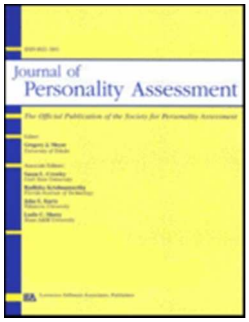

Introducing a Bayesian Approach to Determining Degree of Fit with Existing Rorschach Norms

Journal:	*Journal of Personality Assessment*
Manuscript ID:	JPA-2014-144.R1
Manuscript Type:	General Submission
Keywords:	Rorschach < Measures, Norms, International, Bayes

Abstract

This article offers a new methodological approach to investigate the degree of fit between an independent sample and two existing sets of norms. Specifically, with a new adaptation of a Bayesian method, we developed a user-friendly procedure to compare the mean values of a given sample to those of two different sets of Rorschach norms. To illustrate our technique, we used a small, U.S. community sample of 80 adults and tested whether it resembled more closely to the standard, Comprehensive System norms (CS 600; Exner, 2003), or to a recently introduced, internationally-based set of Rorschach norms (Meyer, Erdberg, & Shaffer, 2007). Strengths and limitations of this new statistical technique are discussed.

Keywords: Rorschach; Norms; International; Bayes

Introducing a Bayesian Approach to Determining Degree of Fit with Existing Rorschach Norms

Establishing accurate normative data for the Rorschach (Rorschach, 1921) Comprehensive System method (CS; Exner, 2003) is crucial to its use in clinical and forensic practice. As with other tests, Rorschach interpretation rests on (1) quantitative, nomothetic normative comparisons and (2) qualitative idiographic, individualized inferences. Thus, evaluating deviations from normative expectations is a central component in quantitative interpretation. Yet, quite surprisingly, the debate about optimal norms for the Rorschach is not settled (Viglione & Hilsenroth, 2001; Wood, Nezworski, Garb, & Lilienfeld, 2001a, 2001b). Indeed, while many practitioners currently use the standard CS normative values (Exner, 2003), some authors (e.g., Meyer, Erdberg, & Shaffer, 2007; Meyer, Viglione, Mihura, Erard, & Erdberg, 2011) have recently advocated that a composite set of internationally-based normative data would improve the applicability of the test.

The original US adult, non-patient CS reference sample is comprised of 600 adult Rorschach protocols (CS 600; Exner et al., 2001). It is balanced by gender, and fairly well stratified with regard to geographic distribution and socioeconomic level. The percentage of non-Caucasian is relatively small by today's optimal standards, about 18%. Most of the respondents were collected from workplace, and were relatively young and well-educated (mean age = 31.7; mean years of education = 13.4); all were volunteers. The Rorschach scores obtained from this large, adult, reference sample represent the most recent, "official" CS norms for adults (Exner, 2003), as well as the foundation for the CS computerized interpretation program currently in use, the fifth version of the Rorschach Interpretation Assistance Program (RIAP-5; Exner & Weiner, 2003).

Despite the size of this sample being relatively large, the CS 600 data were collected a long time ago, mostly in the seventies, so that some relevant changes in the community characteristics might have occurred during this time span. For this and other reasons, some researchers have asserted that the mean value of certain variables might be unrepresentative of the current nonclinical, adult population. In line with this position, Shaffer, Erdberg, and Haroian (1999) reported that a non-patient US sample of 123 adults from Fresno, California, produced significantly shorter and less complex records. Similarly, Wood et al. (2001a, 2001b) showed that a number of non-patient samples from the literature produced notably different mean values from the CS norms (CS 600), with effect sizes ranging from small to very large. Form quality (FQ) related (i.e., X+%, X-%) and color related variables (i.e., Afr, FC, WSumC), as well as popular (P), whole, realistic human content (Pure H), diffuse shading (Y), and reflection (Fr, rF) responses were the most problematic variables. Other empirical evidence also showed that the distributions for form quality (FQ) and number of responses (R) among non-patient samples might diverge from the CS normative expectations (Viglione & Hilsenroth, 2001). Importantly, the direction and size of all these differences suggest that the CS 600 might make normal adults appear maladjusted.

To address these issues, Exner and Erdberg (2005) collected a new normative reference sample, comprised of 450 non-patient adults (CS 450). Many variables' scores in this new sample were midway between the CS 600 and divergent nonpatient samples (Shaffer, Erdberg, and Haroian; 1999; Viglione & Hilsenroth, 2001; Wood et al. 2001a, 2001b), and some other variables were closer to one or the other. For example, the CS 450 mean values of Popular responses (P), the Affective Ratio (Afr), Level 2 Cognitive Special Scores (Lvl 2), and Form Dominated Color (FC) were relatively close to the CS 600 values. However, there were still notable differences between the CS 450 and CS 600 for Form Quality (FQ), Unusual Detail Locations (Dd), and Experience Actual (EA), and again in the

same direction. Ultimately, for a variety of reasons the CS 450 was not adopted as the normative foundation for the CS (Exner & Erdberg, 2005).

In 2007, Meyer et al. (2007) presented descriptive data from 4,704 Rorschach records from non-patient samples from Argentina, Australia, Belgium, Brazil, Denmark, Finland, France, Greece, Israel, Israel, Italy, Japan, Peru, Portugal, Romania, Spain, the Netherlands, and the United States. The mean age of the entire, combined sample was 36.65 (*SD* = 11.71). Years of education, gender, and race were not reported. Analyses, of these international data revealed that both the CS 450 collected by Exner and Erdberg (2005) and the CS 600 collected by Exner et al. (2001) diverge somewhat from most of the other samples for a large proportion of the variables. Perhaps more importantly, applying CS 600 interpretive routines to all these samples would result in pathologized interpretation of these nonpatients[1]. To provide the Rorschach users with more representative normative benchmarks and to reduce the risk of overly pathological interpretations, Meyer et al. (2007) used these data to compile a new international normative reference sample. With reference to these international data and to previous non-patient studies, Viglione and Meyer (2008) summarized the recurring main differences between the CS 600 and other samples and reported that other samples frequently produced more unusual location responses, inferior form quality, fewer elaborated, positive human representations, less color, and fewer texture responses.

Despite the potential utility of the international norms provided by Meyer et al. (2007), some authors raised concerns regarding the quality and integrity of those data. In particular, Ritzler and Sciara (2009) argued that (a) the majority of the studies included in the final, combined sample used graduate students as examiners, which might reduce the overall complexity of the records; (b) most of the data were only collected in large urban areas, which might limit the generalizability of the findings; (c) there was some variability in the

[1] Additional information on these international data can be found in the 2007 Special Issue of the *Journal of Personality Assessment* devoted to International Reference Samples for the Rorschach Comprehensive System.

exclusion criteria adopted by the different studies, so that it is not clear the extent to which the final, combined sample might represent a normative vs. a non-patient sample; (d) despite the large sample size of the final, combined sample, the sample size of most of the individual studies was rather small, which might create problems in terms of representativeness and accuracy of the stratification; and (e) it is not clear to what extent all the studies included in Meyer et al. (2007) followed the CS guidelines strictly, in terms of administration and warm-up procedures.

In response to these concerns, Meyer and colleagues investigated the extent to which the quality of their data might have affected their overall mean scores (Meyer, Shaffer, Erdberg, Viglione, & Mihura, 2009). Specifically, the authors conducted moderation analyses aimed at exploring whether considering "less optimal samples" vs. "more optimal samples" would lead to different conclusions from what was published in Meyer et al. (2007). Less optimal samples (n = 5) were defined as characterized by use of just one examiner, use of examiners with no previous administration experience, and/or incomplete information on examiners and/or quality control. More optimal samples (n = 4) were defined as characterized by use of experienced examiners and inclusion (and description) of ongoing quality control efforts. All remaining samples (n = 12) were considered as "mid-range." Overall, the results of these analyses revealed that the three quality-based groups were very similar to each other, producing virtually identical mean scores, with the largest differences being within four T-score points.

Some individuals judge the research findings convincing enough to start using the international norms as their primary reference data. In fact, Meyer, Viglione, Mihura, Erard, and Erdberg (2011) used a portion of the international norms to produce the reference data for their recently introduced Rorschach Performance Assessment System (R-PAS). Other

authors (e.g., Ritzler & Sciara, 2009) continue to use the CS 600 and advocate for great caution when considering other norms.

Without attempting to settle such a complex debate, the current study addresses a number of related methodological issues. More in detail, the current article introduces a new method to investigate the degree of fit between an independent sample and two existing sets of Rorschach norms.

The Methodological Focus of the Current Study

The question of which norms to use for the Rorschach is not resolved. Moreover, adequate statistical methods to address this question have not been specified. Testing the representativeness of a set of norms, indeed, is not an easy statistical task. A straightforward approach might be to apply standard inferential statistics to demonstrate that a newly collected community or non-patient sample does not differ from the normative reference data being evaluated. However, this approach involves testing the null hypothesis, which poses statistical challenges and has historically created controversy (Altman & Bland, 1995). This, of course, gets even more complicated when comparing the degree of fit of a newly collected sample with *two* different sets of norms.

To address this methodological problem, we introduce in this article an adaptation of Rouder and colleagues' (Rouder & Morey, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009) statistical approach to measure evidence from data for competing positions. Specifically, we illustrate the use of Bayesian statistics to address normative questions, by testing whether the mean values produced by a small non-patient sample collected in San Diego, California, would more closely resemble the CS 600 (Exner, 2003) or the international (Meyer et al, 2007) normative values. Given the small sample size and other limitations associated with our sample, this work is only a demonstration study with the primary aim to

illustrate a new statistical methodology for evaluating the degree of fit between an independent sample and two different sets of norms.

Method

Participants

Volunteers were included in the sample if they: (a) were English-speaking; (b) had no history of psychiatric hospitalization; (c) were not currently in psychotherapy or counseling; (d) were not currently on any psychotropic medications prescribed by a psychiatrist; (e) were not currently abusing or dependent on drugs or alcohol, as outlined in the DSM-III criteria[2]; (f) had not been administered the Rorschach in the previous year.

Initially 98 adults living in San Diego, California volunteered for this study. Eventually, three were excluded because they reported being on psychotropic medications prescribed by a psychiatrist, nine because of incomplete administrations, and six because they had less than 14 responses on their Rorschach administration. Thus, the final sample for the study included 80 participants. Ages ranged from 21 to 79 years, with a mean age of 37.9 (*SD* = 15.2)[3], and around 59% of the sample were women (*n* = 47). Additional demographic information is reported in Table 1. No participant had been administered the Rorschach during the year before their participation in the study but 10 participants had taken it more than a year earlier, four for research purposes, three for student practice, and three for unknown purposes.

Because our sample was collected in the U.S., one might expect our San Diego sample to more closely resemble the American, CS 600 norms than the international sample dominated by countries outside the United States. Compared to the CS 600 sample, however, our sample was significantly older, $t(83.4) = 3.4$, $p < .01$, $d = .55$, more educated, $t(676) = 12.2$, $p < .01$, $d = 1.46$, and included a larger proportion of women, *phi* = .11, $p = .01$, and

[2] When this study was initiated, many clinicians were still using the DSM-III, despite the fact that DSM-IV had already been published.
[3] Five records were missing age information but are known to be adults.

Caucasians, *phi* = .09, $p = .02$. In contrast, no significant age differences emerged when comparing the San Diego sample with the international normative data, $t(75.4) = 0.8$, $p = .41$, $d = .12$.[4] In an attempt to control these variables, we explored their impact on the results in follow-up analyses.

Procedure

Participants were recruited through flyers, announcements, and word of mouth in San Diego. Research assistants screened volunteers with a uniform screening protocol addressing age, sex, history of psychiatric hospitalization, current medications, and whether or not they had a current problem with drugs or alcohol. Participants were told that (a) participation would be anonymous and voluntary, (b) they could terminate their participation at any time, and (c) they would not be compensated for their participation in the research.

Research assistants met potential participants individually on campus or in psychological clinics, homes of participants or administrators, private rooms, or public places, e.g., libraries. Upon meeting, participants were asked to read and sign an informed consent, and anonymity and the confidentiality of the records was explained at this point. Next, participants completed a demographic form, including questions regarding the exclusion criteria. Finally, they were administered the Rorschach, according to standard CS procedures (Exner, 2003). As previously noted, some Rorschach records were eventually excluded from the analysis because of incomplete records of administrations or less than 14 responses, and three records were eventually excluded because the respondents, in contrast from what they reported over the phone, admitted to current use of psychotropic medications.

Rorschach Administration and Scoring

Rorschachs were administered and scored according to the CS guidelines by seven advanced graduate students. They were aware that they were collecting a local non-patient

[4] For the t-tests, when homoscedasticity could not be assumed, Welch-Satterthwaite method was used to adjust degrees of freedom. Also, years of education and distribution of gender and race were not reported in Meyer et al. (2007).

sample, but were unaware of the purpose of the current study. All these individuals and others involved later in checking scores or providing independent reliability scoring were trained in CS techniques and had completed at least two courses involving Rorschach training. Administration procedures, in line with CS guidelines, included warm-up procedures aimed at establishing rapport and addressing potential factors affecting the quality of the administration. The administration and scoring was supervised by the second author, and any questions about both were discussed with him. As an additional scoring check, all scores were checked a second time by other graduate students, also blind to the purpose of the study, and any disagreements in scoring were then resolved by the second author. It should be pointed out that this scoring procedure was completely independent of the scoring procedure used to establish reliability.

From the 113 available CS variables, we selected the 28 "divergent variables," i.e., the 24.8% of the Rorschach variables for which the CS 600 and the international norms differed by at least a Cohen's *d* effect size of .5. As explained later in this paper, indeed, an assumption of the Bayesian approach that we adopted postulates that the two sets of norms do differ from each other. Accordingly, the 85 "non-divergent" variables (75.2%), which are essentially the same in the CS 600 and the international norms, are excluded from analysis. The choice of $d = .5$ as a cut-off score for divergent variables is consistent with Cohen's recommendations (1988), as well as with assessment literature, which characterizes a difference of 5 *T* points ($d = .5$) on the MMPI as a notable difference (e.g., Greene, 2000).

To establish inter-rater reliability, 20 records were randomly selected and scored by two raters blind to the initial coding. For these records, two-way random effects model, single measures intraclass correlation coefficients (ICCs) were calculated for the 28 CS variables included in the analysis (see below). All variables other than WDA% showed at least

adequate inter-rater reliability (see Table 2). Caution when interpreting results related to WDA% is warranted.

Data Analysis

We tested whether the CS 600 or the international norms provide a closer fit to our San Diego sample. To do so, we applied a Bayesian approach to each divergent variable by calculating the ratio of the probability of obtaining our data under the hypothesis that the San Diego sample is a sub-sample of the CS 600 normative population, to the probability of obtaining our data under the alternative hypothesis that the San Diego sample is a sub-sample of the international normative population. For the sake of readability, we label this odds ratio as *Odds Ratio CS 600 over Int'l*, which can be expressed by:

$$\text{Odds Ratio CS 600 over Int'l} = \frac{\Pr(\text{data} \mid H_0 \text{ CS 600})}{\Pr(\text{data} \mid H_0 \text{ Int'l})},$$

where *Pr (data | H_0 CS 600)* is the conditional probability of obtaining our data under the hypothesis that the San Diego sample is a sub-sample of the CS 600 norms, and *Pr (data | H_0 Int'l)* is the conditional probability of obtaining our data under the hypothesis that the San Diego sample is a sub-sample of the international norms.

This odds ratio has the advantage of being directly interpreted according to Jeffreys's (1961) thresholds: values greater than 3 indicate that there is "some evidence" for one hypothesis over another (i.e., one hypothesis is three times more probable than the competing one, odds are three to one); values greater than 10 indicate that there is "strong evidence" for one hypothesis over another (i.e., one hypothesis is ten times more probable than the competing one); and values greater than 30 indicate that there is "very strong evidence" for one hypothesis over another (i.e., one hypothesis is ten times more probable than the competing one). For example, if the *Odds Ratio CS 600 over Int'l* is equal to 3, then the hypothesis that the San Diego sample is a sub-sample of the CS 600 normative population is 3 times more probable than the hypothesis that the San Diego sample is a sub-sample of the

international normative population, given the data. According to Jeffreys's criteria, such a result would therefore be considered as "some evidence" that the CS 600 norms fit the data of the San Diego sample better than the international norms. Conversely, if the *Odds Ratio CS 600 over Int'l* is equal to .33 (i.e., 1/3), then the hypothesis that the San Diego sample is a sub-sample of the international normative population is 3 times more probable than the hypothesis that the San Diego sample is a sub-sample of the CS 600 normative population. Such a result would be considered as "some evidence" that the international norms fit the data of the San Diego sample better than the CS 600 norms.

Synthesizing this information, *Odds Ratio CS 600 over Int'l* values of 3, 10, or 30 indicate that the data provide, respectively, "some evidence," "strong evidence" or "very strong evidence" that the San Diego sample more closely fits with the CS 600 norms, while values of .33, .10, or .03 indicate that the data provide, respectively, "some evidence," "strong evidence" and "very strong evidence" that the San Diego sample more closely fits with the international norms.

The Bayesian Approach. Bayesian analyses are still uncommon in the psychological and Rorschach literature (although see Reese, Viglione, & Giromini, 2014). However, a number of statisticians have recently demonstrated that classic null-hypothesis significance tests (NHSTs) are biased toward rejection, in that they underestimate the support for the null hypothesis, and overstate the evidence against it (e.g., Berger & Sellke, 1987; Edwards, Lindman, & Savage, 1963; Goodman, 1999; Rouder & Morey, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Sellke, Bayarri, & Berger, 2001; Wagenmakers, 2007; Wagenmakers & Grünwald, 2006). In fact, when the null is false, increasing the sample size decreases the *p*-values (as one should expect), but when the null is true, increasing the sample size does not increase the evidence for the null hypothesis (Rouder et al., 2009). That is, the NHST approach does not allow a researcher to gain evidence for the null by increasing the

sample size. Also, with large samples (e.g., $N = 500$) and small effect sizes (e.g., around .2), the probability to obtain a small *p*-value (e.g., around .04 or .05) is very high, even though under similar circumstances the null is about 10 times more probable than the alternative (for details see, for example, Rouder & Morey, 2011). For all these reasons, especially when testing the null hypothesis (as it is the case when testing whether a sample comes from a given population) the Bayesian statistics, which provide a straightforward methodology for measuring evidence from data for competing positions, can be considered to be a more appropriate approach than the NHSTs (Rouder et al., 2009).

Computing the *Odds Ratio CS 600 over Int'l.* For each variable under investigation, to calculate the *Odds Ratio CS 600 over Int'l* we followed a three-step procedure. First, we calculated the ratio of the conditional probability of obtaining our data under the hypothesis that the San Diego sample is a sub-sample of the CS 600 norms, to the conditional probability of obtaining our data under the hypothesis that the San Diego sample is not a sub-sample of the CS 600 norms. These two hypotheses can be seen as the null and the alternative hypotheses of the classic one-sample *t*-test, where the null is that the means of the San Diego sample are equal to those of the CS 600 population, and the alternative is that they are different. In the Bayesian approach, such a ratio, i.e., $Pr\ (data \mid H_0) / Pr\ (data \mid H_1)$, is some times denoted by *B* and termed the *Bayes factor* (Jeffreys, 1961; Kass & Raftery, 1995). In our study, we label this ratio as *B CS 600.*

Computationally, to calculate the value of *B CS 600*, we adopted procedures described by Rouder et al. (2009) and utilized the web-based program provided by the authors. In this method, the *B* values are termed *JZS B* and calculated according to Rouder et al.'s (2009) equation 1 for the one-sample case[5]. As compared to other approaches for calculating *B, JZS*

[5] In calculating the *JZS B*, the experimenter has to define a scale factor related to prior probabilities, which is denoted by *r*. Rouder et al.'s (2009) recommend setting $r = 0.5$ in situations where small differences are important. Because small differences in sets of norms are likely to be interpretatively important we set *r* at 0.5 (for details, see Rouder et al. 2009).

B has a number of advantages: "It makes intuitive sense, it has beneficial theoretical properties, it is not dependent on the measurement scale of the dependent variable, and it can be conveniently computed" (Rouder & Morey, 2011, p. 685).

The second step of our three-step procedure to compute the *Odds Ratio CS 600 over Int'l* consisted of computing the *Bayes factor* for the comparison between the San Diego sample and the international norms. This second *Bayes factor*, which we label as *B Int'l*, was obtained using the same procedures adopted to calculate the *B CS 600* (for details, see Appendix A).

Finally, in the third step of our three-step procedure, we calculated the *Odds Ratio CS 600 over Int'l* according to the following formula:

$$\text{Odds Ratio CS 600 over Int'l} = \frac{\text{JZS B CS 600}}{(1+\text{JZS B CS 600})} * \frac{(1+\text{JZS B Int'l})}{\text{JZS B Int'l}}$$

where *JZS B CS 600* is the *Bayes factor* obtained in our first step (i.e., the *B CS 600*), and *JZS B Int'l* is the *Bayes factor* obtained in our second step (i.e., the *B Int'l*). The derivation of this formula is simple and straightforward, and is detailed in Appendix A. Important to our goal, this formula provides the ratio of the probability of obtaining our data under the hypothesis that the San Diego sample is a sub-sample of the CS 600 norms, to the probability of obtaining our data under the hypothesis that the San Diego sample is a sub-sample of the international sample.

When interpreting the *Odds Ratio CS 600 over Int'l*, however, one must keep in mind that the statistical procedure to produce this index assumes that *either* the San Diego sample is a sub-sample of the CS 600 norms *or* it is a sub-sample of the international norms. Although such an assumption might make sense in many circumstances and certainly is useful so as to make a decision about the degree of fit of one versus the other sets of norms, it might also be misleading, in some situations, and potentially violated in others. The most obvious violation of this assumption occurs when variable means are nearly the same in two

different sets of norms. To avoid violating this assumption and to focus our analysis on the variables that differed in the two sets of norms, we confined our analysis to the 28 "divergent variables."

Results

The 28 divergent variables are divided into interpretively less important and more important groups (Exner, 2003). In the CS Structural Summary sheet (Exner, 2003), these two groups of variables are separated. Interpretively important variables are found in the bottom half in the "Ratios, Percentages, and Derivations" section. For all 28 variables, Table 3 includes the mean and standard deviation of the San Diego sample, the *JZS B CS 600*, *JZS B Int'l*, and *Odds Ratio CS 600 over Int'l*, and the Cohen's d values corresponding to the differences between the San Diego sample and both the sets of norms.

Examination of Table 3 reveals that none of the 28 *Odds Ratio CS 600 over Int'l* is equal to or greater than 3. According to Jeffreys's thresholds, thus, for no variables do the CS 600 norms fit the data of the San Diego sample better than the international norms. Conversely, for 22 of the 28 variables under investigation, the *Odds Ratio CS 600 over Int'l* is lower than .03, which indicates "very strong evidence" that the international norms provide a closer fit. For two other variables, the *Odds Ratio CS 600 over Int'l* value is lower than .33 but greater than .10, thus indicating "some evidence" that the international norms provide a closer fit. For the remaining four variables, neither normative sample provides a better fit.

Cohen's *d* values also confirm the pattern that the international norms provide a better fit for the San Diego sample. The mean absolute *d* value for the difference between the San Diego data and the CS 600 norms is .96. The corresponding value is only .19 when the San Diego sample is contrasted to the international norms. For no variables does the difference between the San Diego sample and the international norms yields a Cohen's *d* equal to or

greater than .5. Conversely, 22 of such medium to large effect sizes emerge when comparing the San Diego data with the CS 600 norms.

Analyses of Possible Confounds

Participant Demographics. As noted earlier, the San Diego sample was older, more educated, and included a greater proportion of women and Caucasians than the CS 600 norms. To investigate the possible impact of these demographic variables on our results, we evaluated the relationship of these demographic variables with the Rorschach variables. We could not examine these variables in the normative data samples themselves because we did not have the individual participant's data. However, within the San Diego sample, the correlations of the 28 divergent variables under investigation with age, education, and gender (dummy variable) were negligible, low, or moderate, $|r| < .33$. Among the 84 tested correlations, only 4, i.e., about 5%, produced uncorrected p-values below .05, a proportion that is fully consistent with chance. In fact, no correlation approached significance after Bonferroni's correction. Because almost all of the San Diego participants were Caucasian Americans, it was not possible to explore the impact of ethnicity on the results, so that more research is needed on this topic.

Changes in CS FQ Coding Guidelines. As shown in Table 3, the international norms provided a much better fit to our San Diego sample for a large number of FQ related variables. The FQ coding guidelines, however, have evolved over time (Meyer & Archer, 2001), and the CS 600 norms may never have been rescored with the updated guidelines. In contrast, both the San Diego sample and the international norms have been collected with updated guidelines. Thus, one may speculate that the increased similarity between the San Diego sample and the international norms may be due to their sharing these revised FQ coding procedures.

To evaluate this possible confound, we selected the two key marker variables of FQ, i.e., X-% and X+%.[6] We tested them once again, this time substituting the set of 450 non-patient data described by Exner and Erdberg in 2005 (CS 450) for the CS 600 and repeated the Bayesian analyses. Since the CS 450 were collected using the updated FQ coding procedures, these additional analyses serve as a test of whether the updated coding procedures had an impact on the main results of the present study.

The results of these analyses are summarized in Table 4. As for X-%, the difference between the San Diego and CS 450 data yielded a notably greater effect size (Cohen's d = .82) than the comparison between the San Diego sample and the international norms (Cohen's d = .18). Similarly to the *Odds Ratio CS 600 over Int'l*, the *Odds Ratio CS 450 over Int'l* also is < .01, thus still providing "very strong evidence" that the San Diego sample has a greater degree of fit with the international norms than with the CS 450. A similar result was observed also when X+% was investigated. The Cohen's d for the difference between the San Diego sample and the CS 450 is 1.22, a notably greater value than the Cohen's d of .15 found when comparing the San Diego sample to the international norms. Again, the *Odds Ratio CS 450 over Int'l* was still less than .01, so that the San Diego sample more closely resembles the international norms.

All in all, it is very unlikely that the observed similarity between the San Diego sample and the international norms is due to FQ coding related changes over time.

Complexity. According to Ritzler and Sciara (2009), a major concern regarding the generalizability of the international norms is that most of its constituent studies used students as examiners. They speculated that these relatively inexperienced student examiners might reduce the overall complexity of the records. Because the San Diego sample also used students as examiners, it is possible that some of the convergence between the San Diego data

[6] XA% was also considered. It is essentially the complement of X-%, and in our sample was correlated with X-% at -.97, p < .01. Thus, it is redundant with X-% so that it was not included.

and the international norms – and some of the divergence between the San Diego data and the CS 600 norms – might be the result of a similar methodological limitation reducing complexity.

To explore possible confounds, we investigated the two marker variables for complexity identified by Ritzler and Sciara (2009), i.e., WSumC and Lambda. Results are reported in Table 5. As for WSumC, the Cohen's *d* for the difference between the San Diego and CS 600 data was -.36; the difference between the San Diego and international data was .28; and the *Odds Ratio CS 600 over Int'l* was .68 (i.e., neither set provided a better fit for the San Diego sample). Lambda has a mean of .60 (*SD* = .31) within the CS 600, .86 (*SD* = .95) within the international norms, and.58 (*SD* = .47) within the San Diego sample. The Cohen's *d* for the difference between the San Diego and CS 600 data was therefore .06, that for the difference between the San Diego and international data was .30, and the *Odds Ratio CS 600 over Int'l* was > 100 (i.e., very strong evidence that the CS 600 provided a better fit than the international norms for the San Diego sample). It should be pointed out, however, that the difference in Lambda between the international and CS 600 norms only yields a Cohen's *d* of .29, a relatively small value.

According to these findings, the level of complexity in the San Diego sample is by no means closer to that of the international norms than to that of the CS 600. In fact, all differences under investigation yielded small Cohen's *d* effect size values for both the variables under investigation, i.e., WSumC and Lambda. Furthermore, the Bayesian analysis indicated that the CS 600 norms provided a better fit for the San Diego sample than the international norms in respect to Lambda.

Discussion

This article offers a new methodological approach to investigate the degree of fit of two different sets of Rorschach norms with an independent sample. To illustrate this

statistical technique, we tested whether the international or CS norms (CS 600) would provide a better fit for a small, newly collected, U.S. community sample of 80 adults. Specifically, we adapted a Bayesian method to calculate the odds ratio of the probability of obtaining our data under the hypothesis that the San Diego sample is a sub-sample of the CS 600 norms, to the probability of obtaining our data under the hypothesis that the San Diego sample is a sub-sample of the international sample. Among the 28 divergent variables under investigation, the international norms provided a greater degree of fit for 24 of these variables. For the remaining four variables, neither normative sample provided a better fit. Taken together, thus, these findings indicate that our small sample more closely resembled the international norms than it did the CS 600 norms.

Previous research publications (Meyer et al, 2007; Shaffer, Erdberg, & Haroian, 1999; Viglione & Hilsenroth, 2001; Wood et al., 2001a, 2001b) argued that the CS norms are problematic for some variables and that using the CS 600 as a benchmark might pathologize interpretations. Somewhat in line with this position, with our small sample the *JZS B CS 600* and related Cohen's *d* values for Form Quality (X+%, FQxo, XA%, X–%, WDA%, Xu%), human representations (Poor HR, MQo, MQ–), and Unusual Location responses (Dd) depart most dramatically from CS expectations (see Table 3, 4th and 5th columns).

Our findings, however, do not support the conclusion that the international norms perfectly fit the data produced in our San Diego sample. In fact, according to the *JZS B Int'l* values found in the sixth column of Table 3, there are seven variables (out of 28) for which there is at least some evidence that a difference between our samples and the International norms exists. For these variables, the mean absolute *d* value of their discrepancy from the international expectations is .30. Three of them (i.e., Bt, MQ+, and Sum Color) are of secondary interpretive importance, whereas four variables are among the interpretively more

important CS Ratios, Percentages, and Derivations (Exner, 2003). Specifically, they are three color and human movement related variables (i.e., EA, WSumC, and FC), and XA%.

When considering these seven variables, for which the international norms do not perfectly fit the data of the San Diego sample, it is important to appreciate the real meaning of the *Odds Ratio CS 600 over Int'l*. As shown in Table 3, indeed, the *Odds Ratio CS 600 over Int'l* value for Bt, MQ+, and XA% is lower than .01, thus indicating that there is "very strong evidence" from the data, that the international norms provide a greater degree of fit than do the CS 600. Though this is true, in a relative way, one should also acknowledge that, in fact, neither normative sample provided a perfect fit for the San Diego sample.

As stated above, given the small sample size and other limitations associated with our sample, this work is only a demonstration study. Our primary aim was to illustrate a new statistical methodology to compare two sets of norms, in this case, Rorschach norms. Our main focus, more precisely, was on the procedure involved in such a challenging statistical task, rather than on our very limited, observed data. By elaborating Rouder et al.'s (2009) guidelines to compute the *JZS B* values for the one-sample t-test case, we developed and demonstrated a new methodological approach to measure and to judge evidence from data for two competing hypotheses.

As noted in the Method section of this paper, unlike the classic null-hypothesis significance tests (NHSTs), Bayesian statistics are not biased toward rejection of the null, and they allow researchers to compare evidence from data for *two* competing hypotheses (i.e., that the null is to be *rejected* versus *accepted*). Relative to NHSTs, thus, this method of comparison is most advantageous with large samples, and applicable to evaluating normative data because these samples are likely large. Noteworthy, the Bayesian approach we introduce in this article offers an important advantage also over other, less complex approaches that only use simple comparisons based on the Cohen's *d* values. One might contend, for

example, that comparing the Cohen's d values for the fit of each variable with two normative reference groups would work just as well as our Bayesian approach. However, because the effect size is independent from the sample size, such an approach would, in fact, overestimate the value of a Cohen's *d* obtained with a small sample size, and underestimate that of a *d* observed with a large sample size. For example, a Cohen's *d* of 1.0 obtained with a very small sample size (e.g., n = 10) would lead to the exact same conclusions as a Cohen's *d* of 1.0 obtained with a much larger sample size (e.g., of thousands of subjects). This behavior is, evidently, undesirable. Unlike the Cohen's *d*, the odds ratio index introduced in this article is based on the *JZS B* (Rouder et al., 2009), and, therefore, it has similar theoretical properties: the greater the sample size, the greater the weight of the evidence from the data.

Limitations and Final Considerations

By no means does this study provide conclusive evidence of the superiority of one set of Rorschach norms over another. Rather, this study's aim was to introduce and to illustrate a "user-friendly," handy solution to evaluate evidence for the degree of fit of an independent sample with two different sets of norms. Indeed, all of these findings may be idiosyncratic to our sample. For instance, they may result from peculiarities in the administration procedures (e.g., some records were collected in homes of participants or administrators, or in public places such as libraries, which deviates from standard administration procedures), or reflect "local" characteristics associated with American culture and development within the U.S. society.

In addition, our participants were relatively old and included a great proportion of Caucasians. We examined the associations with our sample for age and found no support that this variable accounted for our findings. However, without access to the individual participant's data in the two normative samples, we could not test whether the Caucasian

proportion in our sample affected the findings. Thus, racial differences may have influenced our results.

Because the FQ coding guidelines have evolved over time, in our section titled "Analyses of Possible Confounds", we tested whether the increased similarity between the San Diego sample and the international norms would hold true also when considering the more recent, CS 450 norms (Exner & Erdberg, 2005). Although our results indicated that the international norms continued to provide a better fit for X-% and X+% when considering the CS 450 sample, future studies should further investigate the extent to which the international versus the CS 450 norms provide a greater degree of fit for independent samples, perhaps also by considering a wider range of Rorschach variables.

We also tested whether the convergence between the San Diego data and the international norms might be attributable to the use of students as examiners. Specifically, we explored whether the general level of complexity among the San Diego records was lowered by the fact that we used graduate students as examiners. Contrary to our concerns, our results indicated that by no means was the level of complexity in the San Diego sample closer to that of the international norms than to that of the CS 600. In fact, as compared to the international norms, the CS 600 norms provided even a better fit for our San Diego sample, when the Lambda variable was taken into consideration. The degree to which the use of student versus more experienced examiners may influence other variables, however, is presently unknown, and future studies should address this potential limitation.

Lastly, some of our inter-rater reliability values were relatively low, compared to other data reported in the literature (e.g., Meyer et al., 2002; Viglione, Blume-Marcovici, Miller, Giromini, & Meyer, 2012; Viglione & Taylor, 2003). Though both the original and re-scored records were coded by (graduate) students, the initial coding was also carefully supervised by a senior clinician and researcher who has been using the Rorschach for years.

Given that, it is unlikely that coding issues may account for our findings. Nevertheless, caution is warranted when interpreting results related to variables with low inter-rater reliability.

Despite all these limitations, this study offers some initial information concerning the degree of fit of the international and CS 600 norms for a small, independent, U.S. sample, and, most importantly, it introduces a new, statistical approach to evaluate which norms, between two different sets, would provide a greater degree of fit to a given sample. We anticipate that this new approach could be used also for other purposes, in addition to evaluating different sets of Rorschach norms. For example, it could be adapted to investigate whether given test norms from a specific non-U.S. country (e.g., Italy) provide a better fit for a sub-group of immigrants to the U.S. from that specific country (e.g., Italian-Americans). Similarly, one may want to use this statistical approach to investigate whether a specific sample of adolescents more closely resembles the normative reference data for children or adults.

References

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal, 311*, 485.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association, 82*, 112–122.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193–242.

Exner, J. E. (2003). *The Rorschach: A comprehensive system, Volume 1* (4th ed.). New York: Wiley.

Exner, J. E. (with Colligan, S. C., Hillman, L. B., Metts, A. S., Ritzler, B., Rogers, K. T., Sciara, A., D., & Viglione, D. J.) (2001). *A Rorschach workbook for the Comprehensive System* (5th ed.). Asheville, NC: Rorschach Workshops.

Exner, J. E., & Erdberg, P. (2005). *The Rorschach: A Comprehensive System, Volume 2: Advanced Interpretation* (3rd ed.). Oxford: Wiley.

Exner, J. E., & Weiner, I. B. (2003). *RIAP 5 – Rorschach Interpretation Assistance Program: Version 5 for Windows*. Odessa, FL: PAR Psychological Assesment.

Goodman, S. N. (1999). Toward evidence-based medical statistics: I. The p value fallacy. *Annals of Internal Medicine, 130*, 995-1004.

Greene, R. L. (2000). *The MMPI-2: An interpretive manual (2nd ed.)*. Boston: Allyn & Bacon.

Jeffreys, H. (1961). *Theory of probability (3rd ed.)*. New York: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

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

Meyer, G. J., Erdberg, P., Shaffer, T. W. (2007). Toward international normative reference data for the Comprehensive System. *Journal of Personality Assessment, 89(S1)*, S201-S216.

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

Meyer, G. J., Shaffer, T. W., Erdberg, P., Viglione, D. J., & Mihura, J. L. (2009). *Issues in the Development and Use of International Norms for Adults*. Paper presented at the European Rorschach Association conference, Prague, Czech Republic.

Meyer, G. J., Viglione, D. J., Mihura, J. L., Erard, R. E., & Erdberg, P. (2011). *Rorschach Performance Assessment System: Administration, coding, interpretation and technical manual*. Toledo, OH: Rorschach Performance Assessment System.

Reese, J. B., Viglione, D. J., & Giromini, L. (2014). A Comparison Between Comprehensive System and an Early Version of the Rorschach Performance Assessment System Administration with Outpatient Children and Adolescents. *Journal of Personality Assessment*, in press.

Ritzler, B., & Sciara, A. (2009). Rorschach Comprehensive System international norms: Cautionary notes. Unpublished manuscript. Retrieved from

http://www.rorschachtraining.com/wp-content/uploads/2011/10/Rorschach-Comprehensive-System-International-Norms.pdf

Rorschach, H., 1921. *Psychodiagnostik*. Bern, Bircher.

Rouder, J. N., & Morey, R. D. (2001). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*, 682–289.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician, 55*, 62-71.

Shaffer, T. W., Erdberg, P., & Haroian, J. (1999). Current nonpatient data for the Rorschach, WAIS-R, and MMPI-2. *Journal of Personality Assessment, 73*(2), 305-316.

Shrout, P. E., & Fliess, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Viglione, D., Blume-Marcovici, A. C., Miller, H., L., Giromini, L., & Meyer, G. (2012). An Inter-Rater Reliability Study for the Rorschach Performance Assessment System. *Journal of Personality Assessment, 94(6)*, 607-612, DOI: 10.1080/00223891.2012.684118

Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions, and future. *Psychological Assessment, 13*(4), 452–471.

Viglione, D. J., & Meyer G. J (2008). An overview of Rorschach psychometrics for forensic practice. In C. Gacono, F. Evans, N. Kaser-Boyd, & L. Gacono (Eds.) *The Handbook of Forensic Rorschach Assessment* (pp. 21-53). New York, NY: Routledge/Taylor & Francis Group.

Viglione, D. J., & Taylor, N. (2003). Empirical support for interrater reliability of the Rorschach Comprehensive System coding. *Journal of Clinical Psychology, 59(1)*, 111–121.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin &Review, 14*, 779–804.

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science, 17*, 641-642.

Wood, J. M., Nezworski, M. T., Garb, H. N., & Lilienfeld, S. O. (2001a). The misperception of psychopathology: Problems with norms of the Comprehensive System for the Rorschach. *Clinical Psychology: Science & Practice, 8*(3), 350–373.

Wood, J. M., Nezworski, M. T., Garb, H. N., & Lilienfeld, S. O. (2001b). Problems with the norms of the Comprehensive System for the Rorschach: Methodological and conceptual considerations. *Clinical Psychology: Science & Practice, 8*(3), 397–402.

Table 1

Composition of the San Diego Sample (N = 80).

Variable	Descriptive Statistics M (SD) / n (%)
Age[a]	37.9 (15.2)
Education	15.0 (1.7)
Gender	
Women	47 (59%)
Men	33 (41%)
Ethnicity	
Caucasian	73 (91%)
African American	1 (1%)
Hispanic American	1 (1%)
Asian American	1 (1%)
Not indicated	4 (5%)
Employment Status	
Employed	57 (71%)
Unemployed	12 (15%)
Retired	7 (9%)
Not indicated	4 (5%)
Marital Status	
Married	43 (54%)
Divorced	10 (13%)
Separated	3 (4%)
Single	19 (24%)
Widowed	4 (5%)
Not indicated	1 (1%)

[a] Five records were missing age information but are known to be adults

Table 2

Inter-rater Reliability of the 28 Selected Rorschach Variables.

Variables with excellent inter-rater reliability (ICC ≥ .75)	FQx+; MQ+; Afr; D; Sum Color; EA; WSumC; AG; Dd; Bt; MQo; MQu; FC; SQual–; FQxo; Poor HR; Good HR; FQxu; COP; X+%.
Variables with good inter-rater reliability (.60 ≤ ICC < .75)	Populars; Xu%; CF; MQ–.
Variables with fair inter-rater reliability (.40 ≤ ICC < .60)	FQx–; X–%; XA%.
Variables with poor inter-rater reliability (ICC < .40)	WDA%.

Notes. N = 20. ICC = intraclass correlation. The characterization of the ranges of the reliability coefficients is derived from Cicchetti (1994) and Shrout and Fliess (1979). Interested readers may contact the corresponding author for more details about the exact ICC of each variable.

Table 3

Degree of fit of the CS 600 vs. International Norms for the San Diego Sample: JZS B, Cohen's d, and Odds Ratio Values

Divergent Variables	San Diego Sample Data (n = 80)		CS 600 Sample (n = 600)		Int'l Sample (n = 4704)		San Diego Sample vs. CS 600 Norms		San Diego Sample vs. Int'l Norms		Odds Ratio CS 600 over Int'l[c]
	Mean	SD	Mean	SD	Mean	SD	*JZS B*	Cohen's *d*[b]	*JZS B*	Cohen's *d*[b]	
Interpretively Less Important											
Bt	1.09	1.08	2.37	1.32	1.41	1.44	< 0.01	-0.99	0.24	-0.22	< 0.01
CF	1.95	1.47	2.41	1.31	1.65	1.55	0.17	-0.35	1.27	0.19	0.26
FQx–	3.96	2.62	1.56	1.20	4.43	3.23	< 0.01	1.67	1.81	-0.15	< 0.01
FQx+	0.11	0.50	0.71	0.88	0.21	0.68	< 0.01	-0.71	1.49	-0.15	< 0.01
FQxo	11.84	4.22	16.44	3.34	11.11	3.74	< 0.01	-1.33	1.97	0.19	< 0.01
FQxu	6.55	3.92	3.49	2.03	6.20	3.93	< 0.01	1.31	4.36	0.09	< 0.01
MQ+	0.04	0.25	0.44	0.68	0.12	0.43	< 0.01	-0.62	0.12	-0.19	< 0.01
MQo	2.51	1.62	3.57	1.84	2.26	1.66	< 0.01	-0.58	2.41	0.15	< 0.01
MQu	1.08	1.46	0.21	0.51	0.69	0.99	< 0.01	1.26	0.48	0.39	< 0.01
Sum Color	4.96	2.22	6.09	2.44	3.91	2.53	< 0.01	-0.47	< 0.01	0.42	0.35
Interpretively More Important											
Dd[a]	2.90	2.82	1.16	1.67	3.33	3.37	< 0.01	0.94	2.50	-0.13	< 0.01
MQ–[a]	0.49	0.87	0.07	0.27	0.63	1.05	< 0.01	1.08	2.20	-0.13	< 0.01
SQual–[a]	0.99	1.28	0.25	0.56	0.87	1.15	< 0.01	1.08	4.29	0.10	< 0.01
Afr	0.55	0.21	0.67	0.16	0.53	0.20	< 0.01	-0.72	3.59	0.10	< 0.01
AG	0.75	1.07	1.11	1.15	0.54	0.86	0.11	-0.32	1.44	0.24	0.17

COP	1.31	1.20	2.00	1.38	1.07	1.18	< 0.01	-0.51	1.31	0.20	< 0.01
D	9.93	6.10	12.88	3.77	9.89	5.81	< 0.01	-0.72	5.86	0.01	< 0.01
EA	7.83	3.28	8.66	2.38	6.84	3.76	0.58	-0.33	0.22	0.26	2.04
FC	2.75	1.80	3.56	1.88	1.91	1.70	0.01	-0.43	< 0.01	0.49	1.62
Good HR	3.75	2.13	4.93	1.78	3.70	2.18	< 0.01	-0.65	5.74	0.02	< 0.01
Poor HR	2.93	2.63	1.53	1.46	2.86	2.52	< 0.01	0.85	5.73	0.03	< 0.01
Populars	5.75	1.93	6.58	1.39	5.36	1.84	0.01	-0.57	1.31	0.21	0.02
WDA%	0.84	0.09	0.94	0.06	0.82	0.11	< 0.01	-1.56	1.21	0.18	< 0.01
WSumC	3.72	1.81	4.36	1.78	3.11	2.17	0.07	-0.36	0.11	0.28	0.68
X–%	0.17	0.09	0.07	0.05	0.19	0.11	< 0.01	1.78	1.14	-0.18	< 0.01
X+%	0.54	0.14	0.77	0.09	0.52	0.13	< 0.01	-2.37	2.75	0.15	< 0.01
XA%	0.82	0.10	0.92	0.06	0.79	0.11	< 0.01	-1.52	0.18	0.27	< 0.01
Xu%	0.28	0.12	0.15	0.07	0.27	0.11	< 0.01	1.68	4.26	0.09	< 0.01

[a] These variables should not be included in most parametric analyses, according to Exner (2004). Thus, our analyses are likely to be less accurate with these variables. [b] Positive Cohen's *d* values indicate higher means in the San Diego sample; negative Cohen's *d* values indicate higher means in the norms. [c] Odds Ratio CS 600 over Int'l = Pr (data | CS 600 H_0) / Pr (data | Int'l H_0); see Appendix A for details.

Table 4

Degree of fit of the CS 450 vs. International Norms for the Marker Variables of FQ: JZS B, Cohen's d, and Odds Ratio Values.

Variable	San Diego Sample Data (n = 80)		CS 450 Sample (n = 450)		Int'l Sample (n = 4704)		San Diego Sample vs. CS 450 Norms		San Diego Sample vs. Int'l Norms		Odds Ratio CS 450 over Int'l[b]
	Mean	SD	Mean	SD	Mean	SD	*JZS B*	Cohen's *d*[a]	*JZS B*	Cohen's *d*[a]	
X–%	0.17	0.09	.11	.07	0.19	0.11	< 0.01	.82	1.13	-0.18	< 0.01
X+%	0.54	0.14	.68	.11	0.52	0.13	< 0.01	1.22	2.75	0.15	< 0.01

[a] Positive Cohen's *d* values indicate higher means in the San Diego sample; negative Cohen's *d* values indicate higher means in the norms. [b]

Odds Ratio CS 450 over Int'l = Pr (data | CS 450 H_0) / Pr (data | Int'l H_0); see Appendix A for details.

Table 5

Degree of fit of the CS 600 vs. International Norms for the Marker Variables of Complexity: JZS B, Cohen's d, and Odds Ratio Values.

Variable	San Diego Sample Data (n = 80)		CS 600 Sample (n = 600)		Int'l Sample (n = 4704)		San Diego Sample vs. CS 600 Norms		San Diego Sample vs. Int'l Norms		Odds Ratio CS 600 over Int'l[b]
	Mean	SD	Mean	SD	Mean	SD	*JZS B*	Cohen's *d*[a]	*JZS B*	Cohen's *d*[a]	
WSumC	3.72	1.81	4.36	1.78	3.11	2.17	0.07	-0.36	0.11	0.28	0.68
Lambda	0.58	0.47	.60	.31	.86	.95	5.29	-0.06	< 0.01	-0.30	> 100

[a] Positive Cohen's *d* values indicate higher means in the San Diego sample; negative Cohen's *d* values indicate higher means in the norms. [b]

Odds Ratio CS 600 over Int'l = Pr (data | CS 600 H_0) / Pr (data | Int'l H_0); see Appendix A for details.

Appendix A: Statistical Development of the ODDS Ratio CS 600 over International

For each variable under investigation, we used the JZS B statistics (Rouder et al., 2009) to test the null hypothesis that the San Diego sample produces a similar mean to that of the CS 600 norms. This statistic can be expressed as follows:

$$\text{JZS B CS 600} = \frac{\Pr(\text{data} \mid \text{H}_0\ \text{CS 600})}{\Pr(\text{data} \mid \text{H}_1\ \text{CS 600})},$$

where *H_0 CS 600* denotes the null hypothesis that the mean value of the San Diego sample is equal to that of the CS 600 norms, and *H_1 CS 600* denotes the alternative hypothesis that the mean value of the San Diego sample is different to that of the CS 600 norms. Of course, because either the null is true (and the alternative is false) or the null is false (and the alternative is true),

$$\Pr(\text{data} \mid \text{H}_0\ \text{CS 600}) + \Pr(\text{data} \mid \text{H}_1\ \text{CS 600}) = 1.$$

Consequently,

$$\text{JZS B CS 600} = \frac{\Pr(\text{data} \mid \text{H}_0\ \text{CS 600})}{1 - \Pr(\text{data} \mid \text{H}_0\ \text{CS 600})},$$

so that

$$\Pr(\text{data} \mid H_0\ \text{CS}\ 600) = \text{JZS B CS}\ 600 * (1 - \Pr(\text{data} \mid H_0\ \text{CS}\ 600)),$$

$$\Pr(\text{data} \mid H_0\ \text{CS}\ 600) = \text{JZS B CS}\ 600 - \text{JZS B CS}\ 600 * (\Pr(\text{data} \mid H_0\ \text{CS}\ 600)),$$

$$\Pr(\text{data} \mid H_0\ \text{CS}\ 600) + \text{JZS B CS}\ 600 * (\Pr(\text{data} \mid H_0\ \text{CS}\ 600)) = \text{JZS B CS}\ 600,$$

$$\Pr(\text{data} \mid H_0\ \text{CS}\ 600) * (1 + \text{JZS B CS}\ 600) = \text{JZS B CS}\ 600,$$

and, finally,

$$\Pr(\text{data} \mid H_0\ \text{CS}\ 600) = \frac{\text{JZS B CS}\ 600}{(1 + \text{JZS B CS}\ 600)}.$$

For each variable under investigation, we also tested the null hypothesis that the San Diego sample produces a similar mean to that of the international norms. Again, we used the JZS B statistic, so that

$$\text{JZS B Int'l} = \frac{\Pr(\text{data} \mid H_0\ \text{Int'l})}{\Pr(\text{data} \mid H_1\ \text{Int'l})},$$

where *H_0 Int'l* denotes the null hypothesis that the mean value of the San Diego sample is equal to that of the international norms, and *H_1 Int'l* denotes the alternative hypothesis that the mean value of the San Diego sample is different to that of the international norms. Again, of course,

$$\Pr(\text{data} \mid H_0 \text{ Int'l}) + \Pr(\text{data} \mid H_1 \text{ Int'l}) = 1.$$

Also, by adopting a similar approach to that described above, it can be easily demonstrated that

$$\Pr(\text{data} \mid H_0 \text{ Int'l}) = \frac{\text{JZS B Int'l}}{(1 + \text{JZS B Int'l})}.$$

Finally, to test whether the CS 600 or the international norms would provide a better fit for our San Diego sample, for each variable we calculated the following ratio:

$$\text{Odds Ratio CS 600 over Int'l} = \frac{\Pr(\text{data} \mid H_0 \text{ CS 600})}{\Pr(\text{data} \mid H_0 \text{ Int'l})}.$$

This measure, indeed, indicates the ratio between the probability of obtaining our results (i.e., data) under the hypothesis that the CS 600 norms are appropriate for our sample, to the probability of obtaining our results (i.e., data) under the hypothesis that the international norms are appropriate for our sample. On the basis of what was previously reported, the *Odds Ratio CS 600 over Int'l* can be expressed as follows:

$$\text{Odds Ratio CS 600 over Int'l} = \frac{\text{JZS B CS 600}}{(1 + \text{JZS B CS 600})} * \frac{(1 + \text{JZS B Int'l})}{\text{JZS B Int'l}}.$$