocativo (CE+% = 0,940), Qualidade Afetiva (QA+% = 0,940) e Qualidade Formal (QF+% = 0,763), demonstrou níveis de concordância bons a excelentes. Por fim, foram calculadas as correlações entre cada possível par de juízes exclusivamente para as categorias Caráter Evocativo (CE+%) e Qualidade Afetiva (QA+%) utilizando a correlação de ordem de classificação de Spearman *(ρ)*. Os resultados encontrados para o CE+% variaram de ρ = 0,60 a 0,92 e para a QA+% ρ = 0,74 a 0,89, ou seja, todas as correlações resultantes variaram de forte a muito forte. No geral, todos os componentes estudados apresentaram confiabilidade significativa entre os juízes o que assegurou a eficácia do treinamento, das diretrizes escritas e da metodologia de pontuação do CWS.

Crisi e Dentale (2016) investigaram a confiabilidade entre avaliadores das categorias introduzidas pelo CWS, Caráter Evocativo (CE+%), Qualidade Afetiva (QA+%) e Qualidade Formal (QF+%). O estudo consistiu na revisão cega de 30 protocolos clínicos selecionados randomicamente por dois juízes independentes. Os coeficientes de correlação intraclasse foram calculados, com cada variável demonstrando níveis significativos de concordância: CE+% (CCI = 0,74, $p < 0,001$), QA+% (CCI = 0,92, $p < 0,001$) e QF+% (CCI = 0,71, $p < 0,001$).

Como é possível observar, a partir dos resultados das pesquisas apresentadas, as definições elaboradas para a avaliação de um protocolo do WZT, seguindo-se as orientações do CWS no contexto internacional, mostraram-se bem delineadas e permitiram que diferentes avaliadores realizassem avaliações com boa segurança, especialmente quando se consideram as variáveis Qualidade Afetiva (QA+%), Qualidade Formal (QF+%) e, Caráter Evocativo (CE+%). Para um melhor entendimento dessas três variáveis, julgou-se pertinente apresentá-las a seguir, mesmo que de modo suscinto.

A variável QA+% avalia a disposição emocional pessoal, ou seja, a habilidade de entrar em contato com as emoções, o tipo de afeto que caracteriza a vida emocional e em que medida o indivíduo é capaz de se conectar com o ambiente. Em outras palavras, avalia a integração e regulação afetiva (Crisi, 2018; Crisi & Dentale, 2016). Ela é avaliada de acordo com a conotação positiva/agradável, neutra ou negativa/desagradável do conteúdo desenhado em cada campo, bem como a partir das verbalizações feitas pelo indivíduo acerca dos seus desenhos. Aos desenhos de cada campo pode ser atribuída uma pontuação de 1,0, 0,5 ou 0,0, dependendo da conotação positiva/agradável, neutra, ou negativa/desagradável do conteúdo desenhado (Koller, 2023). Como regra geral, uma QA positiva (que recebe a pontuação1,0) refere-se aos desenhos de conteúdos humanos (H), animais (A), de natureza (NAT), botânicos (BOT) e comida (FD). A QA neutra (pontuação 0,5) refere-se aos desenhos de objetos (OBJ), sinais e números (SIG), minerais (MIN) e arquitetura ou construções (ARC). Por fim, QA negativa (que recebe a pontuação 0,0) refere-se aos desenhos de anatomia (ANA), armas (OBJ, WP), explosões (EXP), condições meteorológicas adversas, incluindo nuvens e chuva (CLD) e conteúdos patológicos (PAT) (Crisi, 2018). Para a correta pontuação dessa variável o avaliador é orientado, primeiramente, a observar o tipo de conteúdo e, em seguida, verificar a possibilidade de modificações tanto no sentido positivo quanto no negativo advindas das descrições verbais do cliente a respeito do desenho feito no campo em avaliação. Isto significa que os mesmos conteúdos que poderiam receber uma pontuação 1,0 (QA positiva) ou 0,5 (QA neutra), se explicitamente descritos de maneira negativa, depreciativa ou pejorativa passam a receber a pontuação 0,0 (QA negativa). Do mesmo modo, conteúdos indicados como negativos ou neutros podem ter uma conotação positiva, se descritos com termos carinhosos, assim haveria a mesma modificação na pontuação, só que no sentido inverso (Crisi, 2018).

A variável QF+% estima a qualidade dos processos cognitivos, ou seja, avalia a adequação do teste de realidade, a capacidade de se relacionar com o ambiente, o bom senso, os mecanismos de defesa e o autocontrole consciente (Crisi, 2018; Crisi & Dentale, 2016). Em sua avaliação, é orientado que não se deve medir a capacidade de desenhar bem, mas sim a capacidade que o respondente teve ao tornar claro e óbvio o desenho aos olhos do avaliador. Esta habilidade está estritamente ligada ao nível de funcionamento cognitivo, pois é baseada na capacidade de reconhecer e representar o todo de uma parte (Crisi, 2018). Assim, o avaliador deve atribuir 1,0 (um ponto), quando o desenho é evidente e seu significado é imediatamente percebido, sem a necessidade de recorrer aos esclarecimentos dados pelo examinando, seja por meio dos títulos ou ao que respondeu na pergunta "Eu desenhei". A pontuação de 0,5 (meio ponto) será dada, quando o significado do desenho não for imediatamente percebido, induzindo muitas interpretações diferentes que precisam ser esclarecidas. Nessa condição, o avaliador tem a necessidade de recorrer ao título, bem como às explicações dadas pelo sujeito sobre o desenhado (pergunta "eu desenhei"). Já a pontuação 0,0 ponto (zero) é atribuída quando o desenho se apresenta incompreensível, impreciso ou arbitrário e nem mesmo o inquérito ou a explicação dada pelo sujeito permite uma interpretação apropriada (Crisi, 2018; Crisi & Dentale, 2016).

A variável CE+% avalia a integridade das funções perceptivas e associativas, a adaptação ao pensamento convencional e à realidade, ou seja, a capacidade de se relacionar com o ambiente (Crisi, 2018; Crisi & Dentale, 2016), podendo-se tomá-la como uma medida de ajustamento social. Essa variável é indicativa da capacidade do sujeito de sintonizar os estímulos vindos do mundo exterior e se adaptar a eles adequadamente (Di Renzo et al, 2021). Em termos avaliativos, busca-se avaliar a ressonância e a afinidade do cliente com cada sinal do teste, incluindo sua capacidade de atender efetivamente às demandas organizacionais gestálticas inerentes a cada um dos oito sinais gráficos. Assim como as variáveis QA+% e QF+%, pontuação dessa variável também pode receber uma pontuação que varia entre 1,0, 0,5 e 0,0 (zero)pontos (Crisi, 2018).

Na avaliação da variável CE+% encontra-se presente uma regra geral. Essa, indica que um campo receberá pontuação 1,0 se o desenho e/ou verbalização do cliente se encaixa com a sugestão do estímulo. Uma pontuação de 0,5 será dada quando a resposta

do cliente se alinha apenas parcialmente com a sugestão do sinal. E, uma pontuação 0,0 (zero) será dada quando o cliente ignora completamente ou não parece considerar o caráter evocativo particular de um campo. Além desses aspectos, para os campos complexos (3, 5, 6 e 7, em que o estímulo consiste em dois ou mais sinais) há uma regra adicional e necessária para receber uma pontuação de 1 ou 0,5. Nesse sentido, o cliente deve conceber os dois ou mais sinais do campo como parte do estímulo e então será avaliado a partir da regra geral. Se a condição necessária descrita não for atendida, a pontuação para estes campos específicos será 0,0 (zero) (Crisi, 2018). Avalia-se positivamente se o sujeito tem a capacidade de apreender o Caráter Evocativo do sinal de estímulo presente em cada campo, ou seja, se o desenho capta a sugestão implícita do sinal de estímulo, a saber: Campo 1 = centralidade e relevância; Campo 2 = vitalidade e movimento; Campo 3 = direcionalidade e progressão dinâmica; Campo 4 = peso e estabilidade; Campo 5 = superação dinâmica de um obstáculo; Campo 6 = síntese e estruturação; Campo 7 = delicadeza; Campo 8 = arredondamento e fechamento (Di Renzo et. al, 2021);

Objetivo

Diante do exposto e, considerando que na atualidade, o cenário nacional não possui um sistema de avaliação do WZT preciso e válido, bem como o CWS constituiu-se no único sistema de avaliação que apresenta parâmetros psicométricos consistentes no contexto internacional, tornou-se objetivo desse trabalho realizar um estudo de concordância entre avaliadores para o CWS, com foco nas três variáveis já descritas desse sistema, a saber: Qualidade Afetiva (QA+%), Qualidade Formal (QF+%) e Caráter Evocativo (CE+%). A escolha dessas três variáveis baseou-se no fato delas serem cruciais/centrais para o sistema foco desse trabalho, ou seja, sua mensuração é fundamental para que se possa realizar uma análise mais detalhada da personalidade do indivíduo que se submete ao instrumento, seguindo-se a metodologia do CWS. Hipotetizou-se que a concordância a ser aferida deveria apresentar níveis compatíveis com os estudos de concordância apresentados para o sistema de pontuação no âmbito internacional, bem como deveria ser superior a $r = 0,60$, conforme indicado na Resolução 031/2022 do CFP.

Métodos

Participantes

A amostra foi composta por dois avaliadores, denominados de Juiz 1 e Juiz 2. Estes, eram de ambos os sexos, psicólogos, mestres em Avaliação Psicológica e professores com experiência em técnicas projetivas ou de autoexpressão.

Materiais

Foram utilizados 90 protocolos do WZT aplicados segundo as orientações do CWS. Esse material era proveniente da pesquisa empreendida por Koller (2023), que buscou investigar evidências de validade com medidas externas do WZT na avaliação da depressão, sob a ótica do CWS. A citada pesquisa foi realizada mediante a aprovação do Comitê de Ética em Pesquisa com Seres Humanos do Instituto de Psicologia da Universidade São Paulo - USP, sob o número CAAE 30592820.3.0000.5561. Além dos 90 protocolos em pdf, cada avaliador recebeu um arquivo em word com os requisitos/critérios para avaliar cada uma das três variáveis em estudo e uma planilha em Excel para lançar suas avaliações.

Procedimentos

Os dois juízes foram inicialmente convidados a participar do estudo e, mediante o aceite, assinaram o Termo de Consentimento Livre e Esclarecido. Após essa etapa, foram treinados separadamente nos critérios de correção, objetivando-se o esclarecimento de dúvidas. Em seguida, cada um recebeu por meio digital uma cópia (pdf) dos 90 protocolos, os critérios de codificação das três variáveis do CWS (vide abaixo) e uma planilha eletrônica em que deveriam digitar suas avaliações feitas. Definiu-se o prazo de 30 dias úteis para que enviassem o material aos pesquisadores responsáveis.

Após o recebimento do material enviado pelos juízes, realizou-se o cálculo das pontuações efetuadas em termos percentuais. Esse cálculo foi necessário em razão das pontuações para as três variáveis variarem entre 0,0, 1,0 ou 1,5 para cada um dos oito campos do WZT. Para tanto, utilizou-se a fórmula geral, ou seja, para cada uma das três variáveis em estudo:

$$\text{Variável+\%} = \frac{\Sigma(\text{variável})}{(n)} \times 100$$

- Onde:

Variável + % = QA (Qualidade Afetiva), QF (Qualidade Formal) ou CE (Caráter Evocativo);

Σ (variável) = soma das pontuações atribuídas (0,0, 0,5 ou 1,0) pelos avaliadores para cada variável nos oito campos do WZT; n = Número de campos do WZT, no caso 8;

A partir da utilização dessa fórmula, foi possível chegar a um percentual único para cada uma das variáveis. Esse percentual final foi utilizado para realizar os cálculos relativos ao *r* de Pearson, ao coeficiente Kappa e ao CCI.

Plano de análise de dados

Neste trabalho, optou-se por utilizar três métodos estatísticos — o Coeficiente de Correlação de Pearson, o Kappa de Cohen e o Coeficiente de Correlação Intraclasse (ICC), para avaliar a concordância das avaliações feitas pelos dois juízes para as três variáveis em foco. A utilização conjunta dessas três estatísticas teve por objetivo proporcionar uma análise abrangente da concordância entre juízes, permitindo uma avaliação mais robusta, com vistas a garantir que as avaliações realizadas fossem confiáveis e representativas. Além desse aspecto, o cálculo da concordância pelos três métodos, foi adotado pelo fato de permitir uma comparação efetiva/direta com os estudos de concordância apresentados pois, em sua grande maioria, fizeram o uso das respectivas estatísticas. Para o cálculo, foi utilizado o software SPSS, versão 20 para Windows.

Resultados

Inicialmente procedeu-se ao cálculo do *r* de Pearson. O objetivo de calculá-lo foi entender a força e a direção da relação entre as classificações efetuadas pelos dois juízes, proporcionando uma visão inicial da consistência entre eles. A tabela 1 mostra as correlações apuradas para as três variáveis em estudo.

Tabela 1.

Correlações de Pearson (r) entre as classificações realizadas pelos juízes para as três variáveis do CWS.

		Juiz 2	
	Variáveis	*R*	I.C. 95%
Juiz 1	QA+%	0,98*	0,967 – 0,995
	QF+%	0,99*	0,971 – 0,998
	CE+%	0,98*	0,954 – 0,993

* As correlações foram significativas no nível de 0,01 (bicaudal).

Os resultados apresentados na Tabela 1 indicam uma forte relação positiva (Akoglu, 2018) entre as avaliações realizadas pelos dois juízes para as três variáveis do CWS. Essas, apresentaram intervalos de confiança de 95% relativamente estreitos, variando de 0,954 a 0,998. Assim, refletem uma alta concordância nas classificações atribuídas, indicando que as variações observadas entre suas avaliações são mínimas. Além disso, essas correlações são estatisticamente significativas no nível de 0,01 (bicaudal), reforçando que as relações entre as variáveis avaliadas pelos juízes não ocorreram ao acaso.

Esses valores sugerem, também, que as direções das avaliações realizadas foram extremamente consistentes. Em outros termos, quando um juiz classificou uma variável com uma pontuação mais alta ou mais baixa, o outro juiz tendeu a fazer o mesmo, o que reflete uma harmonia significativa na aplicação dos critérios de avaliação utilizados. Em síntese, as altas correlações apuradas indicam uma forte consistência nas avaliações dos dois juízes, apontando possivelmente para a claridade e precisão dos critérios que eles seguiram, bem como uma interpretação compartilhada dos parâmetros das variáveis. Isso contribui para a confiança nos resultados e na precisão das avaliações realizadas.

Considerando a análise do *r* de Pearson como um passo preliminar que permitiu compreender a força e a direção da relação entre os dados, buscou-se aprofundar a análise e entender melhor a concordância entre os juízes apresentada. Para tanto, fez-se o uso da estatística Kappa. O coeficiente Kappa de Cohen (κ) oferece uma medida de concordância que considera a possibilidade de acordos ao acaso, permitindo assim uma avaliação mais precisa da consistência entre as classificações realizadas, além da força e direção

proporcionada pelo *r* de Pearson. Os resultados para as três variáveis são apresentados na tabela 2.

Tabela 2.

Coeficiente Kappa de Cohen (κ) entre os dois juízes para cada uma das três variáveis do CWS

		Juiz 2	
Variáveis	QA+%	QF+%	CE+%
Juiz 1	κ=90%	κ= 85,6%	κ= 90%

Pelos resultados da tabela 2, verificou-se um alto nível de concordância para as três variáveis do CWS (κ = ≥ 0,85, p < 0,001), indicando uma consistência forte e significativa nas avaliações realizadas pelos dois juízes. Assim, é sugestivo que os critérios de avaliação seguidos pelos juízes foram bem delineados e compreensíveis, permitindo que ambos os avaliadores aplicassem as mesmas regras e interpretassem os aspectos avaliados de maneira bastante semelhante. Cabe salientar que um valor de Kappa acima de 0,80 é comumente interpretado como um nível de concordância substancial a quase perfeita Dettori & Norvell, 2020). Nesse sentido, sugere a presença de um baixo grau de discordância entre os juízes, mesmo ao se considerar a possibilidade de concordância ao acaso.

Estabelecida a concordância via a estatística Kappa, procedeu-se a análise da fidedignidade por intermédio do Coeficiente de Correlação Interclasse (CCI) (tipo concordância absoluta; modelo aleatório de duas vias; Koo& Li, 2016), com o objetivo de avaliar a concordância entre os dois juízes frente à classificação que realizaram dos 90 protocolos. A tabela 3 apresenta esses resultados.

Tabela 3.

Coeficientes de Correlação Interclasse (CCI) para as três variáveis do CWS

Variáveis	CCI	I.C. 95%		Teste F com valor verdadeiro 0,0			
		Mín.	Máx.	Valor	df1	df2	sig
QA+%	0,993*	0,989	0,995	133,787	89	89	0,000
QF+%	0,994*	0,992	0,996	187,205	89	89	0,000
CE+%	0,989*	0,983	0,993	90,289	89	89	0,000

Nota: * Essa estimativa é calculada presumindo-se que o efeito de interação esteja ausente, pois não é estimável de outra forma.

Os CCI apresentados na tabela 3 indicam uma excelente concordância entre os juízes na avaliação das três variáveis do CWS (QA+%, QF+%, CE+%). Segundo os critérios de interpretação do CCI sugeridos por Koo e Li (2016), valores de CCI superiores a 0,9 são indicativos de uma confiabilidade excepcional, o que reforça a consistência entre os avaliadores. Os resultados demonstram valores de CCI com intervalos de confiança bastante estreitos (I.C. 95% variando de 0,983 a 0,996), que apontam para o fato de que os juízes aplicaram os critérios de avaliação de maneira altamente consistente e que a variação observada entre eles é praticamente desprezível.

Além do exposto, os testes *F* realizados com valor verdadeiro de 0,0 mostram significância estatística ($p < 0,001$), o que confirma que a concordância entre os juízes não foi fruto do acaso. As variáveis foram avaliadas com precisão e consistência, e os resultados evidenciam que os critérios de pontuação foram bem estabelecidos e seguidos de forma rigorosa, minimizando discrepâncias e aumentando a confiabilidade das avaliações. Em suma, os altos valores de CCI demonstram que os escores obtidos pelos juízes foram praticamente idênticos, o que corrobora a eficiência dos critérios de avaliação adotados e a robustez da metodologia utilizada.

Discussão

Ao se considerar os resultados apurados em termos das correlações de Pearson (*r*) no presente estudo com os valores apurados por Daini et al. (2012), observou-se que ambos apontaram para uma excelente consistência e confiabilidade entre os juízes na avaliação das variáveis, especialmente nas categorias CE+%, QA+% e QF+%. Embora o trabalho de Daini et al. (2012) não detalhe as correlações específicas para cada variável

(CE+%, QA+%, QF+%) como nesse estudo, a confiabilidade média geral excelente de 0,95 sugere uma consistência muito próxima à observada nas correlações de Pearson (0,98–0,99). Essa semelhança reforça a ideia de que os juízes nos dois estudos interpretaram e aplicaram os critérios de avaliação de forma similar, mesmo em amostras e contextos diferentes. Portanto, ao correlacionar os resultados, ambos os estudos demonstraram um nível elevado de confiabilidade entre juízes, apontando que as variações nas classificações entre eles são mínimas, e que os critérios de avaliação foram aplicados de forma clara e precisa. A elevada concordância entre os juízes, indicada pelos altos coeficientes do *r* Pearson e pela confiabilidade geral média excelente, sugere que as diretrizes foram bem seguidas, proporcionando resultados consistentes em ambos os casos,

Ao correlacionar os coeficientes de Pearson (*r*) desse estudo com os coeficientes de Spearman (ρ) de Crisi, Palm &Lops (2016), observa-se que ambos fornecem fortes evidências de consistência nas avaliações entre os juízes, apesar de utilizarem abordagens estatísticas diferentes. As correlações de Pearson para as três variáveis (QA+%, QF+%, CE+%) variam entre 0,98 e 0,99, refletindo uma relação quase perfeita entre os juízes. Essas correlações são lineares, indicando que, à medida que um juiz atribuiu uma pontuação alta ou baixa, o outro juiz fez o mesmo, com mínima variação entre eles. No estudo de Crisi, Palm &Lops (2016), o coeficiente de Spearman (ρ), que mede a associação de ordem de classificação entre pares de juízes, apresentou resultados que variaram de $\rho = 0{,}60$ a 0,92 para o CE+% e de $\rho = 0{,}74$ a 0,89 para o QA+%. Esses valores indicaram uma forte relação monotônica, ou seja, as classificações dos juízes seguem um padrão consistente, mas sem exigir uma relação estritamente linear como a observada no *r* de Pearson.

Ao comparar diretamente os coeficientes de correlação dos dois estudos, percebeu-se que ambos indicaram um elevado grau de concordância entre os juízes, com os resultados de Crisi, Palm & Lops (2016) apresentando variabilidade maior. Essa diferença pode ser explicada pela natureza das correlações, já que o ρ Spearman não se baseia na suposição de linearidade, e as variações nas classificações podem ser refletidas de forma diferente. No entanto, ambas as abordagens apontam para consistência significativa entre os juízes. Portanto, os resultados dos dois estudos corroboram a confiabilidade dos critérios de avaliação, seja em termos de relações lineares (como

demonstrado pelas correlações de Pearson) ou de relações monotônicas (como indicado pelas correlações de Spearman). Ambos sugerem que os juízes aplicaram os critérios de forma consistente e seguindo diretrizes claras, o que reforça a precisão do sistema de avaliação utilizado no CWS.

Na comparação com os resultados apurados por Crisi (2011), que analisou a concordância de três juízes independentes utilizando tanto o Kappa de Cohen quanto o Kappa de Fleiss, cujos resultados apontaram para CE+%, o Kappa médio de 0,64, e para QA+%, o Kappa de 0,83, com uma concordância substancial, observou-se que os valores para essas estatísticas foram inferiores aos encontrados no presente estudo. Tal perspectiva sugere que os juízes deste estudo talvez tenham compreendido as instruções mais claramente ou que a categorização dos dados tenha sido mais objetiva, resultando em uma maior concordância. No entanto, a alta concordância observada para QA+% no estudo de Crisi (2018) (0,83) é bastante próxima dos 90% encontrados aqui, indicando que a categoria QA+% é frequentemente bem pontuada entre diferentes juízes.

Crisi, Palm, & Lops (2016), ao avaliarem juízes americanos e italianos, encontraram coeficientes Kappa de 0,73 a 0,83 para as variáveis CE+%, QA+%, e QF+%. Embora esses valores reflitam uma concordância substancial, os coeficientes do presente estudo (0,85 a 0,90) indicaram um grau ainda mais elevado de concordância. A diferença pode ser atribuída a variáveis contextuais, como diferenças culturais entre os juízes ou na formação que receberam, uma vez que os juízes deste estudo parecem ter uma compreensão altamente uniforme dos critérios de avaliação.

Os coeficientes Kappa observados refletem uma excelente concordância entre os juízes, similar aos achados dos estudos comparados, especialmente aqueles que envolvem especialistas no uso do CWS. A concordância quase perfeita obtida aqui, particularmente nas variáveis QA+% e CE+%, encontra-se em linha com o esperado quando juízes experientes estão envolvidos na avaliação de protocolos do Wartegg pelo CWS. Cabe salientar que a importância do Kappa reside em sua capacidade de compensar a concordância ao acaso, o que torna seus valores mais informativos do que o *r* de Pearson isolado. Os altos valores de Kappa observados indicaram que a consistência entre os juízes vai além do que seria esperado por mera sorte destacando, mais uma vez, a clareza das instruções e a qualidade dos critérios de avaliação.

A análise da fidedignidade utilizando o Coeficiente de Correlação Interclasse (CCI) revelou uma concordância excepcional entre os dois juízes nas três variáveis do CWS (QA+%, QF+%, CE+%) para o presente estudo. Esses achados corroboram a literatura, que aponta para altos níveis de CCI em estudos semelhantes. Por exemplo, no estudo de Crisi (2011), os CCI variaram de 0,77 a 0,97, com QA+% apresentando um valor de 0,90 e CE+% de 0,77, sugerindo uma confiabilidade substancial a quase perfeita entre os juízes. No presente estudo, os valores ainda mais elevados, especialmente para QA+% e CE+%, indicaram uma avaliação ainda mais consistente entre os juízes. Outro estudo relevante é o de Crisi, Vari, Carlesiano, Guzzi e Zavattini (2014), que também encontrou CCIs elevados ao avaliar 84 protocolos, com valores de 0,79 para QA+%, 0,89 para CE+% e 0,80 para QF+%. Esses resultados, classificados entre substanciais e quase perfeitos, são comparáveis aos encontrados no presente estudo, em que as variáveis QA+% e CE+% também apresentaram níveis excepcionais de concordância.

Crisi, Palm e Lops (2016) conduziram uma avaliação com seis juízes, em uma amostra americana, encontrando CCIs de 0,94 para QA+% e 0,94 para CE+%, enquanto QF+% apresentou um valor um pouco menor (0,763). Embora esses valores já representem uma excelente concordância, os resultados do presente estudo indicam uma consistência ainda maior, especialmente para QF+%, que apresentou um CCI de 0,99, sugerindo que os juízes seguiram criteriosamente as diretrizes de pontuação. Por fim, Crisi e Dentale (2016) encontraram níveis semelhantes de confiabilidade em 30 protocolos clínicos, com QA+% apresentando um CCI de 0,92, CE+% de 0,74 e QF+% de 0,71, confirmando uma boa concordância entre os avaliadores. No entanto, os resultados do presente estudo destacam uma maior precisão nas avaliações, o que pode ser atribuído à clareza dos critérios de pontuação e ao treinamento dos juízes.

Em resumo, os resultados obtidos com o CCI, com valores altos e intervalos de confiança estreitos, demonstram a robustez dos critérios de avaliação adotados, além de confirmar a fidedignidade das avaliações feitas pelos juízes. Esses achados reforçam que o sistema de avaliação para astrês variáveis do CWS em foco, permite obter resultados consistentes e replicáveis. Ao se considerar os valores apurados em relação ao valor mínimo preconizado pelo CFP para estudos de precisão, ou seja 0,60, observou-se que esse critério foi plenamente atendido. Tal condição reforça a ideia de que, se o presente estudo de fidedignidade fosse apresentado à Comissão Consultiva de Avaliação

Psicológica (CCAP), o CWS teria, a princípio, seu uso aprovado no que concerne a esse aspecto.

Conclusão

Este estudo teve como objetivo avaliar a concordância entre avaliadores do sistema de avaliação CWS, focando nas variáveis Qualidade Afetiva (QA+%), Qualidade Formal (QF+%) e Caráter Evocativo (CE+%). Os resultados demonstraram uma elevada consistência e confiabilidade nas avaliações, com coeficientes de correlação de Pearson (r), indicando uma concordância quase perfeita entre os juízes. Além disso, os valores do coeficiente Kappa refletiram também uma excelente concordância, superando os achados de estudos anteriores. A análise do Coeficiente de Correlação Interclasse (CCI) corroborou esses resultados, mostrando uma robustez dos critérios de avaliação, especialmente para QA+% e CE+%. Desse modo, confirmou-se a hipótese levantada de que a concordância aferida apresentaria níveis compatíveis com os estudos de concordância no âmbito internacional, bem como deveria ser superior ao valor mínimo exigido pelo CFP para estudos de fidedignidade para o contexto brasileiro.

Entretanto, o estudo apresenta algumas limitações. Dentre essas, salienta-se à amostra de protocolos pequena, que não abrangeu o território brasileiro em sua totalidade, limitando a generalização dos resultados. Outra limitação a ser destacada, se refere a ausência de uma avaliação às cegas. Nesse sentido, embora o treinamento prévio tenha contribuído para os altos índices de concordância observados, uma avaliação às cegas, sem esse tipo de treinamento prévio, seria oportuna para aproximar a avaliação da realidade, refletindo a performance de avaliadores sem treino ou experiência na análise das variáveis do CWS. Considerando-se ainda o treinamento prévio, seria oportuno realizar também um estudo comparativo de concordância entre juízes, em que a comparação do desempenho entre treinados (especialistas) e não treinados seja o foco. Essa estratégia poderia dimensionar o efeito que o treinamento na avaliação das variáveis consideradas exerce.

Por fim, é sugestivo também a ampliação dos estudos de concordância entre avaliadores para outras variáveis do CWS não contempladas no presente estudo. O intuito seria o de verificar se as mesmas tendências de concordância observadas para as variáveis

em foco se aplicam, ou se essas tendências se mostrariam diferentes, apontando para a clareza ou não de todos os critérios/variáveis do CWS.

Referências

Akoglu H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Alsaqr, A. M. (2021). Remarkson the use of Pearson's and Spearman's correlation coefficients in assessing relationships in ophthalmic data. *African Vision and Eye Health*, *80*(1), 612-622. https://doi.org/10.4102/aveh.v80i1.612

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. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1413-03942010000100006&lng=pt&tlng=pt.

Bakeman, R. (2023). Kappa Acc: A program for assessing the adequacy of kappa. *Behavior Research Methods*, (55), 633–638. https://doi.org/10.3758/s13428-022-01836-1

Berlinck, V. R. (2000). *O Teste de Completamento de Desenhos de Wartegg em universitários de São Paulo.* (Dissertação de Mestrado, Universidade de São Paulo, São Paulo, Brasil).Recuperado em 2024-10-02, de www.teses.usp.br

Berlinck, V. R. (2006). *O Teste de Completamento de Desenhos de Wartegg em profissionais adultos com nível de escolaridade fundamental e médio.* (Tese de Doutorado, Universidade de São Paulo, São Paulo, Brasil).Recuperado em 2024-10-02, de www.teses.usp.br

Conselho Federal de Psicologia - CFP (2022). *Define e regulamenta o uso, a elaboração e a comercialização de testes psicológicos e revoga a Resolução CFP nº 09/2018*. Resolução nº 31/2022.https://atosoficiais.com.br/cfp/

Crisi, A. (2011). Reliability and validity of the Crisi Wartegg System. In A. Crisi, *The Crisi Wartegg System (CWS). Manual for administration, scoring and interpretation.* (pp. 48-49). New York, NY: Routledge.

Crisi, A. (2018). *The Crisi Wartegg System (CWS). Manual for administration, scoring and interpretation.* New York, NY: Routledge.

Crisi, A., Vari, C., Velotti, P., Carlesimo, S., Guzzi, C., & Zavattini, G. C. (2014). Depression and negative affect in dermatology patients: the use of the WDCT according to the Crisi Wartegg System. In A. Crisi, *The Crisi Wartegg System (CWS). Manual for administration, scoring and interpretation.* (pp. 120-123). New York, NY: Routledge.

Crisi, A., Palm, J., & Lops, I. (2016). Interrater reliability of the Crisi Wartegg System: an initial United States sample. *N* A. Crisi, *The Crisi Wartegg System (CWS). Manual for administration, scoring and interpretation.* (pp. 51-55). New York, NY: Routledge.

Crisi, A., & Dentale, F. (2016). The Wartegg Drawing Completion Test: inter-rater agreement and criterion validity of three new scoring categories. *International Journal of Psychology and Psychological Therapy*, *16*(1), 83-90. www.ijpsy.com/volumen16/num1/435.html.

Daini, S., Lai, C., Festa, G. M., Maiorino, F., Pertosa, M., & De Risio, S. (2006). Impulsivity in Eating Disorders: Analysis Through Wartegg Test. *Journal of Projective Psychology & Mental Health, 13*(2), 107-119. https://www.proquest.com/scholarly-journals/impulsivity-eating-disorders-analysed-through/docview/222301318/se-2.

Daini, S., Petrongolo, L., Manzo, A., & Bernardini, L. (2012). Personality and nursing work: a comparison between professional and student nurses. *Journal of Projective Psichology & Mental Health, (1)*19, 18-31.SIS JournalofProjectivePsychology& Mental Health, 2012, Vol 19, Issue 1, p18

de Raadt, A., Warrens, M.J., Bosker, R.J., &Kiers, H. A. L. (2021). A Comparison of Reliability Coefficients for Ordinal Rating Scales. *Journal of Classification, 38*,519–543. https://doi.org/10.1007/s00357-021-09386-5

Dettori, J. R., &Norvell, D. C. (2020). Kappa andBeyond: Is ThereAgreement?. *Global spinejournal*, *10*(4), 499–501. https://doi.org/10.1177/2192568220911648

Di Renzo, M., Bianch di Castelbianco, F., Crisi, A., Zaza, F. Marini, C., Racinaro, L., & Rea, M. (2021). A psyhological Reading of autismo spectrum disorders through the Wartegg – CWS. *Journal of Psychological and Educational Research JPER, 29* (2), 90-110. e-ISSN: 2821-4099.

Han, X. (2020). On Statistical Measures for Data Quality Evaluation. *Journal of Geographic Information System, (12)*3, 178-187. doi: 10.4236/jgis.2020.123011.

Koller, L. (2023). *Teste de Completamento de Desenhos de Wartegg: um estudo de validade de critério para o Sistema Crisi de Interpretação*. (Dissertação de Mestrado, Instituto de Psicologia, Universidade de São Paulo, São Paulo, Brasil). doi:10.11606/D.47.2023.tde-08122023-170414.Recuperado em 2024-10-02, de www.teses.usp.br

Konstantinidis, M., Le, Lisa. W., & Gao, X. (2022). Na Empirical Comparative Assessment of Inter-Rater Agreement of Binary Outcomes and Multiple Raters. *Symmetry*, *14*(2), 262. MDPI AG. http://dx.doi.org/10.3390/sym14020262

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Ogbodo-Adoga, R. (2020). Psychological Tests in Counselling. *Prestige Journal Of Counselling Psychology, 3*(1), 247-258. ISSN: 2651-5709 (Online)

Pereira, D. F. (2006). *Um estudo sobre o Wartegg como medida de criatividade em seleção de pessoal*. (Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil). UFRGS Lume Repositório Digital. https://lume.ufrgs.br/handle/10183/8712.

Pérez, J.A., Martin, P. S. P. Martin (2023). Coeficiente de correlación intraclasse. *Medicina de Familia, 49*(3), 1-4. https://doi.org/10.1016/j.semerg.2022.101907

Pessotto, F. (2015). *Estudos de sistemas de codificação do Teste de Wartegg e suas relações com o Rorschach (R-PAS)*. (Tese de Doutorado, Universidade São Francisco, Itatiba, São Paulo, Brasil). https://www.usf.edu.br/galeria/getImage/427/6722144706571676.pdf.

Pessotto, F., Primi, R. (2017). Precisão entre juízes para um novo sistema de codificação do Teste de Wartegg. *Revista de Psicologia*, *8*(2), 19-26.

Pessotto, F., & Primi, R. (2018). Evidências de validade de critério para o Teste de Wartegg. *Avaliação Psicológica, 17*(3), 279-291. http://dx.doi.org/10.15689/ap.2018.1703.13941.01

Ramon, R. R. (2006). *Wartegg: Precisão entre avaliadores e evidência de validade com o Método de Rorschach*. (Dissertação de Mestrado, Universidade São Francisco,

Itatiba, São Paulo, Brasil). http://pepsic.bvsalud.org/scielo.php?script=sci_nlinks&ref=080809&pid=S1413-03942010000010000600020&lng=pt.

Schneider, A. M. A., Bandeira, D. R., & Meyer, G. J. (2020). Rorschach Performance Assessment System (R-PAS) Interrater Reliability in a Brazilian Adolescent Sample and Comparisons With Three Other Studies. *Assessment, 29*(5), 859-871.https://doi.org/10.1177/1073191120973075

Souza, C. V. R., Primi, R., & Miguel, F. K. (2007). Validade do Teste Wartegg: correlação com 16PF, BPR-5 e desempenho profissional. *Avaliação Psicológica*, *6*(1), 39-49. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712007000100006&lng=pt&tlng=pt.

Tarigan, M., & Fadillah, F. (2022). Inter-rater and Intra-Rater Reliability Test with Goodenough-Harris Drawing Test. *The Open Psychology Journal*, *15*(1), 1-5. doi: 10.2174/18743501-v15-e2207130

Toffoli, S. F. L., Andrade, D. F., Bornia, A. C., & Quevedo-Camargo, G. (2016). Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça. *Educação. Pesquisa, 42*(2), 343-358.doi: http://dx.doi.org/10.1590/S1517-9702201606135887

Villemor-Amaral, A. E., Cardoso, L. M. (2019). Avaliação da personalidade no Brasil utilizando métodos projetivos. Em Makilim Nunes Baptista... [et al.] (Orgs), *Compêndio de Avaliação Psicológica* (pp. 475-482). Petrópolis, RJ: Vozes.