Archives of Assessment Psychology, Vol. 8, No. 1, (57-65)
Printed in U.S.A. All rights reserved

© 2018 American Board of Assessment Psychology

Inter-Rater Reliability of the Raphael Projective System (RPS) of Scoring Projective Drawings

R. Lauren Miller, Psy.D., J.D., Ann Marie Paolino, Psy.D., Theresa Ascheman Jones, Psy.D., Alan J. Raphael, Ph.D., and Charles Golden, Ph.D.

Abstract

Purpose: The current study was conducted to examine the inter-rater reliability of the Raphael Projective System (RPS), a standardized system for the administration, scoring, and interpretation of projective drawings. In anticipation of completion of the comprehensive manual of the RPS, the authors examined the inter-rater reliability of raters who were trained on scoring House, Tree, and Person drawings. Methods: Three of the authors refined the scoring system and reached consensus on scoring for 30 projective drawing test protocols (ten House drawings, ten Tree drawings, and ten Person drawings), which they had independently scored. Five graduate psychology students were trained on the system and independently scored the same 30 drawings scored by the authors. The percentages of agreement and inter-rater reliability (IRR) between the students and the authors and between all rater pairs were calculated. Due to the high prevalence of negative responses, agreement was measured using both kappa and adjusted kappa statistic for skewed prevalence. Results: The average percentages of agreement between the student raters and the authors were as follows: 93.2% for House drawings, 85.4% for Tree drawings, 87.8% for Person drawings, and 89.7% for all drawings. Inter-rater reliability between the student raters and the authors' consensus was substantial overall, with an average k =0.88 for House drawings, k =0.72 for Tree drawings, and k =0.76 for Person drawings, with an overall average k = 0.79. An average prevalence index was measured at 0.57, and positive agreement was averaged at 0.94. Discussion: The findings suggest strong inter-rater reliability for the RPS, even for novice raters with minimal instruction on the system. The results are similar in strength to previous studies examining the reliability of projective drawing scoring systems.

Introduction

Delineating details on any projective measure is very much like describing snowflakes in that all are basically unique. Any oddity of inclusion or omission and any distortion should be considered for interpretation by the examiner. It is extremely difficult to describe each specific response obtained and its corresponding interpretation. Variables such as age, ethnicity, gender, physical status, education, and reason for evaluation must be considered in the accurate scoring and interpretation of all psychological measures. There are ever-increasing demands placed on the social sciences to create better, quicker, more accurate, and less expensive methods of assessing what is real about an individual in traits such as honesty, fitness for duty, competence, dangerousness, self-control, and more.

The seemingly infinite nature of potential responses to projective drawing prompts has made it difficult to standardize an interpretive system. This has contributed to reluctance among practitioners to use these time-honored methods amidst the current emphasis on objective and quantitative assessment measures. The Raphael Projective System (RPS) is an approach to administering and scoring projective drawings, including the Bender Gestalt Test, House-Tree-Person drawings, Kinetic Family Drawings, and Free Drawings, that aims to improve reliability and accuracy in interpretation of results. A comprehensive manual for the administration, scoring, and interpretation of projective drawings using the RPS approach is in development. In anticipation of its completion, the authors conducted this investigation into the inter-rater reliability of the scoring system.

Method

Ratings. Dr. Raphael, Dr. Ascheman, and Dr. Miller expounded and refined Dr. Norman Reichenberg's preliminary system for scoring projective drawings. Through numerous and lengthy discussions, scoring criteria were clarified to reduce subjectivity in determination of binary scoring (i.e., a score of 1 = *present* or *yes* and 0 = *absent* or *no*). Such clarifications included changing relative terms of criteria such as "close" or "disproportionate" to specific units of measurement such as "within a quarter of an inch" or "more than one third of the overall size." The scoring criteria underwent several iterations before the authors determined that subjectivity within each item had been sufficiently minimized or eliminated such that the resultant criteria could easily be taught to psychological assessment practitioners with varying levels of experience, including trainees, and uniformly applied to the interpretation of drawings from clients as young as age eight years. The number of criteria for each category of drawings varied, and are as follows: House drawings = 55, Tree drawings = 33, and Person drawings = 25.

Materials and Participants. Next, drawings were selected from a sample derived from an outpatient private practice in a metropolitan area (Miami, Florida) that had been collected over a 30-year period. Dr. Ascheman chose ten House drawings, ten Tree drawings, and ten Person drawings that were observed to contain many of the scoring criteria. Drs. Raphael, Ascheman, and Miller met to independently score each drawing and then jointly reviewed their responses until consensus was achieved regarding the correct score. Dr. Raphael then provided training on the scoring system for approximately one hour to five graduate students in a clinical psychology doctoral program with assistance from Dr. Ascheman.

Procedure. Immediately following the training, the students independently scored the same ten House drawing tests, ten Tree drawing tests, and ten Person drawing tests using the newly learned scoring system. The students' ratings were then compared to the correct ratings as determined by consensus. The number of correct scores was counted for each drawing scored by each rater. The correct scores were summed for each rater on each category of drawings (i.e., House, Tree, and Person), and total percentages correct were calculated for raters on each category. The percentages correct were averaged for the five raters for each category of drawing. The number of correct responses for each category and student rater can be found in Appendix A.

Since percentages of agreement do not correct for agreements that would be expected by chance, an inter-rater reliability (IRR) analysis was performed to assess the degree to which the five student raters consistently assigned the correct binary ratings (*present/absent*) to each of the 1,130 items (ten House drawings x 55 items, + ten Tree drawings x 33 items, + ten Person drawings x 25 items). Based on our data, Cohen's (unweighted) kappa (Cohen, 1960) was deemed an appropriate index of IRR. Kappa was therefore computed for each rater's individual item scores paired with the correct scores (as agreed upon by the authors prior to the study), then averaged to provide a single index of IRR (Conger, 1980).

We were also interested to examine how consistent the students' and authors' scores were with each other, and computed kappa for each coder pair (both students and authors) with an averaged single index.

Results

Inter-rater Agreement. Examination of the students' percentage of agreement with the authors found agreement ranging from 91.6% to 94.2% (average of 93.1% for all five students) on the House drawings. The students demonstrated 82.1% to 87.9% agreement with the authors (average of 85.4% agreement for all five students) on Tree drawings, and 83.6% to 91.2% agreement with the authors (average of 87.8% agreement for all five students) on Person drawings. The specific percentages for each rater and category of drawings are displayed below in Table 1.

Table 1

Percentages Correct by Student Raters

Drawing	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
House	94.2	91.6	94.0	94.2	91.8	93.2
Tree	85.5	82.1	87.9	85.2	86.4	85.4
Person	91.2	89.6	87.2	83.6	86.4	87.8
All Drawings	90.9	88.5	90.7	89.3	89.2	89.7

Inter-rater Reliability (IRR). In calculating the IRR, the marginal distributions of our binary ratings showed a substantially greater number of *absent* or *no* ratings, indicating a prevalence problem. The difference between the probability of *yes* and the probability of *no* is the Prevalence Index (PI). When *yes* and *no* are equally probable, the PI = 0, which gives more reliable kappa values (PI's are presented in Table 2). If the PI is high, then kappa is reduced accordingly (Cicchetti & Feinstein, 1990). Therefore, an adjusted kappa statistic for skewed prevalence (PABAK) was used to determine a more accurate level of agreement (Byrt, Bishop, & Carlin, 1991). Although PABAK has received some criticism in the literature, there seems to be agreement that kappa values can be considered reliable when good agreement is obtained despite skewed prevalence. In the Table 2 below, we listed our original Cohen's kappa statistics, the adjusted kappa, and the prevalence rates, in order to highlight the prevalence effect and allow a clearer interpretation of results than would have been available with a single index of agreement.

The resulting kappa values for each student rater's agreement with the authors' consensus (as indicated in the table below) indicate excellent agreement on House drawings, k = .864, Tree drawings, k = .710, and Person drawings k =.764, with a substantial average kappa for all drawings, k=.779.

Table 2

Inter-Rater Reliability (Student Raters x Consensus)

Rater/Drawing	Kappa	Adjusted Kappa	CI*	Prevalence** Index
Rater 1				
House	.81	.88		.63
Tree	.59	.71		.55
Person	.77	.86		.45
Rater 1 All Drawings	.73	.82	.80 - .84	.57
Rater 2				
House	.72	.84		.64
Tree	.52	.64		.51
Person	.75	.80		.44
Rater 2 All Drawings	.67	.77	.75 – .79	.56
Rater 3				
House	.80	.88	.	.64
Tree	.64	.76		.57
Person	.66	.74		.51
Rater 3 All Drawings	.72	.81	.79 – .83	.59
Rater 4				
House	.81	.88		.63
Tree	.57	.71		.57
Person	.61	.68		.43
Rater 4 All Drawings	.69	.79	.77 – .81	.57
Rater 5				
House	.73	.84		.63
Tree	.59	.73		.59
Person	.68	.74		.46
Rater 5 All Drawings	.68	.78	.76 – .80	.58
House Drawings Average	.77	.86	---	.63
Tree Drawings Average	.58	.71	---	.59
Person Drawings Average	.69	.76	---	.46
All Drawings	.70	.80	---	.57

Note. * CI: Confidence Interval 95% ; Prevalence Index = Difference between probability of “yes” and probability of “no.”

`

` We then compared the level of agreement between each student and author with each of the other students and authors. As depicted in Appendix B, there was excellent agreement between most pairs of raters (students and the authors) with an overall average kappa of $k = 0.79$.

Indices of Positive and Negative Agreement. Approximately only 20% of all items were scored *yes* in the drawings. With such a relatively small amount of positive ratings, we were interested in the probability that any two randomly assigned raters would both assign a *yes* rating. We calculated Cicchetti and Feinstein’s (1990) indices of positive and negative agreement ($p_{pos} = 2a / [N + a – d]$ and $p_{neg} = 2d / [N – a + d]$), as we felt separate indices of agreement for positively and negatively scored items would contribute to providing more transparency to our results (see Table 3 below for results). The results show that it is extremely likely (>90%) that if one rater were to assign a positive rating, a second rater would also rate a *yes.*

Table 3

Indices of Student Raters' Positive and Negative Agreement with Consensus Scores

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Positive Agreement	.94	.93	.94	.93	.93
Negative Agreement	.79	.74	.77	.75	.74

Discussion

These authors are encouraged by the results, which suggest strong inter-rater reliability among even novice assessment practitioners. These statistics are comparable to that of other studies examining the reliability of scoring methods for projective drawings. For example, Van Hutton (1994) tested the inter-rater reliability of her system by comparing consistency between two clinicians' scoring of the House-Tree-Person drawings of 20 children. After revisions of several items, the raters were in agreement 93.2% of the time, and each scoring item met the cutoff criteria of being consistently rated 80% of the time on the drawings of ten additional child subjects. Similarly, inter-rater studies of the Advanced Scoring System for the Bender Gestalt Test-Revised (ABGT-R) (Raphael, Golden, & Raphael, 2012) also yielded results suggesting that raters can easily and quickly learn the 207-item scoring system and produce high rates of agreement. After a 90-minute training, nine raters scored three BGT protocols and demonstrated 100% agreement on 156 of the 207 items, with an average of 94% specific item agreement between raters, which ranged from 56% to 100% across items. The overall agreement on the three protocols ranged from 93% to 95% (Aucone et al., 1999). Moreover, the ABGT-R was found to have satisfactory test-retest reliability when administered to outpatients diagnosed with schizophrenia twice with a mean interval of 6.4 years between administrations. After completing a 90-minute training on the system, five doctoral students and two Master's level students in counseling psychology scored 80 BGT protocols. The mean reliability was .74 and ranged from .71 to .80. Lending further support to previous inter-rater reliability testing, raters demonstrated an average of 85.97% agreement, ranging from 79% to 97% (Aucone et al., 2001).

While this study found adequate reliability of the RPS scoring system, the results have not yet been replicated. Upon release of the RPS manual, including scoring criteria, the authors encourage other assessment practitioners and researchers to learn, utilize, and further study the method. It will be especially important to investigate the reliability of the method among practitioners who were not trained by its developer.

Furthermore, additional inquiries of research may include correlating the RPS interpretive results with results from other measures. Past results by other researchers on relationships between projective and objective measures have been mixed in this regard. For example, Gillespie (1994) promoted validation by advising examiners to compare results of Mother-and-Child drawings with MMPI profiles; however, the author did not provide empirical analysis of statistical relationships between the drawing test and the objective MMPI, which could have provided more support for the validity of drawing tests. In contrast, the ABGT-R has been empirically validated to have predictive power in evaluating personality and neuropsychological functioning. The ABGT-R has been empirically correlated with the MMPI and MMPI-2 (Raphael & Golden, 2002). Future studies investigating the correlation between projective and objective methods will continue to shed light on the utility of projective methods for answering clinical and forensic assessment questions.

About the Authors

R. Lauren Miller, Psy.D., J.D. is with International Assessment Systems, Inc.Correspondence concerning this article should be addressed to Lauren Miller, lmiller.psyd.jd@gmail.com .

Ann Marie Paolino, Psy.D. is with the Neuropsychology Assessment Center, Nova Southeastern University.

Theresa Ascheman Jones, **Psy.D.** is with South Florida Evaluation & Treatment Center

Alan J. Raphael, Ph.D., ABAP is President of International Assessment Systems, Inc.

Charles J. Golden, Ph.D., ABAP, ABPP is Director of Neuropsychology at Nova Southeastern University.

The authors wish to thank Monica Gamez, M.S., Chantel Sorochuk, M.S., Scott Harcourt, M.S., Hilary Hayhurst, M.S., and Jacqueline Marsh, M.S. for serving as raters in this study.

References

Aucone, E. J., Raphael, A. J., Golden, C. J., Espe-Pfeifer, P., Seldon, J., Pospisil, T., Dornheim, L., Proctor-Weber, Z., & Calabria, M. (1999). Reliability of the Advanced Psychodiagnostic Interpretation (API) scoring system for the Bender Gestalt. *Assessment*, *68*(3), 301-303.

Aucone, E. J., Wagner, E. E., Raphael, A. J., Golden, C. J., Espe-Pfeifer, P., Dornheim, L., Seldon, J., Pospisil, T., Proctor-Weber, Z., & Calabria, M. (2001). Test-retest reliability of the Advanced Psychodiagnostic Interpretation (API) scoring system for the Bender Gestalt in chronic schizophrenics. *Assessment*, *8*(3), 351-353.

Byrt, T., Bishop, J. & Carlin, J. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*, 423-429.

Cicchetti, D.V., & Feinstein, A.R. (1990) High agreement but low kappa II: Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*, 551-558.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322-328.

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, March, 33(1)*, 159-74.

Raphael, A. J., & Golden, C. J. (2002). Relationships of objectively scored Bender variables with MMPI scores in an outpatient psychiatric population. *Perceptual and Motor Skills*, *95*(3), 1217-1232.

Raphael, A. J., Golden, C., & Raphael, M. A. (2012). *The Advanced Scoring System for the Bender Gestalt Test- Revised (ABGT-R): Ages 8-80*. Deer Park, NY: Linus Publications, Inc.

Van Hutton, V. (1994). *House-Tree-Person and Draw-A-Person as measures of abuse in children: A quantitative scoring system.* Odessa, FL: Psychological Assessment Resources, Inc.

Appendix A
Total Number of “Correct” Scores by Student Raters

Table A1

Total Number of “Correct” Scores by Student Raters

Drawing	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
House 1	51	52	52	54	51
House 2	50	50	53	51	52
House 3	54	53	53	53	52
House 4	54	53	54	54	54
House 5	50	49	49	51	48
House 6	51	49	51	50	49
House 7	53	52	54	51	51
House 8	52	52	51	51	49
House 9	54	50	55	55	53
House 10	46	44	47	49	47
Total House	515	504	519	519	506
Tree 1	31	31	31	32	30
Tree 2	25	27	25	25	23
Tree 3	30	29	27	30	27
Tree 4	28	29	30	25	26
Tree 5	27	29	28	26	28
Tree 6	24	25	29	28	30
Tree 7	29	28	32	28	32
Tree 8	30	27	30	30	31
Tree 9	28	23	29	28	29
Tree 10	31	26	30	30	30
Total Tree	283	274	291	282	286
Person 1	20	23	20	21	19
Person 2	20	21	23	21	21
Person 3	24	22	22	20	23
Person 4	23	22	19	21	24
Person 5	22	21	21	22	21
Person 6	25	24	21	23	23
Person 7	24	24	24	23	23
Person 8	23	21	23	18	19
Person 9	23	24	22	23	24
Person 10	23	23	21	16	23
Total Person	227	225	216	208	220
Total Score All Drawings	1,025	1,003	1,026	1,009	1,012

Appendix B

Inter-rater Reliability Between All Pairs of Raters Using Adjusted Kappa

Table B1

Inter-rater Reliability Between All Pairs of Raters Using Adjusted Kappa

House Drawings

	Rater1	Rater2	Rater3	Rater4	Rater5	Author1	Author2	Author3	Average
Rater1	1.00	.862	.887	.905	.865	.876	.887	.902	.883
Rater2	.862	1.00	880	.876	.836	.833	.851	.851	.856
Rater3	.887	.880	1.00	.909	.891	.865	.905	.891	.890
Rater4	.905	.876	.909	1.00	.873	.876	.902	.916	.894
Rater5	.865	.836	891	.873	1.00	.822	.869	.855	.859
Author1	.876	.833	.865	.876	.822	1.00	.895	.887	.863
Author2	.887	.851	.905	.902	.869	.895	1.00	.935	.892
Author3	.902	.851	.891	.916	.855	.887	.935	1.00	.891
Average	.883	.856	.890	.894	.859	.865	.892	.891	.879

Tree Drawings

	Rater1	Rater2	Rater3	Rater4	Rater5	Author1	Author2	Author3	Average
Rater1	1.00	.679	.697	.715	.727	.685	.770	.812	.726
Rater2	.679	1.00	.691	.612	.661	.667	.691	.648	.664
Rater3	.697	.691	1.00	.764	.848	.697	.745	.752	.742
Rater4	.715	.612	.764	1.00	.733	.679	.739	.782	.718
Rater5	.727	.661	.848	.733	1.00	.691	.727	.770	.737
Author1	.685	.667	.697	.679	.691	1.00	.697	.715	.690
Author2	.770	.691	.745	.739	.727	.697	1.00	.800	.738
Author3	.812	.648	.752	.782	.770	.715	.800	1.00	.754
Average	.726	.664	.742	.718	.737	.690	.738	.754	.721

Person Drawings

	Rater1	Rater2	Rater3	Rater4	Rater5	Author1	Author2	Author3	Average
Rater1	1.00	.840	.800	.688	.816	.800	.792	.834	.795
Rater2	.840	1.00	.752	.688	.832	.800	.760	.768	.777
Rater3	.800	.752	1.00	.680	.760	.744	.848	.792	.768
Rater4	.688	.688	.680	1.00	.712	.696	.736	.696	.699
Rater5	.816	.832	.760	.712	1.00	.744	.896	.776	.791
Author1	.800	.800	.744	.696	.744	1.00	.720	.856	.666
Author2	.792	.760	.848	.736	.896	.720	1.00	.784	.791
Author3	.832	.768	.792	.696	.776	.856	.784	1.00	.786
Average	.795	.777	.768	.699	.791	.766	.791	.786	.759