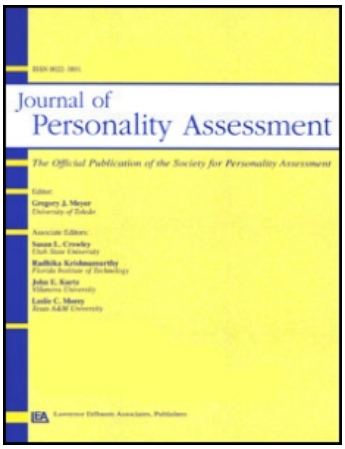

Journal of Personality Assessment
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653663

The Interrater Reliability of Elizur's Hostility Systems and Holt's Aggression Variables: A Meta-Analytical Review

Nicholas J. Katko [a]; Gregory J. Meyer [a]; Joni L. Mihura [a]; George Bombel [a]
[a] Department of Psychology, University of Toledo,

Online Publication Date: 01 July 2009

To cite this Article Katko, Nicholas J., Meyer, Gregory J., Mihura, Joni L. and Bombel, George(2009)'The Interrater Reliability of Elizur's Hostility Systems and Holt's Aggression Variables: A Meta-Analytical Review',Journal of Personality Assessment,91:4,357 — 364
To link to this Article: DOI: 10.1080/00223890902936116
URL: http://dx.doi.org/10.1080/00223890902936116

Journal of Personality Assessment, *91*(4), 357–364, 2009

ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223890902936116

Routledge
Taylor & Francis Group

The Interrater Reliability of Elizur's Hostility Systems and Holt's Aggression Variables: A Meta-Analytical Review

NICHOLAS J. KATKO, GREGORY J. MEYER, JONI L. MIHURA, AND GEORGE BOMBEL

Department of Psychology, University of Toledo

We provide a meta-analytic review of interrater reliability for scoring the 2 most commonly studied Rorschach (2003) aggression measures: (a) The Elizur (1949) Hostility Scale and its main derivative scoring systems (Holtzman, Thorpe, Swartz, & Herron, 1961; Murstein, 1956) and (b) Holt's (1977, 2005) aggression variables. Substantial reliability was observed for both Elizur's hostility score (e.g., weighted mean summary score correlation = .91, $N = 1,279$) and Holt's aggression variables (e.g., weighted mean summary score correlation = .84, $N = 226$). These meta-analytic data suggest that like the contemporary variables included in Exner's Comprehensive System or Gacono and Meloy's (1994) extended aggression scores, the historically important Elizur scoring systems and Holt aggression variables can be scored reliably.

[Supplementary materials are available for this article. Go to the publisher's online edition of the *Journal of Personality Assessment* for the following free supplemental resources: a document of results examining publication source, scoring system, amount of rater training, and severity of aggressive pathology in the sample as potential moderators of reliability.]

The Rorschach provides a unique method for understanding aggression as a personality construct. Exner's Comprehensive System (CS; Exner, 2003) is the most popular means for administering, scoring, and interpreting the Rorschach. Good reliability (McGrath et al., 2005; Meyer et al., 2002) and reasonable validity have been shown for many CS scores (Hiller, Rosenthal, Bornstein, Berry, & Brunell-Neuleib, 1999; Meyer & Archer, 2001; Meyer & Viglione, 2008; Mihura, 2008; Viglione & Hilsenroth, 2001). The CS Aggressive Movement (AG) score is assigned for aggressive actions taking place in the present. Because certain types of aggressive imagery are not captured by AG criteria, Gacono and Meloy (1994) developed an extended set of aggression indexes consisting of scores for aggressive thematic content (AgC), potential aggressive activity (AgPot), the results of aggression (AgPast), and the experience of pleasure combined with aggression (SM for sadomasochism). These indexes have shown good reliability and validity (Baity & Hilsenroth, 1999; Gacono, Bannatyne-Gacono, Meloy, & Baity, 2005). Along with AG, they currently are the most commonly used scales of aggression.

The AG and the extended aggression indexes are relatively recent additions to Rorschach-based assessment. Prior investigators typically examined aggression-related constructs using Elizur's (1949) Hostility Scale and Holt's (1977, 2005) primary (*A*1) and secondary (*A*2) Aggression variables. These measures have enjoyed considerable longevity and accumulated an impressive body of supportive research (see Aronow & Reznikoff, 1976; Goldfried, Stricker, & Weiner, 1971; Holt, 2005; Lerner & Lewandowski, 1975). In fact, they have been studied more extensively than Exner's (2003) and Gacono and Meloy's (1994) scores, and they are still occasionally used (Darolia & Joshi, 2004; Leichsenring, 2004; Liebman, Porcerelli, & Abell, 2005).

To bridge the research on the "old" measures with the "new" measures, it is necessary to understand how the contemporary measures are similar to and different from the older alternatives. One way to explore this would be to examine all of the aggression variables in a factor analysis to see how the variables are related. Once empirical and conceptual parallels are established, the collection of historical research findings can be synthesized and aligned. Before getting to this point, however, it is worthwhile to determine whether the Elizur (1949) and Holt (1977, 2005) scales can be scored with the same degree of consistency as the newer scales. Thorough evidence documenting good to excellent scoring reliability is readily available for AG (McGrath et al., 2005; Meyer et al., 2002; Viglione & Taylor, 2003) and the extended aggression indexes (Gacono et al., 2005). Our goal was to provide similarly thorough evidence for the historically most commonly used scales.

We begin by briefly describing the Elizur (1949) and Holt (1977, 2005) scales. We then outline how the combination of statistic type and computation level may moderate interrater reliability. Following this, we present the methodology used in the meta-analysis and the findings.

THE ELIZUR HOSTILITY SCALE AND ITS DESCENDENTS

Elizur's (1949) Hostility Scale is part of the Rorschach Content Test, which also includes a parallel measure for scoring anxiety imagery. It has been the most widely used system for scoring hostility (Goldfried et al., 1971). Several authors have also slightly modified the original coding criteria. We systematically reviewed relevant books and chapters (Aronow & Reznikoff, 1976; Goldfried et al., 1971; Lerner & Lewandowski, 1975) and identified 16 Elizur-derived hostility scales (see Katko, 2007) including the hostility score that is a part of the Holtzman Inkblot Test (Holtzman, Thorpe, Swartz, & Herron, 1961). For this review, we include only the most frequently researched scales as estimated by the Social Science Citation Index (SSCI). Because

Received August 20, 2007; Revised January 31, 2009.

Editor's note: Robert E. McGrath served as Editor with full decision authority on this manuscript.

Address correspondence to Nicholas J. Katko, Department of Psychology, University of Toledo, Toledo, Ohio 43606–3390; Email: nkatko@UTNet.UToledo.Edu

the Holtzman et al. (1961) system assesses many variables besides hostility, it was cited more often than Elizur (1949): 377 times versus 103, respectively. There was a notable break in the citation frequency data after Elizur, so we decided to focus on Elizur's original scale and Holtzman et al.'s (1961) modification (Hs) of it. However, Holtzman et al.'s Hs score is a hybrid of Elizur's original scale and Murstein's (1956) modification of Elizur's scale, so we added the Murstein scale to the meta-analysis despite its lower popularity (35 citations in our search).

Three levels of hostility exist in Elizur's (1949) system. Responses that directly express hostility are scored 2 points. They contain clear-cut instances of hatred, dislike, criticism, and derogation such as "animals fighting" or "butterfly who got its wing torn." Less definitive instances of hostility are assigned a value of 1. Examples include "freaks," "scissors," or "war mask." Neutral or ambiguous responses, such as "animal skin," "dangerous place," or "dead leaf" are not scored. For each protocol, all hostile response points are tabulated into a summary score representing the total hostility of the subject.

Murstein's (1956) Rorschach Hostility System (RHS) uses an 8-point scale ranging from 0 (*no hostility*) to 7 (*severe hostility*) to measure severity. The scale encompasses two dimensions. One considers the "overt-covertness" of the action, with values increasing as the perceptions move from abstract, vague, impersonal expressions to more active, direct manifestations of violence (Murstein, 1956). Thus, "two bears vying for a piece of fish receives a smaller hostility score than two bears fighting" (Murstein, 1956, p. 420), as the second instance is clearly more overt. The second dimension reflects the complexity and humanness of the organism. Thus, "two men fighting" receives a greater score than the previous example of "two bears fighting." The RHS and the Elizur (1949) Hostility Scale show a high degree of conceptual and empirical overlap, with observed correlations between .77 and .84 (Megargee, 1967; Murstein, 1956).

Like Murstein's (1956) RHS, the Hs score increases "from vague or symbolic expression or actions to more direct violent one" (Hill, 1972, p. 71). Additionally, higher scores are given as the objects involved in the percept move from inanimate objects to animals and then to humans. Hs is based on a 4-point scale ranging from 0 (*no hostility*) to 3 (*severe hostility*). Megargee (1967) reported Hs had a correlation of .75 (N = 75) with Elizur's (1949) Hostility Scale and .94 with Murstein's (1956) RHS, which Megargee considered essentially parallel forms.

Primary Process System

Holt's (1977, 2005) A1 and A2 variables are part of the Primary Process System (pripro system; Holt, 2005). Holt (2005) recently reviewed pripro research including the aggression variables. Although the review contained data obtained from primary studies, unlike this investigation it was not numerically summarized. A1 responses evidence aggressive primary process drives, particularly murderous and palpably sadomasochistic manifestations of aggression such as "this is what's left of a stomach after an armor piercing bullet has hit it—the blood and the guts splattered there and there" (Baity & Hilsenroth, 1999, p. 94). A2 responses reflect aggressive secondary processes; they contain nonlethal forms of aggression or hostility that is expressed in socially tolerable ways such as "it looks like two bears fighting," or "a sword" (Baity & Hilsenroth, 1999, p. 94). Responses can also be scored for object-relational qualities based on whether the examinee identifies with the aggressor, the victim, or the aftermath of an action, which is the primary way Holt's (1977, 2005) system differs from the Elizur-based (1949) scales.

Authors typically compute interrater reliability using Σ Ag, which is the sum across all responses containing aggressive content, or separately report results for A1 and/or A2. Although interrater reliability can be reported by object relation content subcategories (e.g., Victim score, Aggressor score), we found just three studies that did (Benfari & Calogeras, 1968; Saunders, 1991; Webster, 1995), so we limited this review to Σ Ag, A1, and A2.

Research Questions

Our primary goal was to summarize interrater reliability results. A secondary goal was to identify study characteristics that may have differentially influenced reliability. Here we focus on the type of reliability statistic used and whether the results were computed at the response level or the protocol level. Four other moderators were examined (publication source type, scoring system, amount of rater training, and degree of aggressiveness in the sample), and these results are provided in a supplement that can be found online at the *Journal of Personality Assessment* publisher's Web site.

Statistic Type

The statistic used to compute reliability can produce variable results. Percent agreement is simply the sum of all the identically scored responses divided by the total number of responses. Cohen's (1960) kappa is more conservative; it corrects percent agreement for the degree of agreement expected by chance. Percent agreement and the traditional unweighted kappa coefficient are best applied to dichotomous or multicategory data. They are not well suited to ordered, multicategory data or dimensional scales because they do not take into account degrees of agreement. Correlation-based statistics are preferred for these kinds of data. They include Pearson's r, which measures the degree of association between two sets of scores; Spearman's rho (r_s), which is a correlation based on rank ordering; the absolute agreement intraclass coefficient (ICC), which controls for any systematic differences between raters (McGraw & Wong, 1996); and weighted kappa (wκ), which is asymptotically equal to the absolute agreement ICC. These statistics do not operate under the assumption that two raters must come into exact agreement for each response to receive credit. Rather, dimensional statistics weigh the magnitude of any disagreement. In turn, they typically yield higher values on dimensional data than do categorical statistics.

Unweighted kappa was hypothesized to yield lower reliability values than the nonchance-corrected percent agreement statistic. Second, because all the scales under consideration were scored on at least a 3-point continuum at the response level (e.g., for Holt (1977, 2005): 0, A2, A1; for Elizur (1949): 0, h, H), we hypothesized categorical statistics that are insensitive to this dimensionality (i.e., unweighted kappa, percent agreement) to produce lower reliability values than dimensionally sensitive statistics (i.e., r, r_s, ICC, wκ).

Computation Level

Generally, Rorschach score reliability is either computed at the response or summary score level. At the response level,[1] the statistical calculation is made using the total number of jointly scored responses. At the summary level, reliability is computed using the total number of jointly scored protocols and not jointly scored responses. This method benefits from aggregation: When several ratings are combined, the random error associated with each measurement tends to average out (Rushton, Brainerd, & Pressley, 1983). Thus, dimensionally assessed reliability is typically greater when computed at the summary-score level rather than at the response level.

Typically, unweighted kappa and percent agreement are computed at the response level, whereas correlations are calculated at the summary level. Exceptions exist, particularly when response-level data can be coded dimensionally (e.g., Murstein's, 1956, 8-point Hostility Scale). Following statistical convention, we hypothesized that summary-level aggregation would produce greater reliability values than response-level analyses.

Method

Literature Search

We used PsycINFO, PsycINFO Historical, MEDLINE, SSCI, and ProQuest databases to identify potentially relevant English language journal articles and dissertations published between January 1949 and August 2005. The start of this period corresponded to the completion of Elizur's (1949) dissertation describing the Hostility Scale; the end reflected the start of this project.

We then applied two search strategies to the PsycINFO, MEDLINE, and ProQuest databases. For the first search, the terms *Rorschach, inkblot,* or *Holtzman* had to appear either in the abstract or title. In addition, an article had to contain one of three components in the abstract, title, or keywords: (a) the respective scale author's last name (e.g., Elizur, Holt), (b) reference to relevant target scales (e.g., pripro, Level 1), or (c) mention of a construct related to hostility or aggression (e.g., assault, violence, anger). Also in the search, we excluded any study that examined Holtzman rats or the Somatic Inkblot Series (Cassell, 1980).[2] We believed that limiting the review to just Rorschach's and Holtzman et al.'s (1961) inkblots was adequate. Our second search focused on the Holtzman et al.'s Hostility variable, Hs. Because Hs is a standard score in the Holtzman system, this search only required the term *Holtzman* to appear with a term that suggested reliability data was present in the article (e.g., interrater, agreement, scoring, kappa).[3] Finally, we searched the SSCI using bibliographic information for the seminal sources describing each scoring system to find all subsequent articles that cited them. After excluding duplicate citations, the combined searches yielded abstracts for 781 articles and 313 dissertations. Because of the expense, we only considered dissertations available through interlibrary loan.

Coding Procedure

We systematically reviewed and coded abstracts obtained from the PsycINFO, MEDLINE, and ProQuest searches based on preestablished criteria. We obtained full text articles and dissertations if the source was not a case study, review of previous research, or other nonempirical study and if it clearly or potentially (a) used a Holt (1977, 2005) or Elizur (1949) family scoring system to measure a hostility or aggression construct with the Rorschach or Holtzman et al. (1961) inkblots or (b) computed reliability for the Holtzman et al. Inkblot Test. To determine the reliability of applying these criteria, N. Katko and G. Bombel independently rated 50 randomly selected abstracts of the available citations ($N = 642$). Kappa was .81 and within the excellent range (Cicchetti, 1994), whereas agreement was 92%.

We reviewed abstracts from the SSCI using different criteria. Because these articles cited seminal studies (e.g., Elizur, 1949), we applied exclusion criteria, rather than inclusion criteria, to omit studies that clearly were not relevant. An article was obtained unless it was clear that (a) the Rorschach or Holtzman inkblots were not used; (b) it was a case study, review of previous research, or other nonempirical study; (c) test instruments were only scored for a nonhostility scale; or (d) the study assessed anxiety, and there was no positive information indicating that hostility constructs also were scored. Based on 50 randomly selected abstracts from the total available ($N = 139$), kappa was .67 and within the good range (Cicchetti, 1994). Agreement was 84%.

After applying both sets of abstract rating criteria, 450 articles and 153 dissertations remained for further screening; however, only 63 of the dissertations were available for review via loan. We obtained the full text of these articles and dissertations; and if the study provided relevant data, we coded the reliability values, the number of protocols or responses used for reliability, the number of judges (i.e., raters), and the moderator variables. We assessed the reliability of these classifications by computing percent agreement and kappa on categorical judgments (i.e., statistic type, computation level, publication source, and scoring system) and ICCs on dimensional judgments (i.e., reliability values, number of scored units, number of judges, amount of rater training, pathology severity). Interrater reliability was uniformly high in a sample of 24 articles containing data that were inde-

[1] We are grateful to the Editor for pointing out how standard reliability coefficients do not address dependencies that are inherent in response-level reliability data, which is hierarchically organized such that the responses within a person are not independent of each other.

[2] The specific terms for this search were ((rorschach OR (inkblot NOT somatic ADJ inkblot) OR (holtzman NOT holtzman ADJ rat)) AND (elizur OR murstein OR holt OR (rorschach ADJ content ADJ test) OR primary ADJ process OR secondary ADJ process OR pripro OR pri-pro OR level ADJ 1 OR level ADJ 2 OR aggress$ OR hostil$ OR primitive OR sadis$ OR violen$ OR attack$ OR assault$ OR anger OR angry OR rage OR masochis$ OR sadomasochis$ OR sado-masochis$ OR destructi$)). The symbol "$" identifies any word that has the letters preceding it. Thus, *aggress*$ identifies the terms *aggression* and *aggressive*. Holtzman rats are a strain that is commonly used in experimental research, and the Somatic Inkblots are designed to elicit body-related imagery.

[3] The specific search terms were ((holtzman AND inkblot NOT somatic ADJ inkblot) AND (reliability OR agreement OR rating OR coding OR scoring OR interrater OR inter-rater OR interscorer OR inter-scorer OR interjudge OR inter-judge OR intercoder OR inter-coder OR kappa OR intraclass OR ICC).

TABLE 1.—Characteristics of studies included in the Elizur Scoring Systems Hostility Scale interrater reliability meta-analysis.

Study	R_{xx}	Statistic Type	Comp Level	k	Score Sys	Amount of Rater Training	Severity/Setting	Protocols(P) Responses(R)
				Articles				
Barger & Sechrest (1961)	.65	*r*	R	2	H	Unknown	Nonpatient	100_R
Costa & McCrae (1986)	.93	*r*	S	2	H	Unknown	Nonpatient	29_P
Darolia & Joshi (2004)	.88	*r*	S	2	H	Unknown	Nonpatient	100_P
Elizur (1949)	.93	*r*	S	9	E	Some training	Mixed psych	15_P
Fehr (1976)	.89	% A	R	2	H	Unknown	Nonpatient	$3{,}240_R$
Forsyth (1959)	.93	*r*	S	2	E	Moderate/extensive	Nonpatient	60_P
Hess, Hess, & Hess (1999)	.99	% A	R	2	H	Unknown	Nonpatient	$(840)_R$
Holtzman	.96	*r*	S	2	H	Moderate/extensive	Inpatient	40_P
Thorpe, Swartz, & Herron (1961)	.88	*r*	S	2	H	No training	Nonpatient	96_P
Leichsenring (2004)	.86	*r*	S	2	H	Moderate/extensive	Mixed psych	50_P
McCraw & Tuma (1977)	.96	r_s	S	2	E	Unknown	Nonpatient	50_P
Megargee (1967)	.95	*r*	S	2	H	Unknown	Mixed psych	75_P
Murstein (1956)	.96	*r*	S	3	M	Some training	Nonpatient	40_P
	.70	*r*	S	3	E	Unknown		20_P
Sanders & Cleveland (1953)	.98	*r*	S	2	E	Unknown	Nonpatient	9_P
Singh & Sehgal (1979)	.94	r_s	S	2	M	Unknown	Nonpatient	31_P
Singh & Kapur (1984)	.91	r_s	S	2	M	Unknown	Nonpatient	25_P
Singh & Singh (1991)	.89	r_s	S	2	M	Unknown	Nonpatient	30_P
Spigelman, Spigelman, & Englesson (1991)	.91	% A	R	2	E	Unknown	Nonpatient	$(2{,}160)_R$
				Dissertations				
Abrams (1958)	.88	% A	S	2	E	Unknown	Inpatient	68_P
Cummings (1954)	.92	*r*	S	2	E	Unknown	Nonpatient	70_P
Fulgenzi (1965)	.86	*r*	S	2	E	Unknown	Nonpatient	120_P
Gallagher (1951)	.90	*r*	S	3	E	Some training	Mixed psych	40_P
Jackson (1990)	.89	kappa	R	2	E	Some training	Mixed psych	$(600)_R$
Leigh (1986)	.90	*r*	S	2	E	Unknown	Inpatient	20_P
Lurie (1988)	.75	*r*	S	2	H	Unknown	Inpatient	10_P
Speare (1972)	.92	% A	R	2	H	Unknown	Inpatient	108_R
Stroad (1978)	.95	ICC	S	3	E	Some training	Inpatient	49_P
Wolf-Dorlester (1976)	.98	% A	R	2	H	Some training	Nonpatient	$(4{,}005)_R$
	.94	*r*	S	2	H			86_P

Note. Numbers in parentheses indicate an estimate. $R_{XX'}$ = Reliability; Comp Level = Computation Level (R = response level, S = summary score level); k = number of judges; Score Sys = scoring system (E = Elizur, H = Holtzman, M = Murstein); Severity/Setting (Mixed psych = mixed psychiatric); % A = percent agreement; ICC = intraclass correlation. For Protocols or Responses numbers in parentheses indicate an estimate.

pendently coded by N. Katko and G. Bombel. Agreement for the categorical variables was excellent, with agreement ranging from 92% to 100% and kappa ranging from .82 to 1.0. Similarly, reliability for dimensionally coded variables was excellent, with ICCs from .92 to 1.0. All initial disagreements were reconciled among the coders.

At times, the same interrater reliability data were presented in different articles, so we limited the analyses to a single set of findings for the reliability sample. In most instances, these were readily identified because authors cited the article in which interrater reliability originally appeared. In one instance, identical interrater reliability information (i.e., observed reliability value, *N*, sample, statistic type, computation level, etc.) was presented in a pair of studies (Ackerman, Hilsenroth, Clemence, Weatherill, & Fowler, 2000; Blais, Hilsenroth, Fowler, & Conboy, 1999). The raters who scored the protocols in these studies, M. Hilsenroth and C. Fowler, were contacted for clarification. Although they were confident the samples were different, the original data was no longer available, so we took a conservative position and just used the data from the Ackerman et al. article. In total, we obtained interrater reliability data from 26 articles and 16 dissertations. Three studies reported reliability in two samples, culminating in a total of 45 samples.

Effect Sizes

We computed weighted summary effect sizes across all reliability coefficients. The formula n (k–1) was used to weight each sample. This formula assigns greater weights to studies with larger samples (n) and/or more raters (k) so that samples contributed to the final estimate in proportion to their number of independently paired judgments. Because rho and r have similar statistical properties when applied to coding decisions, we combined them to form a single coding category (r_s/r). We then performed separate analyses for findings based on ICCs and r_s/r values. Because the results were comparable (see Tables 1 and 2), summary findings are based on a single category of dimensionally based correlation statistics (i.e., r_s/r/ICC).

Data Analyses

If authors reported the number of protocols used in the reliability analyses but not the number of responses, each protocol was estimated to contain 20 responses. For the Holtzman et al. (1961), each protocol was estimated to have 1 response for each card administered unless stated otherwise. We excluded four studies because they contained reliability coefficients too ambiguous to definitively code (Benfari & Calogeras, 1968;

TABLE 2.—Characteristics of studies included in the Holt aggression scoring interrater reliability meta-analysis.

Study	R_{xx}	Statistic Type	Comp Level	k	Score Sys	Amount of Rater Training	Severity/Setting	Protocols (P) Responses (R)
				Articles				
Ackerman, Hilsenroth, Clemence, Weatherill, & Fowler (2000)	.87	% A	R	2	ΣAg	Unknown	Outpatient	$(400)_R$
	.98	% A	R	2	A1			
	.96	% A	R	2	A2			
Baity & Hilsenroth (1999)	.63	kappa	R	2	A1	Some training	Outpatient	543_R
	.64	kappa	R	2	A2			
Fowler, Hilsenroth, & Handler (1995)	.81	% A	R	2	ΣAg	Unknown	Mixed psych	$(600)_R$
Fowler, Hilsenroth, & Handler (1996)	.87	*r*	S	2	A1	Unknown	Mixed psych	20_P
	.89	*r*	S	2	A2			
Fowler, Hilsenroth, & Nolan (2000)	.87	ICC	S	2	ΣAg	Unknown	Violent off	20_P
	.98	% A	R	2	ΣAg			$(400)_R$
Liebman, Porcerelli, & Abell (2005)	.64	kappa	R	2	A1	Unknown	Mixed off	$(2,418)_R$
	.91	kappa	R	2	A2			
Rosegrant (1982)	.91	*r*	S	2	A1	Some training	Nonpatient	20_P
Saunders (1991)	.70	*r*	S	2	ΣAg	Some training	Violent off	$(10)_P$
Wiseman & Rehyer (1973)	.89	*r*	S	2	ΣAg	Some training	Violent off	10_P
				Dissertations				
Caldwell (1993)	.52	kappa	R	2	A1	Unknown	Nonpatient	$(400)_R$
Greco (1989)	.95	ICC	S	2	ΣAg	Moderate/extensive	Mixed off	33_P
Hoffman (1976)	.92	*r*	S	2	ΣAg	Moderate/extensive	Nonpatient	25_P
	.79	*r*	S	2	A1			
Johnson (1980)	.95	*r*	S	2	ΣAg	Unknown	Mixed psych	15_P
Safrin (1974)	.71	*r*	S	2	ΣAg	Moderate/extensive	Nonpatient	63_P
Phillips (1988)	.90	ICC	S	2	A1	Unknown	Unknown	$(10)_P$
	.92	ICC	S	2	A2			

Note. Numbers in parentheses indicate an estimate. $R_{XX'}$ = reliability; Comp Level = computation level (R = response level, S = summary score level); k = number of judges; Score Sys = scoring system (Σ Ag = total aggression, A1 = primary aggression, A2 = secondary aggression); Severity/Setting (Mixed off = mixed offender, Mixed psych = mixed psychiatric, violent off = Violent-offender); % A = percent agreement; ICC = intraclass correlation. For Protocols or Responses numbers in parentheses indicate an estimate.

Ganzer, Sarason, Green, & Rinke, 1970; Siegel, 1956; Vernallis, 1953). Two studies departed from statistical convention and computed reliability at a level of analysis seldom observed for that particular statistic type. One computed *r* at the response level (Barger & Sechrest, 1961), and the other computed percent agreement at the summary-score level (Abrams, 1957).

RESULTS

For the Elizur (1949) system analyses, we relied on 17 articles and 10 dissertations. Three studies provided two samples; so in total, 30 samples contributed data. A summary of the coding for each sample is presented in Table 1. Reliability values ranged from .65 to .99, with 1 sample using kappa, 6 using percent agreement, and 23 using r_s/*r*/ICC. A total of 7 samples computed reliability at the response level and 23 at the summary-score level. A total of 25 samples computed reliability with two raters, four used three raters, and one used nine raters.

Following Cicchetti's (1994) benchmarks for interpreting kappa, which are also applicable to ICC, the weighted mean reliability was excellent (>.74) for kappa at the response level ($M = .89$; $N = 600$) and for correlation-based statistics at the summary-score level ($M = .91$; $N = 1279$, $SD = .05$). Reliability was considered good for the one study reporting a correlation at the response level ($r = .65$, $N = 100$). Although there are not established benchmarks for interpreting percent agreement, it was 94% across studies examining it at the response level ($N = 10, 353$, $SD = .04$) and 88% for the one study that computed it at the summary-score level ($N = 68$).

The Holt (1977, 2005) results were drawn from nine articles and six dissertations. Six of the sources reported reliability for more than one type of statistic, scale, or level of analysis; and in total, there were 23 different reliability statistics (see Table 2). Reliability values ranged from .52 to .98. A total of 5 statistics were reported as kappa, 5 as percent agreement, and 13 as *r*/ICC. A total of 10 statistics were computed at the response level and 13 at the summary-score level. All used two raters.

Because there were a limited number of results, we initially examined if the available information could be maximized by collapsing across the three aggression scoring categories (Σ Ag, A1, and A2). We averaged the *r*/ICC results for Σ Ag, A1, and A2 separately at the summary-score level, which encompassed results from most studies. The separate A1 ($M = .86$, $N = 176$, $SD = .05$) and A2 ($M = .90$, $N = 30$, $SD = .01$) results were similar to each other. Further, when we averaged the A1 and A2 variables within studies, the value ($M = .86$, $N = 75$, $SD = .05$) approximated the value for Σ Ag ($M = .83$, $N = 176$, $SD = .11$). We took these results as reasonable support for pooling Σ Ag, A1, and A2 in subsequent analyses.

The overall results indicate good weighted mean reliability for kappa at the response level ($M = .73$, $N = 3, 361$, $SD = .13$) and for correlation-based statistics at the summary-score level ($M = .84$, $N = 226$, $SD = .10$). Mean response level agreement was 89% ($N = 1, 943$, $SD = .06$).

DISCUSSION

In our meta-analytic review, we found reliability of substantial magnitude at the summary-score level for the Elizur (1949)

Scoring Systems ($M\ r_s/r$/ICC = .91) and the Holt (1977, 2005) aggression variables (Mr/ICC = .84). Furthermore, response level reliability was high for percent agreement (Elizur: .94; Holt [1977, 2005]: .89) and good for kappa (Elizur: .89; Holt [1977, 2005]: .73). The methodological features of our study, including an extensive literature search, reliable judgments for study inclusion, and conservative statistical analyses, provide credibility for these results.

Coefficients were consistently higher for the Elizur (1949) system than the Holt (1977, 2005) system, which suggests that raters have an easier time consistently applying coding rules for the former than the latter. However, this prospect has yet to be empirically tested using the same raters with the same degree of training in each system.

Our good reliability findings parallel results for the more recently developed Rorschach measures of aggression. For instance, for the CS AG score, Meyer et al. (2002) found an ICC of .90 across 219 protocols, Viglione and Taylor (2003) recorded an ICC of .89 across 84 protocols, and McGrath et al. (2005) reported a kappa of .76 across 1,588 observations at the response level. Gacono et al. (2005) combined the mean reliability values across six published articles, six dissertations, and one book chapter for each of their aggression scores. The weighted mean kappa (presumably computed at the response level) was excellent for AgC (.88; range = .80–.95), AgPot (.83; range = .66–1.0), and AgPast (.86; range = .65–.94). Reliability also was excellent for SM, although it was only examined in one sample (kappa = .91). ICC values were computed in just one small study (presumably at the protocol level), and they were .97, .92, and .95 for AgC, AgPot, and AgPast, respectively. Our study thus provides evidence that the Elizur (1949) and Holt (1977, 2005) approaches to scoring aggression, both of which have a long history of use in the Rorschach literature, can be coded as reliably as other contemporary scales.

A limitation of this research is that the observed reliability is bound to a particular universe of studies (i.e., published literature, dissertations), which may not be identical with the universe of actual instances in which the scales are applied (Hunt, 1997). Also, to the degree that raters in the studies (e.g., researchers, psychology graduate students) are unrepresentative of all raters coding the systems, the figures may not be fully generalizable. These concerns may also be understood in terms of research reliability versus field reliability. Importantly, evidence indicates that reliability tends to be largely consistent across settings at least for CS variables (McGrath et al., 2005; Meyer et al., 2002). As noted in our online supplement examining potential moderators, varying levels of rater training did not markedly influence reliability results; this also provides partial support for the generalizability of these findings.

Given that sufficient interrater reliability has been demonstrated for the Elizur (1949) and Holt (1977, 2005) variables, as well as Exner's (2003) AG score, and the Gacono and Meloy (1994) extended aggression indexes (Baity & Hilsenroth, 1999; Gacono et al., 2005), subsequent research should attempt to clarify existing validity issues. These include identifying the unique and incremental validity that each scale may have for predicting observable displays of aggressive behavior as well as the extent of conceptual and empirical overlap among these different approaches to scoring aggression. To the extent that shared variance can be documented through the joint factor analyses of the various scales (e.g., Baity & Hilsenroth, 1999; Liebman et al., 2005), the many years of validity research on the older Holt (1977, 2005) and Elizur scales may be synthesized and aligned with the more recently developed scales of Exner and Gacono and Meloy (1994).

Acknowledgments

We thank Jeanne Brockmyer for her helpful input on this project.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Abrams, J. (1957). *Chlorpromazine in the treatment of chronic schizophrenia: A comparative investigation of the therapeutic value of chlorpromazine in effecting certain psychological and behavioral changes in chronic schizophrenic patients.* Unpublished doctoral dissertation, New York University.

*Ackerman, S. J., Hilsenroth, M. J., Clemence, A. J., Weatherill, R., & Fowler, J. C. (2000). The effects of social cognition and object representation on psychotherapy continuation. *Bulletin of the Menninger Clinic, 64,* 386–408.

Aronow, E., & Reznikoff, M. (1976). *Rorschach content interpretation*. New York: Grune & Stratton.

*Baity, M., & Hilsenroth, M. (1999). Rorschach aggression variables: A study of reliability and validity. *Journal of Personality Assessment*, 72, 93–110.

*Barger, P. M., & Sechrest, L. (1961). Convergent and discrimination validity of four Holtzman Inkblot Test variables. *Journal of Psychological Studies, 12,* 227–236.

Benfari, R. C., & Calogeras, R. C. (1968). Levels of cognition and conscience typologies. *Journal of Projective Techniques and Personality Assessment, 32,* 466–474.

Blais, M. A., Hilsenroth, M. J., Fowler, J. C., & Conboy, C. A. (1999). A Rorschach explanation of the *DSM–IV* borderline personality disorder. *Journal of Clinical Psychology, 55,* 563–572.

*Caldwell, E. (1993). A longitudinal study of self-image and regression in aging architects of varying degrees of creativity (Doctoral dissertation, University of Montreal, 1995). *Dissertation Abstracts International, 55*(7-B), 3007.

Cassell, W. A. (1980). *Body symbolism and the somatic inkblot series*. Anchorage, AK: Aurora Publishers.

Cicchetti, D. V. (1994). Guidelines, criteria and rules of thumb for evaluating normed and standardized instruments in psychology. *Psychological Assessment, 6,* 284–290.

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

*Costa, P. T., & McCrae, R. R. (1986). Age, personality, and the Holtzman Inkblot Technique. *International Journal of Aging and Human Development, 23,* 115–125.

*Cummings, C. P. (1954). *The role of various psychological variables in children's nailbiting behavior*. Unpublished doctoral dissertation, Pennsylvania State University, University Park.

*Darolia, C. R., & Joshi, H. L. (2004). Standardization of Holtzman Inkblot Technique on Indian high school students. *Psychological Studies, 49,* 52–58.

*Elizur, A. (1949). Content analysis of the Rorschach with regard to anxiety and hostility. *Rorschach Research Exchange and Journal of Projective Techniques, 13,* 247–287.

Exner, J. (2003). *The Rorschach: A Comprehensive System* (4th ed.). New York: Wiley.

*Fehr, L. A. (1976). Construct validity of the Holtzman Inkblot anxiety and hostility scores. *Journal of Personality Assessment, 40,* 483–486.

*Forsyth, R. P. (1959). The influences of color, shading and Welsh anxiety level on Elizur Rorschach content test analyses of anxiety and hostility. *Journal of Projective Techniques, 23,* 207–213.

*Fowler, C., Hilsenroth, M. J., & Handler, L. (1995). Early memories: An exploration of theoretically derived queries and their clinical utility. *Bulletin of the Menninger Clinic, 59,* 79–98.

*Fowler, C., Hilsenroth, M. J., & Handler, L. (1996). A multimethod approach to assessing dependency: The early memory dependency probe. *Journal of Personality Assessment, 67,* 399–413.

* Fowler, C., Hilsenroth, M. J., & Nolan, E. (2000). Exploring the inner world of self-mutilating borderline patients: A Rorschach investigation. *Bulletin of the Menninger Clinic, 64,* 365–385.

*Fulgenzi, L. B. (1965). *The repression-sensitization personality dimension and aggressive behavior.* Unpublished doctoral dissertation, University of Oklahoma, Oklahoma City.

Gacono, C. B., Bannatyne-Gacono, L., Meloy, J. R., & Baity, M. R. (2005). The Rorschach extended aggression scores. *Rorschachiana, 18,* 164–190.

Gacono, C. B., & Meloy, J. R. (1994). The aggression response. In C. B. Gacono & J. R. Meloy (Eds.), *The Rorschach assessment of aggressive and psychopathic personalities* (pp. 259–278). Hillsdale, NJ: Lawrence Erlbaum Associates.

*Gallagher, J. (1951). *An investigation into factors differentiating college students who discontinue non-directive counseling from college students who continue counseling*. Unpublished doctoral dissertation, Pennsylvania State University, University Park.

Ganzer, V. J., Sarason, I. G., Green, C. T., & Rinke, C. (1970). Effects of model's and observer's hostility on Rorschach, interview and test performance. *Journal of Projective Techniques and Personality Assessment, 34,* 302–315.

Goldfried, M., Stricker, G., & Weiner, I. (1971). *Rorschach handbook of clinical and research applications*. Englewood Cliffs, NJ: Prentice Hall.

*Greco, C. M. (1989). Rorschach ego development in homicidally-aggressive youth (Doctoral dissertation, University of Virginia, 1989). *Dissertation Abstracts International, 50*(9-B), 4218.

*Hess, T. H., Hess, K. D., & Hess, A. K. (1999). The effects of violent media on adolescent inkblot responses: Implications for clinical and forensic assessments. *Journal of Clinical Psychology, 55,* 439–445.

Hill, E. (1972). *The Holtzman Inkblot Technique*. San Francisco: Jossey-Bass.

Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (1999). A comparative meta-analysis of Rorschach and MMPI validity. *Psychological Assessment, 11,* 278–296.

*Hoffman, R. A. (1977). Relationships between primary process functioning and aspects of adjustment in elementary school boys (Doctoral dissertation, New York University, 1976). *Dissertation Abstracts International, 37*(9-B), 4684.

Holt, R. (1977). A method for assessing primary process manifestations and their control in Rorschach responses. In M. A. Rickers-Ovsiankina (Ed.), *Rorschach psychology* (pp. 375–420). New York: Krieger.

Holt, R. (2005). The Pripro scoring system. In R. F. Bornstein & J. M. Masling (Eds.), *Scoring the Rorschach: Seven validated systems* (pp. 191–235). Hillsdale, NJ: Lawrence Erlbaum Associates.

*Holtzman, W. H., Thorpe, J. S., Swartz, J. D., & Herron, E. W. (1961). *Inkblot perception and personality: Holtzman Inkblot Technique.* Austin: University of Texas Press.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Sage.

*Jackson, A. E. (1990). *Psychoanalytic conceptualizations of borderline personality disorder: A Rorschach study of borderline adolescents.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

*Johnson, D. R. (1980). *Cognitive organization in paranoid and nonparanoid schizophrenia: A study of self-other representations in improvisational role-playing and on the Rorschach*. Unpublished doctoral dissertation, Yale University, New Haven, Connecticut.

Katko, N. J. (2007). *The interrater reliability of scoring Elizur's Hostility Scale for the Rorschach*. Unpublished master's thesis, University of Toledo, Toledo, Ohio.

*Leichsenring, F. (2004). The role of structure in the assessment of psychopathology. *European Journal of Psychological Assessment, 20,* 275–282.

*Leigh, D. L. (1986). *The relation between "pathogenesis" in parents and Rorschach indices of ego functioning in preadolescent daughters.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

Lerner, P. M., & Lewandowski, A. J. (1975). The measurement of primary process manifestations: A review. In P. Lerner (Ed.), *Handbook of Rorschach scales* (pp. 181–214). Oxford, England: International Universities Press.

*Liebman, S. J., Porcerelli, J., & Abell, S. C. (2005). Reliability and validity of Rorschach aggression variables with a sample of adjudicated adolescents. *Journal of Personality Assessment, 85,* 33–39.

*Lurie, I. (1988). *Personality differences in Israeli and American Hebrew-English Bilinguals.* Unpublished doctoral dissertation, The California School of Professional Psychology, Berkley.

*McCraw, R. K., & Tuma, J. M. (1977). Rorschach content categories of juvenile diabetics. *Psychological Reports, 40,* 818.

McGrath, R. E., Pogge, D. L., Stokes, J. M., Cragnolino, A., Zaccario, M., Hayman, J., et al. (2005). Comprehensive System scoring reliability in an adolescent inpatient sample. *Assessment, 12,* 199–209.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

Megargee, E. I. (1967). The relation of TAT and inkblot aggressive content scales with each other and with criteria of overt aggressiveness in juvenile delinquents. *Journal of Projective Techniques and Personality Assessment, 31,* 48–60.

Meyer, G. J., & Archer, R. P. (2001). The hard science of Rorschach research: What do we know and where do we go? *Psychological Assessment, 13,* 486–502.

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Fowler, C. J., Piers, C. C., et al. (2002). An examination of interrater reliability for scoring the Rorschach CS in eight data sets. *Journal of Personality Assessment, 78,* 219–274.

Meyer, G. J., & Viglione, D. J. (2008). An Introduction to Rorschach Assessment. In R. P. Archer & S. R. Smith (Eds.), *A guide to personality assessment: Evaluation, application, and integration* (pp. 281–336). New York: Routledge**.**

Mihura, J. L. (2008, March). *A review of the validity research on the Rorschach Comprehensive System variables*. Workshop presented at the Society for Personality Assessment, New Orleans, LA.

*Murstein, B. (1956). The projection of hostility on the Rorschach and as a result of ego-threat. *Journal of Projective Techniques, 20,* 418–428.

*Phillips, S. H. (1988). Personality functioning in bulimia: A study of defense mechanisms and drive expression in the Rorschach's of bulimic, borderline, and normal women (Doctoral dissertation, Northwestern University, Evanston, Illinois, 1988). *Dissertation Abstracts International, 49*(11-B), 5030.

*Rosegrant, J. (1982). Primary process patterning in college students' Rorschach responses. *Journal of Personality Assessment, 46,* 578–581.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

*Safrin, R. (1974). Primary process thought in the Rorschach's of girls at the oedipal, latency and adolescent stages of development (Doctoral dissertation, New York University, 1974). *Dissertation Abstracts International, 35*(11-B), 5650.

*Sanders, R., & Cleveland, S. E. (1953). The relationship between certain examiner personality variables and subjects' Rorschach scores. *Journal of Projective Techniques, 17,* 34–50.

*Saunders, E. A. (1991). Rorschach indicators of chronic childhood sexual abuse in female borderline inpatients. *Bulletin of the Menninger Clinic, 55,* 48–71.

Siegel, S. M. (1956). The relationship of hostility to authoritarianism. *Journal of Abnormal and Social Psychology,* 52, 368–372.

*Singh, S., & Kapur, R. (1984). Psychometric and behavioral correlates of group Rorschach measure of hostility. *Indian Journal of Clinical Psychology, 11,* 35–44.

*Singh, S., & Sehgal, M. (1979). Correlates of group Rorschach measures of hostility and their factorial structure. *Projective Psychology, 24,* 25–33.

*Singh, S., & Singh, D. (1991). Personality correlates of Rorschach measure of hostility. *British Journal of Projective Psychology, 36,* 17–24.

*Speare, J. (1972). *HIT hostility, MMPI control, and behavioral aggression.* Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

*Spigelman, G., Spigelman, A., & Englesson, I. (1991). Hostility, aggression, and anxiety levels of divorce and nondivorce children as manifested in their responses to projective tests. *Journal of Personality Assessment, 56,* 438–452.

*Stroad, M. (1978). *Signs of hostility and anxiety in Rorschach test content differentiating Borderline and acute schizophrenic individuals.* Unpublished doctoral dissertation, California School of Professional Psychology, Berkeley.

Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions, and future. *Psychological Assessment, 13,* 452–471.

Viglione, D. J., & Taylor, N. (2003). Empirical support for interrater reliability of Rorschach Comprehensive System coding. *Journal of Clinical Psychology, 59,* 111–121.

Vernallis, F. F. (1953). Teeth-grinding: Some relationships to anxiety, hostility, and hyperactivity. Unpublished doctoral dissertation, Pennsylvania State University, University Park.

Webster, S. E. (1995). *Rorschach performance of African American Adolescents referred for outpatient psychological evaluation*. Unpublished doctoral dissertation, University of Virginia, Charlottesville.

*Wiseman, R. J., & Reyher, J. (1973). Hypnotically induced dreams using the Rorschach inkblots as stimuli: A test of Freud's theory of dreams. *Journal of Personality & Social Psychology, 27,* 329–336.

*Wolf-Dorlester, B. (1976). *Creativity, adaptive regression, reflective eye movements, and the Holtzman movement responses.* Unpublished doctoral dissertation, City University of New York, New York.