International Journal of Psychology and Psychological Therapy, 2016, 16, 1, 83-90

The Wartegg Drawing Completion Test: Inter-rater Agreement and Criterion Validity of Three New Scoring Categories

Alessandro Crisi, Francesco Dentale*
Sapienza Universitá di Roma, Italia

Abstract

Three new scoring categories for the Wartegg Drawing Completion Test (WDCT) were introduced. Each of them permits to assess specific personality characteristics: the Evocative Character (EC) that is related to social adjustment; Form Quality (FQ) that is connected to reality testing ability and Affective Quality, (AQ) that is linked to general mood state. Inter-rater agreement and criterion validity of the new categories were investigated on a sample composed by healthy, anxious and psychotic individuals. For EC, FQ and AQ, results showed an adequate level of inter-rater agreement and a satisfactory capacity to discriminate among subjects, supporting their reliability and criterion validity. In particular, as expected, significant differences were found for each category among all groups, with higher mean scores for healthy subjects, medium for anxious and lower for psychotics (except for AQ category that showed significant differences only in the comparison between healthy and pathological subjects but not between anxious and psychotic ones). Possible approches to assess WDCT validity in future researches were discussed.
Key words: Wartegg Drawing Completion Test, evocative character, affective quality, form quality.

Novelty and Significance

What is already known about the topic?

- Wartegg Drawing Completion Test is a drawing projective technique.
- Traditional scoring procedures presented some difficulties that limited its use both in clinical and non-clinical domains.

What this paper adds?

- Introduces three new scoring categories for the Wartegg Drawing Completion Test.
- Inter-rater agreement and criterion validity of the three new scoring categories are investigated.

The Wartegg Drawing Completion Test (WDCT) was created by the German psychologist Ehrig Wartegg (1897-1983), a follower of the School of Gestalt psychology in Leipzig that was the main centre of the psychology of totality. The WDCT derives from the Sander Phantasie Test (Berger, 1939) and, as reported by Roivanen (2009), was published for the first time in 1926, even though a complete handbook saw the light only many years later (Wartegg, 1953). Since the publication of the first manual (Wartegg, 1953), several studies have been conducted using the WDCT in different contexts, such as clinical (e.g. Kinget, 1980; Pfeiffer, 1984), educational (e.g. Avé-Lallemant, 1994), and organizational settings (e.g. Souza, Primi, & Miguel, 2007).

The WDCT is a drawing projective technique whose graphic elements are semi-structured signs on which the individual is prone "to project contents and specific dynamics of his personality which are, then, revealing of his organization" (Rapaport,

* Correspondence: Sapienza Universitá di Roma, Via dei Marsi, 78, 00184 Roma, Italia. E-mail: francesco.dentale@libero.it

1977, p. 31). In accordance with Bornstein's (2007) view, we may define the WDCT as a performance-based personality test that "can be classified as a stimulus-attribution test in which the examinees give meaning through interpretation" (Bornstein, 2007; p. 203). The WDCT is based on a form divided in eight boxes, each containing a different graphical sign. During the administration procedure, the client sits opposite to the examiner. The examiner gives him the WDCT form and a pencil No. 2 without an eraser, and tells the following instruction: "As You can see, this form is subdivided in 8 Boxes; in each there is a little sign. Start by making a drawing in each box with a completed meaning, preferably the first which comes in your mind and avoiding, if it's possible, abstract drawings. It's not necessary for you to follow the numerical order; work at your own pace: there are not time restrictions." To any questions of the examinee, the psychologist will respond avoiding to influence the subject's response and reinforcing the belief that everything will be good because there are not right things or wrong things to do. One of the most important features of the WDCT is that its instructions are easier to understand than those of other commonly accepted projective tests. This simplicity decreases the probability that the respondent will misunderstand the instructions. As a consequence, WDCT is particularly suitable and helpful in childhood (it can be administered starting around 4 years and half) and, furthermore, it can be administered successfully to individuals with certain specific disabilities (deafness, cognitive delay, etc.). Moreover, the WDCT is easy, short, quick to administer (most individuals complete it in 5-10 minutes), quick to score (10-15 minutes) and to interpret (30 minutes).

In spite of WDCT usefulness and assessment potentials, the original scoring rules, proposed by Wartegg, presented two difficulties that limited its use both in clinical and non-clinical domains. Firstly, the scoring procedures conceived by Wartegg were very complex and arduous for the psychologists. Secondly, Wartegg referred to a theoretical model (the Psychology of the Totality of Krueger) that was not appropriate to fully support the interpretations of results. To overcome these difficulties, several scoring systems has been developed that are generally categorized into two different traditions (Gronnerod & Gronnerod, 2012). In the analytical systems tradition, the printed signs are considered as quite visual stimuli (e.g. Takala & Hakkarainen, 1953) while in the dynamic systems (e.g. Wass & Mattlar, 2000) they were conceived as bringing symbolic meanings that reflect individuals' personality. Moreover, the scoring systems for the WDCT protocols vary from approaches that focus on qualitative interpretation (e.g. Wartegg 1953; Avé-Lallemant 1978) to other systems that emphasize quantitative interpretation (e.g. Wass & Mattlar, 2000). The scoring categories generally used are: drawing time, the order of the squares drawn, possible refusals, the size of the drawings, the content of the drawings, crossing of the borders of the squares, shading, drawing line quality, and the written title of the drawing. However, in accordance with Gronnerod and Gronnerod (2012), although several systems shared many common elements, the categories and personality functions considered were not ever clearly defined and above all were not ever supported by empirical validation studies.

For these reasons, since 1983, a more functional and useful methodology, tested in several researches (Bianchi, Crisi, & Di Renzo, 1996; Crisi, 1998; 2007) and improved

many times in the last 30 years, was developed. More specifically, on the basis of a huge experience using Wartegg and Rorschach together (more than 3,000 clients examined), a scoring procedure that included the same categories of Rorschach according to Bohm's method (Bohm, 1969) was developed. Such categories are the contents, the frequency (popular and not popular responses), the special scores and the movement responses. Moreover, after several years of clinical application, three new categories were proposed: the Form Quality (FQ), the Evocative Character (EC) and the Affective Quality (AQ). An appropriate computation and a combined analysis of these indices permits to assess the personality organization via different areas of mental functioning: a qualitative and a quantitative evaluation of intelligence, the organization of thinking processes (presence of possible thought disorders), the evaluation of the available energies, the ability to socialize and to maintain interpersonal relationships, the level of maturity and stability of different psycho-affective dimensions (e.g. affections and their modulation), the sexuality, the self-evaluation, the relationship with parental figures and the social feelings. Finally, WDCT is also able to recognize symptoms of many psychological disorders as reactive or deep depression, psychotic states, conflicting diseases, etc.

In the present study an investigation of the three new scoring categories (i.e. EC, AQ and FQ) was performed with a particular focus on their inter-rater agreement and criterion validity. Regarding the EC category, the term evoke comes from the Latin word ex-vocare that means to call out, evoke or bring to mind. EC is, therefore, referred to the capacity of a specific stimulus to recall and facilitate the projection of particular psychic contents. The graphic signs were selected by Wartegg (1953) on the basis of their tendency to elicit specific conceptual domains: centrality (box 1), vitality (box 2), directionality (box 3), stability/heaviness (box 4), contraposition/overcoming (box 5), synthesis/union (box 6), delicacy/ agreeableness (box 7) and finally rounding/ closing (box 8) and that represent the evocative character of the boxes. In this vein, the individual differences in the ability to catch the EC for each category is a potential index of subjective adjustment to common thought and social settings (i.e social intelligence).

The Affective Quality is simply the assessment of the emotional connotation of each drawing. It may be partly compared to GHR (Good Human Response) and PHR (Poor Human Response) indexes that are included in Exner's scoring system (Exner, 1991, 1993, 2001), but in this system, it concerns not only Human contents but also all the others. On a theoretical point of view AQ is aimed at detecting the general emotional disposition, the type of affect that characterizes the emotional life, the degree of harmony that the subject is able to achieve in its relations with the environment and the presence of depression.

In order to assess the Form Quality (FQ) the drawings are evaluated in their clearness and immediate intelligibility, considered as indicators of the subjects' capacity to report the main form elements of their drawings. In this vein, FQ indicates the suitability of the subject to make an adequate reality testing and thus the levels of cognitive functioning, the prevalence of defence mechanisms and the capacity to be in an adequate contact with the environment.

The present research is firstly aimed at evaluating the psychometric properties of the WDCT new categories as well as their inter-rater agreement. Moreover, in order to

offer a first evidence for EC, AQ and FQ criterion validity, their capacity to discriminate among control, anxious and psychotic subjects was tested. In particular, (1) since healthy subjects are expected to show a more marked connection to social settings, a better emotional disposition and a more adequate reality testing than psychopathological patients, mean scores of the new categories are expected to be significantly higher in the former group with respect to psychopathological ones. For similar reasons, (2) significant higher mean scores are expected for anxious with respect to psychotic patients.

METHOD

Participants and Procedure

The sample was composed by 564 participants (290 males and 274 females) with a mean age of 24.6 years (*SD*= 3.54). They were classified into three subgroups that are healthy, anxious and psychotic. Healthy subjects are part of those used for the Italian standardization of WDCT while the anxious and psychotic groups has been classified by a group of psychiatrists according to DSM-IV-TR criteria. The healthy group is composed of 401 individuals (186 females) with a mean age of 19.95 years (*SD*= 3.19). The pathological group is composed by 107 psychotic individuals (females) aged 37.80 years (*SD*= 9.71) and by 56 anxious individuals (14 females) aged 32.23 years (*SD*= 9.42). Statistical tests revealed significant age [F(2, 563)= 453.44, p <.001] and gender [χ^2= 31.38, p <.001] differences among groups, indicating to use them as covariates for the subsequent analyses that are aimed to evaluate the criterion validity of new categories. The anxious group is composed by individuals with obsessive-compulsive (*n*= 18), anxiety (*n*= 24) and panic disorders (*n*= 14). The psychotic group is formed by subjects with paranoid schizophrenia (*n*= 41), schizo-affective psychosis (*n*= 11), undifferentiated schizophrenia (*n*= 6), disorganized schizophrenia (*n*= 13), simple schizophrenia (*n*= 6), delusional disorder eroto-maniac type (*n*= 1), maniac depressive psychosis (*n*= 5) and psychotic disorder NAS (*n*= 24).

Regarding the procedure, after the participants completed a brief socio-demographic questionnaire, an expert psychologist administered the Wartegg Drawing Completion Test and other tools that are irrelevant for the aims of the present research.

Measures

Each of the eight drawing was assessed in their Evocative Character, Affective Quality and Form Quality as well as in other aspects (i.e. the contents, the frequency of popular responses and original responses, the presence of special scores, the movement responses divided in main and secondary) that are not considered in the present study. In the next sections the procedures to score the drawings using the new categories and to compute their global indices were described (A general description of the WDCT and a theoretical framework of EC, AQ and FQ were illustrated in the introduction).

The Evocative Character (EC). In order to rate, in a standardized manner, the extent to which a drawing catch the evocative character of a sign the follow criteria are used.

We assign a score of 1 if the drawing picks the implicit suggestion of the stimulus-sign while .5 and 0 scores are assigned, respectively, when the evocative character has been picked only partially or not at all. To esteem a global measure of Evocative Character that synthesized all the drawing scores, we compute the total Percentage of Evocative Character as follows:

$$CE+\% = \frac{\sum CE}{8} x100$$

From a diagnostic point of view such a global score indicates adjustment to common thought, adaptation to the environment and ability to be connected to social settings (i.e social intelligence).

The Affective Quality (AQ). All drawings are scored using the following criteria. We assigned 1 to positive contents such as human, animal, natural elements, botanical and food. Scores of .5 and 0 are respectively assigned when the drawings show neutral contents (i.e. objects, letters, numbers, symbols, mineral, architectural and abstract) and negative contents (i.e. anatomical, weapons, explosion, pathological, smoke, cloud and rain). To ensure uniformity in the scoring procedure among different raters, a very huge list of drawings with their scores has been developed (Crisi, 2007). Such a list, available on the website of the Italian Institute of Wartegg, is frequently updated in real time on the basis of the judgments of a group of eight experts.To esteem a global measure of Affective Quality that synthesized all the drawing scores, we compute the total Percentage of Affective Quality as follows:

$$AF+\% = \frac{\sum AF}{8} x100$$

From a diagnostic point of view such a global score indicates the general emotional disposition, namely the type of affect that characterizes the emotional life and the presence of depression.

The Form Quality (FQ). For each drawing we assign a score of 1 when the drawing is evident and its meaning is immediately perceived by the rater, without subject's clarifications yet. Score of .5 when the drawing meaning is not immediately perceived, inducing many different interpretations that need to be clarified using the subject's explanations. Score of 0 when the drawing is incomprehensible, inaccurate or arbitrary and nothing -not even the explanation of the client- permits an appropriate interpretation. To esteem a global measure that synthesized all drawings scores, we compute the total Percentage of Form Quality as follows:

$$FQ+\% = \frac{\sum FQ}{8} x100$$

From a diagnostic point of view such a global score indicates the ability to make an adequate reality testing and thus the levels of cognitive functioning and the potential prevalence of defence mechanisms.

RESULTS

Means, standard deviations as well as skewness and kurtosis of the new categories divided by healthy, anxious and psychotic participants are reported in Table 1. Healthy subjects showed the highest mean scores for all categories whereas anxious and psychotic

Table 1. Descriptive Statistics and comparisons among healthy, anxious and psychotic subjects.

Category	Diagnosis	*Mean*	*SD*	Skewness	Kurtosis	*F*	*p*	*Partial eta square*
EC	Psychotic	58.72	15.69	-.41	-.20	33.70	<.001	.11
	Anxious	66.39	16.71	-.41	-.29			
	Healthy	77.17	1.62	-.22	-.10			
AQ	Psychotic	57.92	14.12	.04	-.65	13.65	<.001	.05
	Anxious	60.73	15.77	-.39	-.16			
	Healthy	65.77	12.07	-.05	-.04			
FQ	Psychotic	68.50	16.60	-.23	-.53	12.42	<.001	.30
	Anxious	85.93	14.59	-1.39	2.50			
	Healthy	99.79	1.33	-6.98	52.10			

subjects showed respectively the medium and the lowest ones. All categories exhibited close to normal distributions in each group, with skewness and kurtosis ranging from ± 1, except for FQ scores in the healthy group that revealed a very high mean score and low standard deviation along with a high level of negative asymmetry and positive kurtosis.

To estimate the inter-rater agreement of EC, AQ and FQ categories further 30 WDCT protocols by two independent judges and the Intraclass Correlation Coefficients (ICC) between them were computed. Results showed an adequate level of inter-rater agreement between the expert judges, with an ICC= .74 for the Evocative Character, ICC= .92 for Affective Quality and ICC= .71 for the Form Quality, giving initial evidence for their reliability.

In order to evaluate the criterion validity of the new scoring categories a counter-posed groups approach was used. In order to estimate the power of EC, AQ and FQ to discriminate among healthy, anxious and psychotic subjects a MANCOVA (controlling for age and gender) approach was used for the first two categories, while a non parametric Quade test (controlling for age and gender) was conducted for the FQ as it showed relevant deviation with respect to normal distribution. Post hoc comparisons were conducted for each category using the Sidak test.

Results showed a non significant age effect for the EC category while significant effects emerged for AQ (p <.05) and FQ (p <.001) ones. Significant gender effects emerged for all categories [p <.05 for EC; p <.001 for AQ; p <.001 for FQ]. Moreover, as reported in table 1, significant effects of diagnosis emerged for all categories, with partial eta square higher for FQ (.30) rather than for EC (.11) and AQ (.05), giving a first evidence for their criterion validity. Post hoc comparisons (i.e. Sidak tests) revealed significant differences among all groups on both EC and FC (p <.01), with a higher mean score for healthy subjects, medium for anxious and lower for psychotic ones. Differently, on AQ category healthy subjects exhibited higher mean scores than anxious (p <.01) and psychotic (p <.01) groups, but a similar mean score emerged between pathological subjects.

Discussion

In the present study, three new scoring categories for WDCT (EC, AQ and FQ) were presented and their inter-rater agreement as well as their criterion validity using a counter-posed groups approach were evaluated. In particular the discriminative power of EC, AQ and FQ was tested evaluating mean differences among healthy, anxious and psychotic individuals.

Results showed an adequate level of inter-rater agreement among two expert judges for all the new categories, giving a first evidence for their reliability. However, since inter-rater agreement is considered only as a necessary but not sufficient condition for an adequate reliability, other investigations should be conducted in order to further explore it, for instance using a test-retest approach. Obviously a special attention will be necessary to chose an appropriate interval between test and retest observations in order to avoid confusion between state variation of construct-related variability and fluctuations related to casual error. In fact, for instance, an excessively long interval may decrease test-retest correlations as a consequence of true construct variability and not for casual error variations.

Moreover, results showed that mean scores of EC, AQ and FQ are significantly different among healthy, anxious and psychotic subjects in the hypothesized direction. In particular, as expected, healthy individuals showed higher scores than pathological ones on all categories. Moreover, anxious subjects exhibited higher scores than psychotic subjects on EC and FQ categories. Since EC is an index of social adjustment, AQ of emotional personal disposition and FQ of the ability to test reality, these results give a first evidence for the criterion and construct validity of new categories.

In order to further support the criterion validity of the new scoring indices introduced here, other investigations should be conducted exploring their concurrent validity and their incremental validity in respect to other important dimensions. As regards the concurrent validity, the Evocative Character (as an index of social adjustment) may be correlated with the socialization scale of MMPI or with social intelligence scales such as the Tromsø Social Intelligence Scale (Silvera, Martinussen, & Dahl, 2001). In a similar manner, it may be interesting to evaluate the correlations between the Affective Quality (as an index of emotional disposition) and diagnostic depression scales such as the Beck Depression Inventory (BDI; Beck, Rush, Shaw, & Emery, 1979) or the CES-D (Radloff, 1977). Moreover, it may be examined the correlations between the Form Quality (as an index of reality testing ability) and potential related scales such as the Tellegen Absorption Scale (Jamieson, 2005) that permits to assess the tendency to be absorbed into fantasies and mental images.

Overall, notwithstanding the preliminary nature of the present study, these results provide a first evidence for the reliability and criterion validity of the new scoring categories and encourage to conduct new studies to confirm them.

REFERENCES

Avé-Lallemant U (1978). *Der Wartegg-Zeichentest in der Jugendberatung*. Munchen: Reinhardt

Avé-Lallemant U (1994). *Der Wartegg-Zeichentest in der Lebensberatung*. 2, erw. Aufl. 189 S., 69 Abb.

Beck AT, Rush AJ, Shaw BF, & Emery G (1979). *Cognitive therapy of depression*. New York: Guilford.

Berger E (1939). Der Sandersche phantasietest im rahmen der psychologische Eignuntersuchung jugendlischer. *Archiv Gesellschaft fur Psychologie, 103,* 499-543.

Bianchi F, Crisi A, & Di Renzo M (1996). *Il test di Wartegg nell'età evolutiva. Un contributo psicodiagnostico allo studio clinico della balbuzie, della sordità e dei disturbi dell'apprendimento*. Roma: ES MaGi.

Bohm E (1969). *Manuale di Psicodiagnostica Rorschach*. Firenze: Giunti.

Bornstein R (2007). Toward a process-based framework for classifying personality tests: Comment on Meyer and Kurtz. *Journal of Personality Assessment, 89, 202-207*.

Crisi A (1998). *Manuale del test di Wartegg*. Roma: ES MaGi.

Crisi A (2007). *Manuale del test di Wartegg* (2nd Ed.). Roma: ES MaGi.

Gini G & Iotti G (2004). *La Tromsø Social Intelligence Scale: traduzione e adattamento alla popolazione italiana*. University of Padova, Italy

Exner JE (1991). *The Rorschach: A Comprehensive System Vol. II: Interpretation* (2nd Ed.). New York: Wiley.

Exner JE (1993). *The Rorschach: a Comprehensive System Vol. I: Basic foundations* (3rd Ed.). New York: Wiley.

Exner JE (2001). *A Rorschach workbook for the Comprehensive System* (5th Ed.). Asheville: Rorschach Workshops.

Grønnerød SJ & Grønnerød C (2012). The Wartegg Zeichen Test: A literature Overview and a Meta-Analysis of Reliability and Validity. *Psychological Assessment*, *24*, 476-489.

Jamieson GA (2005). The Modified Tellegen Absorption Scale: A clearer window on the structure and meaning of absorption. *Australian Journal of Clinical and Experimental Hypnosis*, *33*, 119-139.

Kinget GM (1980). The drawing completion test. In EF Hammer (Ed.), *The Clinical application of projective drawings* (pp. 344). Springfield, IL: Thomas.

Pfeiffer W (1984). Diagnosi psichica mediante il reattivo di Wartegg. In *Contributi originali per l'approfondimento del diagnostico di Wartegg (W.Z.T.)* Verona, C.I.S.E.R.P.P.

Radloff LS (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

Rapaport D (1977). *Collected papers of David Rapaport. Il modello concettuale della psicoanalisi: scritti 1942-1960*. Milan: Feltrinelli.

Roivanen E (2009). A brief history of the Wartegg Drawing Test. *Gestalt Theory, 31*, 55-71.

Souza CV, Primi R, & Miguel FK (2007). Validade do Teste Wartegg: correlação com 16PF, BPR-5 e desempenho profissional. *Avaliação psicológica, 6,* 39-49.

Takala M & Hakkarainen M (1953). Über Faktorenstruktur und Validität des Wartegg-Zeichentests. *Annales Academiae Scientiarum Fennicae. B 81*, 95-122.

Silvera DH, Martinussen M, & Dahl TI (2001). The Tromsø Social Intelligence Scale, a self-report measure of social intelligence. *Scandinavian Journal of Psychology*, *42*, 313-319

Wass T, Mattlar C (2000). *Wartegg teckningstest*. Stockholm: Psykologiforlaget AB.

Wartegg E (1953). *Schichtdiagnostik: Der Zeichentest (WZT)*. Gottingen: Verlag fur Psychologie Hogrefe.

Received, September 30, 2015
Final Acceptance, January 10, 2016