Behav Res (2011) 43:710–719
DOI 10.3758/s13428-011-0079-7

Robust regression for single-case data analysis: How can it help?

Daniel F. Brossart · Richard I. Parker · Linda G. Castillo

Published online: 25 March 2011

Abstract This study examined the degree to which outliers were present in a convenience sample of published single-case research. Using a procedure for analyzing single-case data Allison & Gorman (*Behaviour Research and Therapy, 31*, 621–631, 1993), this study compared the effect of outliers using ordinary least squares (OLS) regression to a robust regression method and attempted to answer four questions: (1) To what degree does outlier detection vary from OLS to robust regression? (2) How much do effect sizes differ from OLS to robust regression? (3) Are the differences produced by robust regression in more or less agreement with visual judgments of treatment effectiveness? (4) What is a typical range of effect sizes for robust regression versus OLS regression for data from "effective interventions"? Results suggest that outliers are common in single-case data. The effects of outliers in single-case data are explored, and the implications for researchers and practitioners using single-case designs are discussed.

Keywords Single case · Robust regression · Visual analysis · Outliers · Effect size

Historically, statistical analysis has been infrequently used in single-case research (SCR), with researchers typically preferring visual analysis. In part, this was due to the strong roots SCR has in applied behavior analysis, which relies on visual analysis to detect large changes. Given that there was a strong tradition of visual analysis across several decades (see Busk & Marascuilo, 1992; Kratochwill & Brody, 1978; Parker & Brossart, 2003), it is somewhat understandable that the practice of using visual analysis as the sole means of summarizing SCR data continued in spite of a number of studies that documented the unreliability of visual analysis, showing low-to-moderate interrater reliabilities, in the range of .40–.60 (DeProspero & Cohen, 1979; Harbst, Ottenbacher, & Harris, 1991; Ottenbacher, 1990; Park, Marascuilo, & Gaylord-Ross, 1990). Brossart, Parker, Olson, and Mahadevan (2006) designed a study of visual analysis that avoided most of the design limitations of earlier research, but they still obtained an average individual rater-to-group correlation of .58, a level similar to that in earlier research. Such findings suggest that supplementing visual analysis with statistical analysis should be standard practice.

Changes in the research climate have likely necessitated a move toward incorporating statistical analysis. The current climate values the documentation of treatment effects to meet expectations for accountability and to provide objective evidence for funding agencies. This trend toward objectively measured outcomes and greater scientific rigor can be seen in a wide range of published research (e.g., Kaplan & Groessl, 2002; Newnham & Page, 2010) and policy statements by influential groups such as the National Research Council (Shavelson & Towne, 2002). The call for empirically supported treatments, evidence-based practice (McHugh & Barlow, 2010) and the growing importance of meta-analysis also contribute to the need for statistical analysis in SCR.

In addition, there appears to be an increasing awareness that SCR designs have an important role in developing the evidentiary foundation of many domains, such as behavioral, psychological, rehabilitation, and educational research. This may be due in part to the contemporary

D. F. Brossart (✉) · R. I. Parker · L. G. Castillo
Department of Educational Psychology,
Texas A&M University,
College Station, TX 77843–4225, USA
e-mail: brossart@tamu.edu

dialogue regarding the role of randomized clinical trials (RCTs) in terms of their strengths and important limitations (e.g., Tucker & Reed, 2008; Tucker & Roth, 2006). Some have criticized the push for making the RCT the standard for evaluating psychological treatment as premature and as promoting a tendency toward methodcentric reasoning —"a form of cognitive myopia that leads psychologists to judge their preferred research methodology superior to all others" (Blais & Hilsenroth, 2006, p. 31).

Although SCR designs are often more feasible to conduct than RCTs (e.g., Morgan & Morgan, 2001), the increasing attention SCR is receiving may be due to an upsurge in researchers' awareness that single-case designs can be among "the most effective and powerful" (Shadish, Cook, & Campbell, 2002, p. 171) nonrandomized experimental designs (Shadish, Rindskopf, & Hedges, 2008). With increased use of SCR designs comes the need for researchers to continue to evaluate the performance of statistical techniques for single-case data (e.g., Brossart, Meythaler, Parker, McNamara, & Elliott, 2008; Brossart et al., 2006; Parker & Brossart, 2003; Parker, Cryer, & Byrns, 2006; Parker, Hagan-Burke, & Vannest, 2007).

As was noted by Parker and Brossart (2003), the number of statistical analytic techniques available has tripled since the early 1980s, but little information is available on how these techniques typically perform: their typical effect sizes, their dependability (confidence intervals), and how well they handle atypical data sets. Currently, regression models such as those presented by Center, Skiba, and Casey (1985–1986) and Allison and colleagues (Allison & Gorman, 1993; Faith, Allison, & Gorman, 1996) appear to be among the more promising methods available, although not without limitations (Faith et al., 1996; Parker & Brossart, 2003). Among the strengths of these regression models is their adequate power for short data series, their ability to control for baseline trend, and their ability to address both change in level and change in trend.

The need for robust methods

All of the regression-based techniques noted above use ordinary least squares (OLS) regression, which has important limitations that are too often overlooked. These limitations include the following: (1) Small departures from normality produce low power; (2) even with normal distributions, heteroscedasticity can markedly lower power; (3) with small departures from normality, typical confidence intervals and measures of effect size can be very inaccurate; (4) OLS is not an effective method for ascertaining and examining outliers (Wilcox, 1998a, b); and (5) outlier data points can produce unstable results in multiple regression methods (Hutcheson & Sofroniou, 1999). Yet, in applied work, it is very common for data to show skewness, outliers, unequal variance along the score distribution, and heavy-tailed distributions (Tukey, 1960; Wilcox, 1998a). Single-case researchers often find these undesirable characteristics in their time series data. Furthermore, single-case data sets often contain one or more outliers. Thus, one statistician has concluded that the OLS estimator may be "one of the poorest choices researchers could make" (Wilcox, 1998b, p. 311).

A common strategy to deal with outliers has been to delete them. This makes some sense, because detecting and removing outliers provides a way to reduce heteroscedasticity, but it also results in using the wrong standard error and can lead to low power. Unfortunately, the presence of outliers can result in a failure to detect all of the outliers present in a data set (Wilcox & Keselman, 2004). This problem is called *masking,* and it affects many of our traditional statistical methods (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986). Traditional methods of detecting outliers may also lead to some points labeled as outliers by chance and can result in using the incorrect standard error (Wilcox & Keselman, 2004).

Robustness historically referred to the problem of controlling for Type I error when testing hypotheses. Population parameters were said to be robust if slight changes in the distribution did not have an arbitrarily large effect on their value. Today, this definition has been expanded; suffice it to say, a major goal of robust estimators is to avoid having a few aberrant data points dominate or overly influence one's statistical results. Alternatively, robustness "refers to the ability of a statistic to be insensitive to small deviations from statistical assumptions" (Anderson & Schumacker, 2003, p. 84).

This article demonstrates the use of robust regression as a method for controlling outlier data in SCR; however, our primary task is not to show that robust methods are superior to standard OLS, because, as Wilcox (1998b) has noted, hundreds of articles have shown the deficiencies of standard methods and summaries have been presented in numerous books (e.g., Birkes & Dodge, 1993; Hampel et al., 1986; Hoaglin, Mosteller, & Tukey, 1983, 1985; Wilcox, 1996). In spite of this large literature base, it appears that misconceptions about robust methods persist (Wilcox, 1998a). While there is a preponderance of evidence that robust regression is an improvement over traditional methods, robust methods have yet to be demonstrated with single-case time series data.

To effectively use a statistical technique, the user must have a sense of how the technique performs with real data. Our goal in this article is to help the applied researcher gain a better understanding of how robust regression works with single-case data. We compare applications of OLS regression and robust regression on a convenience sample of 61 single-case data sets from published studies. We also

compare OLS and robust regression against visual analysis by expert single-case researchers.

There are numerous types of robust regression. For example, four common methods include least median of squares (Rousseeuw, 1984), least-trimmed squares (Rousseeuw, 1984), biweight midregression (Wilcox, 1997), and an MM method supported in S+ (TIBCO Software Inc., 2008a). One of the better "all purpose" methods used in this study is the MM regression method supported in current versions of S+ (Anderson & Schumacker, 2003; TIBCO Software Inc., 2008a). It should be noted that studies continue to be conducted, and some investigators suggest there are other robust regression methods that, in some instances, outperform the technique illustrated here (Wilcox & Keselman, 2004). The important point is that while robust methods usually outperform OLS, as was noted by Wilcox (2005), "it seems to be easy to find fault with any estimator that has been proposed" (p. 461). Thus, the robust method used here (if using R, the function MMreg applies the MM estimator used here) should not be viewed as the best or only robust estimator one should consider, because, in any given situation, one particular robust method may be more appropriate than another.

An example of how OLS and robust regression perform differently is illustrated in Fig. 1. In this figure, the data points are not connected between phases, as is typical in graphs of single-case data. Instead, the data points are represented by circles in the baseline phase and triangles in the treatment phase. The lines represent an OLS and robust regression line fit to each phase. There is no difference in the treatment phase, but in the baseline phase, there is a notable difference between the two regression methods. OLS gives far more importance to three data points from sessions 1, 9, and 10. The robust method is less susceptible to being influenced by extreme values, so the slope is not shifted clockwise, as it is with OLS. To confirm that these three data points are outliers, one may create a q-q plot, which tests the assumption that a model's errors are normally distributed. Figure 2 shows two q-q plots when the baseline phase (phase A) is examined: The one on the left is based on OLS, and the one on the right is based on a robust regression method. Points lying outside of the dotted lines above and below the solid regression line suggest nonnormality in the residuals. OLS is unable to detect any outliers (the masking limitation) and suggests that the residuals are normally distributed, whereas the robust method indicates nonnormality in the residuals and indicates three data points as outliers. This example demonstrates that robust methods are not as influenced by problematic variability and that they may be able to detect outliers that traditional methods miss.

As was noted by Wilcox (1998b), there is no need to remain plagued by the limitations of OLS regression when there are modern robust regression methods readily available. While using robust regression methods appears to address some of the limitations of OLS regression, there are many unanswered questions regarding the use of robust methods in SCR. This article will attempt to answer some of the initial questions a single-case researcher may have regarding the use of a robust regression method: (1) To what degree does outlier detection vary from OLS to robust regression? (2) How much do effect sizes differ from OLS to robust regression? (3) Are the differences produced by robust regression in more or less agreement with visual judgments of treatment effectiveness? (4) What is a typical

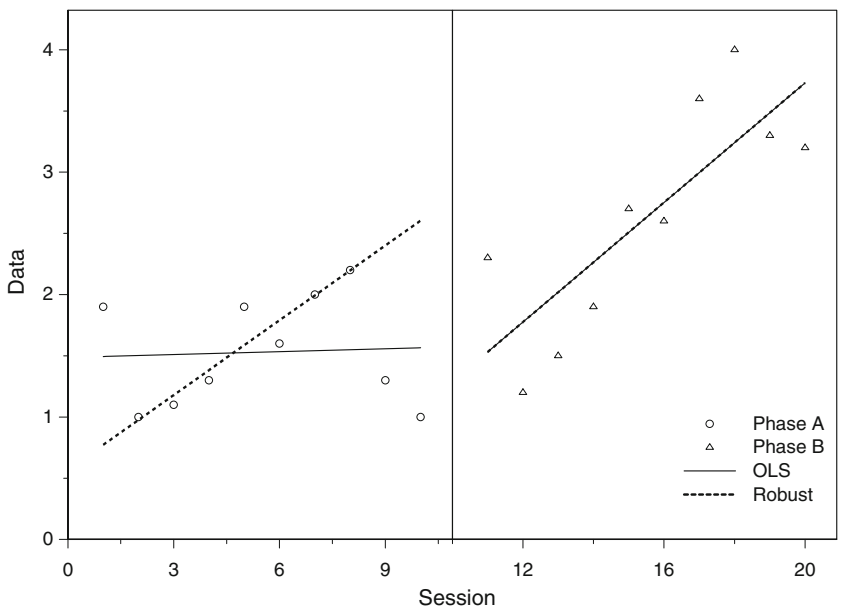

Fig. 1 Comparison of phases A and B for OLS and robust regression (MM)

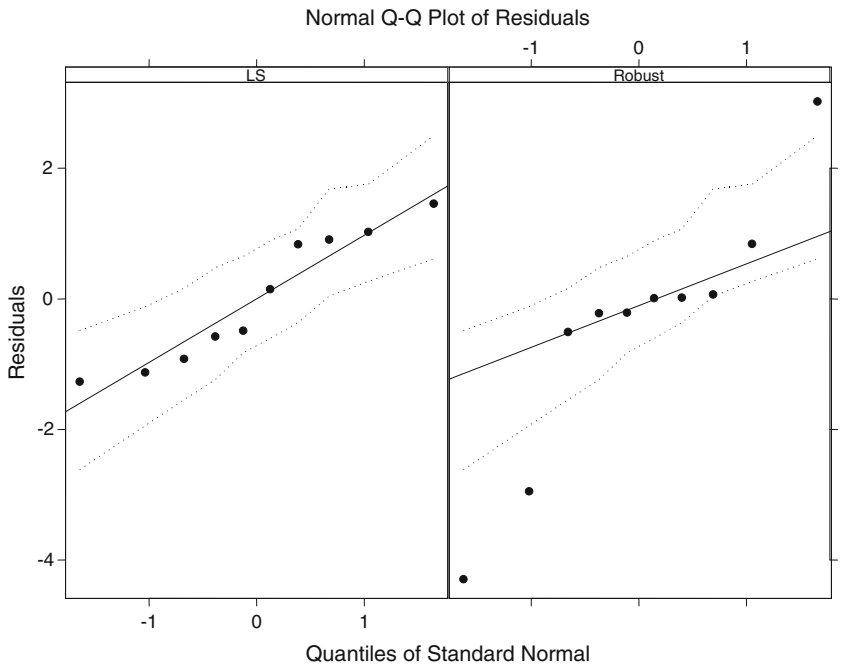

Fig. 2 Normal q-q plot of residuals comparing OLS with robust regression (MM)

range of effect sizes for robust regression versus OLS regression for data from "effective interventions"?

Method

This study used two data sets. One data set addressed the questions about the performance of robust regression when compared with OLS. The second data set was used to address questions about the relationship between robust regression and visual analysis.

A convenience sample of AB data series was obtained from ERIC and PsychINFO searches using the search terms "single case," "single subject," "time series," "baseline," "AB," "ABA," "ABAB," and "ABC." For multiphase designs, the sample included only the initial AB phases. For multiple baseline designs, each data series was treated separately. When the graphs were clear enough for electronic scanning, they were scanned and digitized using *i-extractor* software (Linden Software Ltd., 1998). Additional detail on the procedure is explained in Parker et al. (2005). A total of 61 graphs were used in this study. The final sample represented 21 articles, listed in the Appendix. This data set, consisting of published data, was used to answer the first two questions: (1) To what degree does outlier detection vary from OLS to robust regression? (2) How much do effect sizes differ from OLS to robust regression?

The second data set was initially presented in a previous study (Brossart et al., 2006). In that study, 250 single-case AB (baseline and intervention phases) data sets were created from random number series to represent a range of degrees and types of intervention effects. Each phase was transformed separately by adjusting four levels of four statistical characteristics: (1) variability, (2) trend, (3) mean level, and (4) gap of trend line intercepts between data points 10 and 11. Each data set contained a total of 20 data points, with 10 per phase. Thirty-five graphs were then selected that were representative of the 250 on the four attributes; however, they also had to show comparatively little trend in phase A (to mirror most published graphs).

Brossart et al., (2006) recruited judges who were doctoral students and faculty in an educational psychology department. After interviewing each potential rater, they classified each of the 45 respondents as being experienced or not in teaching graphic analysis. They omitted all respondents who did not have sufficient experience evaluating SCR graphs, leaving 15 judges.

The judges were presented details of a scenario describing the baseline and treatment phases, as well as a description of the instrument used to monitor progress. Acting as consultants, the judges were asked to evaluate intervention effectiveness on the basis of visual analysis of the AB graphs of the target behavior. Thirty-five graphs were evaluated, which the authors reported to be the maximum number for obtaining cooperation and maintaining good concentration with their sample of judges. These data were reanalyzed using robust regression to link the results in that previous study to the performance of OLS regression and the ratings of treatment effectiveness by the judges. Thus, this second data set was used to answer the remaining questions (Are the differences produced by

robust regression in more or less agreement with visual judgments of treatment effectiveness? and What is a typical range of effect sizes for robust regression vs. OLS regression for data from "effective interventions"?).

Procedure The procedure for this study was to evaluate the single-case data sets using a method that removes trend in the baseline phase or phase A (Allison & Gorman, 1993; Faith et al., 1996). The model allows one to test for mean phase differences only (which we will call the *Allison mean*, or AM) or to include both phase and trend components (which we will call the *Allison mean plus trend*, or AMT) after controlling for phase A trend only. This method has shown promise in evaluating single-case data in previous studies (Parker & Brossart, 2003; Parker et al., 2005), and AMT appears to possess enough power for a large proportion of data series found in the literature (Parker et al., 2005). Both OLS and robust regression (MM) were used to produce results with the AM and AMT method on the same 61 data sets, using S+ (TIBCO Software Inc., 2008a). A second data set, originally presented in Brossart et al. (2006), was reanalyzed using robust regression as well, so that comparisons could be made to data sets rated visually by judges.

To run these regression models, both AM and AMT require preliminary detrending in four steps: (1) Create a temporary variable containing the scores for Phase A only; (2) regress this new variable on trend (the time variable); (3) save the predicted output; and (4) subtract these predicted values from the original scores. The resulting differences or residual scores are used, instead of the original scores, in the final regression analysis for AM $R^2_{detY.M}$, and for AMT $R^2_{detY.M.TB}$, (where *detY* is the detrended *Y* scores, subscript *M* is a dummy-coded phase mean shift vector variable, *T* is a time or trend variable, and *TB* is a variable containing trend scores for phase B only).

Results

To what degree does outlier detection vary from OLS to robust regression? There are multiple ways one may evaluate data sets for outliers. For this study, a q-q plot for each of the 61 data sets was created that compared OLS with the robust method. Those data points that fell outside the 95% simulation envelopes for the normal q-q plot, shown as dotted lines, are outliers (see Fig. 2 for an example). This reveals one of the most important advantages of a good robust fit; it clearly exposes outliers, while the least squares fit is highly influenced by outliers in such a way that the outliers are often not clearly revealed in the residuals. The results show that for OLS, 51 (83.6%) of the q-q plots revealed no outliers, whereas the robust method showed that 24 (39.3%) of the graphs had no outliers. Across all graphs, the mean number of outliers detected per graph by OLS was 0.87 (*SD* = 4.3), and for the robust method, the mean number of outliers detected per graph was 2.93 (*SD* = 5.46). Table 1 displays the number of outliers detected for each method. A striking feature of this table is that OLS detects far fewer outliers, especially in those data sets with a single outlier or those with more than 4 outliers. The robust method detected 4 or more outliers in 21.3% of the data sets, whereas OLS detected 4 or more outliers in only 4.9% of the data sets. Table 2 displays the amount of disagreement between the two methods in identifying outliers. There was no disagreement between methods for 24 graphs (39.3%). Yet for 18% of the graphs, the robust method detected 4 or more outliers than were detected by OLS.

Table 1 Number of outliers detected for each method

Method	Detected Outliers	Frequency	%
OLS	0	51	83.6
	1	4	6.6
	2	1	1.6
	3	2	3.3
	≥4	3	4.9
Robust	0	24	39.3
	1	15	24.6
	2	5	8.2
	3	4	6.6
	≥4	13	21.3

How much do effect sizes differ from OLS to robust regression? Table 3 provides summary statistics for the R^2 from OLS and robust regression (TIBCO Software Inc., 2008b). It should be noted that R^2 is a commonly reported effect size, but it may be converted to other effect sizes, such as Cohen's *d*, where it can be computed from *R* $\left(R = \frac{d}{\sqrt{d^2+4}}\right)$ (Rosenthal, 1991; Wolf, 1986). Most single-case researchers will be largely interested in the first two columns, because they address treatment effects in terms of the mean difference between the two phases. For completeness, the third and fourth columns list the results with a

Table 2 Number of outliers detected by robust method minus number of outliers detected by OLS

Difference	Number	%
−2	1	1.6
−1	2	3.3
0	24	39.3
1	17	27.9
2	4	6.6
3	2	3.3
≥4	11	18.0

Table 3 Summary statistics of R^2 values produced by OLS and robust regression

	AM–OLS	AM–Robust	AMT–OLS	AMT–Robust
1st quartile	.36	.23	.53	.37
Mean	.48	.39	.67	.51
Median	.53	.40	.72	.53
3rd quartile	.69	.56	.88	.61
Standard deviation	.29	.21	.26	.18

slope parameter added to the model. Overall, in comparison with OLS, the robust method has a smaller standard deviation and a smaller mean R^2. Figure 3 displays the descriptive information in the form of box plots. Correlations comparing OLS and the robust method effect sizes show that with a phase only model (AM), they correlate highly (r = .86), and with slope added (AMT), the correlation was .77. Subtracting the robust R^2 from the OLS R^2 portrays the magnitude and direction of the differences between the two methods. These results are given in Table 4. For the model with phase or mean level only comparisons (AM), there were 15 data sets (24%) where the robust method produced larger R^2 values than did OLS. For the remaining 47 data sets, OLS produced larger R^2 values than did the robust method for 46 data sets (74%), with 1 data set producing no difference in the R^2 values. When trend was added to the model (AMT), there were 10 data sets (16%) where robust regression produced larger values than OLS. With 3 data sets producing no difference, for the remaining 49 data sets (79%), OLS produced larger R^2 values.

If one assumes that a difference equal to or greater than 10% between the R^2 produced by OLS and robust regression is large enough to be meaningful or important (a 10% difference in the amount of variance accounted for), then, with the AM model, 54.8% (and 75.8% with AMT) of the data sets analyzed exhibited a difference large enough to be considered problematic. This suggests that in many, but not all, data sets, robust regression should be seriously considered over OLS. Thus, OLS was found to frequently over- or underestimate R^2 to such an extent as to be of concern.

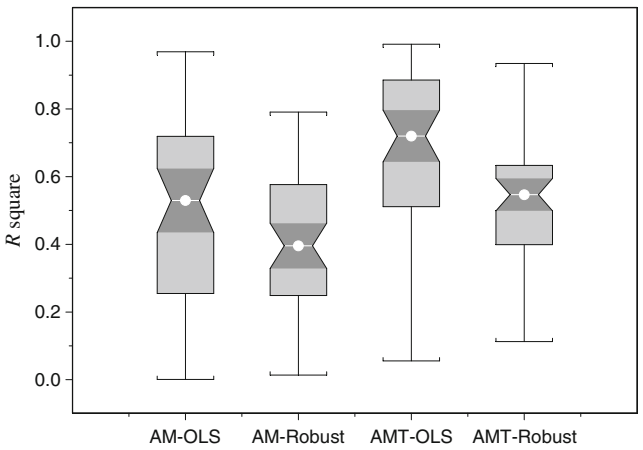

Fig. 3 Box plots of R^2 comparing AM and AMT with OLS and robust regression (MM) methods

Graphs of intervention effectiveness Are the differences produced by robust regression in more or less agreement with visual judgments of treatment effectiveness? To answer this question, we analyzed the data reported in Brossart et al. (2006). Using the ratings of 15 experienced users of single-case graphs, they categorized 35 graphs on the basis of average intervention effectiveness ratings. Using a 5-point judgment scale, mean ratings of 1–2.9 were defined as *not effective or minimally effective* interventions (8 graphs, or 23%). Mean ratings of 3.0–3.5 were defined as *somewhat effective* interventions (13 graphs, or 37%), and mean ratings of 3.6–5.0 were defined as *effective or very effective* interventions (14 graphs, or 40%). Robust correlations between the judges' ratings and the robust statistical regressions produced the following results: AM-OLS, r = .56; AM-robust, r = .49,; AMT-OLS, r = .57 (Brossart et al., 2006); AMT-robust, r = .53. Thus, overall, there was a small reduction in the correlation between the robust method and expert visual judgments. The take-home message from this is that robust methods (including non robust methods like OLS), correlate only moderately with ratings from expert judges using visual analysis.

What is a typical range of effect sizes for robust regression, as compared with OLS regression, for data from "effective interventions"? Using the aforementioned categories, the performance of robust regression, as compared with OLS, is reported in Table 5 for each category of treatment effectiveness. In general, the robust R^2s are smaller for each category than those from OLS and will be discussed further below.

Discussion

This study addressed several practical questions for single-case researchers who are considering using robust statistical methods. Beginning with the first question, we attempted to

Table 4 Differences in R^2 values when robust R^2 is subtracted from the OLS R^2

	R^2 Differences in AM Model	R^2 Differences in AMT Model
Minimum	−.35	−.40
10%	−.11	−.06
25%	0	.14
50%	.09	.17
Mean	.09	.16
75%	.21	.30
90%	.38	.40
Maximum	.48	.58

answer the following: To what degree does outlier detection vary from OLS to robust regression? The results suggest that the traditional methods based on OLS are inadequate to detect outliers. For example, in 84% of the graphs examined, no outliers were detected using OLS. In contrast, using robust regression, at least one outlier was present in 61% of the data sets examined. Furthermore, the robust method detected four or more outliers in just over 21% of the data sets, whereas the traditional method found four or more outliers in only 4.9% of the data sets. Thus, in terms of the limitations of OLS regression noted earlier, the warning that OLS is not an effective method for ascertaining and examining outliers was confirmed.

The second question addressed was the following: How much do effect sizes differ from OLS to robust regression? This study showed that regardless of the model used (AM or AMT), in over 70% of the data sets examined, OLS produced a larger effect size than did the robust method. Furthermore, the difference was larger than 10% (a 10-point difference in the R^2) in a little over half (55%) of the data sets, using a phase or mean level comparison (AM), and in roughly three fourths (76%) of the data sets when the trend component was included (AMT). Thus, the differences between the two methods were large enough to be generally considered nontrivial. These differences highlight a limitation noted earlier—that measures of effect size can be very inaccurate when based on OLS regression.

Next, we asked the following: Are the differences produced by robust regression in more or less agreement with visual judgments of treatment effectiveness? The correlations with visual judgments of treatment effectiveness were smaller but similar to those produced by OLS. This finding was in line with our expectations and was not surprising given the unreliability of expert judges' evaluation of treatment effectiveness using visual analysis (Brossart et al., 2006).

We then sought to answer the following: What is a typical range of effect sizes for robust regression versus OLS regression for data from "effective interventions"? The results suggest that more work needs to be done in this area. Specifically, we examined average judge ratings for each category of treatment effectiveness (not effective, somewhat effective, very effective) and found that. often, the differences between categories were very small. For instance, when examining the full model with mean level and trend included (AMT), there was only a 1-point difference in the average effect size for graphs judged to be *not effective* versus those deemed *somewhat effective*. There was a 7-point difference in the average effect size between the *somewhat effective* category and those graphs rated *very effective*. When trend was not included in the model (AM—mean level only), the categories showed a greater degree of separation, yet there was only a 10-point difference between the average effect size for the *somewhat effective* category (R^2 = .41) and those graphs rated *very effective* (R^2 = .51).

We attempted to link statistical results and visual judgment, yet given the results, it appears that these categories may be of limited use for those trying to interpret an effect size using the method presented here. More work needs to be done before it is clear as to how these effect sizes are related to a given treatment effect.

Table 5 Mean effect size values (averaged over 15 judges) for three groups of graphs, judged as depicting "not effective," "somewhat effective," and "very effective" interventions

Analytic Technique	Not Effective (8 graphs)	Somewhat Effective (13 graphs)	Very Effective (14 graphs)
AM-OLS R^2	.36	.52	.67
AM-robust R^2	.29	.41	.51
AMT-OLS R^{2*}	.65	.73	.87
AMT-robust R^2	.58	.59	.66

*Reported in Brossart et al. (2006)

These results may be viewed as "ball park" estimates or as tentative suggestions for how one may interpret their results, but all effect sizes should be placed in the context of the study from which they were derived for a proper interpretation. It seems worth repeating the caution voiced by other researchers, that Cohen's (1988) guidelines for small, medium, and large effect sizes to do not hold for single-case research and that the effect sizes produced may vary greatly depending on the statistical method used (e.g., Brossart et al., 2006; Parker & Brossart, 2003; Parker et al., 2005).

There are several important limitations to the present study to consider. First, the results were based on 61 previously published data sets. A larger sample is desirable, but the present sample contained a variety of studies and probably provides a "rough sketch" of what the body of published single-case data looks like in terms of providing an initial examination of how robust regression using AM or AMT would perform. Second, the comparisons with judges' ratings were based on a limited sample of graphs. It may be that the distinctions investigators use to judge graphs do actually translate into small effect size differences when distinctions are made between effective and not effective interventions (e.g., an average of .58 for a *not effective* intervention to an average of .59 for a *somewhat effective* intervention, and then to an average of .66 for a *very effective* intervention based on robust AMT), but until more research is conducted in this area, the findings should be viewed as tentative. Third, this article did not address autocorrelation. Further investigation regarding the role of autocorrelation in the robust method presented here is warranted. Lack of independence in single-case data has been found to exist and has been studied by a number of researchers (e.g., Hartmann et al., 1980; Huitema & McKean, 1991; Sharpley & Alavosius, 1988; Suen & Ary, 1987). It should be noted that serial dependence can be removed before conducting one's primary analysis, but in most cases, its removal has little impact on any resulting effect size (Parker, 2006). Even so, it may prove beneficial to examine the effect of removing the autoregressive component in single-case data, using an ARIMA model suggested by Brossart et al. (2006) and Parker et al. (2006). Some researchers have noted that even with the concerns of serial dependence, the advantages of using statistical methods overshadow the concerns (Matyas & Greenwood, 1996).

Given the current state of knowledge about SCR, few researchers continue to advocate the sole use of visual analysis. Yet our knowledge base about the meaning of the effect sizes produced by various methods to analyze single-case data suggests that abandoning some type of visual analysis may be premature. The results presented illustrate that reliance on nonrobust methods (OLS specifically) in single-case data analysis should be questioned and that a robust form of the AM or AMT model has clear advantages over OLS-based AM or AMT. For those investigators who need to document treatment or experimental effects and who wish to supplement visual analysis with an empirical method, MM robust regression appears to be a better choice than OLS-based methods.

Appendix

Allen, S. J., & Kramer, J. J. (1990). Modification of personal hygiene and grooming behaviors with contingency contracting: A brief review and case study. *Psychology in the Schools*, *27*, 244–251.

Anhalt, K., McNeil, C. B., & Bahl, A. B. (1998). The ADHD classroom kit: A whole-classroom approach for managing disruptive behavior. *Psychology in the Schools*, *35*, 67–79.

Bray, M. A., & Kehle, T. J. (1998). Self-modeling as an intervention for stuttering. *School Psychology Review*, *27*, 587–598.

Bujold, A., Ladouceur, R., Sylvain, C., & Boisvert, J.-M. (1994). Treatment of pathological gamblers: An experimental study. *Journal of Behavior Therapy & Experimental Psychiatry*, *25*, 275–282.

Burnette, M. M., Koehn, K. A., Kenyon-Jump, R., Hutton, K., & Stark, C. (1991). Control of genital herpes recurrences using progressive muscle relaxation. *Behavior Therapy*, *22*, 237–247.

Cassisi, J. E., & McGlynn, F. D. (1988). Effects of EMG-activated alarms on nocturnal bruxism. *Behavior Therapy*, *19*, 133–142.

Chadwick, P. (1994). Examining specific cognitive change in cognitive therapy for depression: A controlled case experiment. *Journal of Cognitive Psychotherapy*, *8*, 19–31.

Chadwick, P., & Lowe, C. (1990). Measurement and modification of delusional beliefs. *Journal of Consulting and Clinical Psychology*, *58*, 225–232.

Chadwick, P., & Trower, P. (1996). Cognitive therapy for punishment paranoia: A single case experiment. *Behavior Research and Therapy*, *34*, 351–356.

Hartley, E. T., Bray, M. A., & Kehle, T. J. (1998). Self-modeling as an intervention to increase student classroom participation. *Psychology in the Schools*, *35*, 363–372.

Ladouceur, R., Boisvert, J.-M., & Dumont, J. (1994). Cognitive-behavioral treatment for adolescent pathological gamblers. *Behavior Modification*, *18*, 230–242.

Ladouceur, R., Freeston, M. H., Gagnon, F., Thibodeau, N., & Dumont, J. (1993). Idiographic considerations in

the behavioral treatment of obessional thoughts. *Journal of Behavior Therapy & Experimental Psychiatry*, *24*, 301–310.

Lemanek, K. L., & Gresham, F. M. (1984). Social skills training with a deaf adolescent: Implications for placement and programming. *School Psychology Review*, *13*, 385–390.

Lopez, A., & Cole, C. L. (1999). Effects of a parent-implemented intervention on the academic readiness skills of five Puerto Rican kindergarten students in an urban school. *School Psychology Review*, *28*, 439–447.

O'Kearney, R. (1993). Additional considerations in the cognitive-behavioral treatment of obsessional ruminations: A case study. *Journal of Behavior Therapy & Experimental Psychiatry*, *24*, 357–365.

Pray, B., Kramer, J. J., & Lindskog, R. (1986). Assessment and treatment of tic behavior: A review and case study. *School Psychology Review*, *15*, 418–429.

Rankin, H. (1982). Control rather than abstinence as a goal in the treatment of excessive gambling. *Behavior Research and Therapy*, *20*, 185–187.

Salend, S. J., Whittaker, C. R., Raab, S., & Giek, K. (1991). Using a self-evaluation system as a group contingency. *Journal of School Psychology*, *29*, 319–329.

Shapiro, E. S., Albright, T. S., & Ager, C. L. (1986). Group versus individual contingencies in modifying two disruptive adolescents' behavior. *Professional School Psychology*, *1*, 105–116.

Sharpe, T., & Lounsbery, M. (1997). The effects of a sequential behavior analysis protocol on the teaching practices of undergraduate trainees. *School Psychology Quarterly*, *12*, 327–343.

Tollefson, N., Tracy, D. B., Johnsen, E. P., & Chatman, J. (1986). Teaching learning disabled students goal-implementation skills. *Psychology in the Schools*, *23*, 194–204.

References

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*, 621–631.

Anderson, C., & Schumacker, R. E. (2003). A comparison of five robust regression methods with ordinary least squares regression: Relative efficiency, bias, and test of the null hypothesis. *Understanding Statistics, 2*, 70–103.

Birkes, D., & Dodge, Y. (1993). *Alternative methods of regression*. New York: Wiley.

Blais, M. A., & Hilsenroth, M. J. (2006). Methodcentric reasoning and the empirically supported treatment debates. In S. G. Hofmann & J. Weinberger (Eds.), *The art and science of psychotherapy* (pp. 31–47). New York: Routledge.

Brossart, D. F., Meythaler, J. M., Parker, R. I., McNamara, J., & Elliott, T. R. (2008). Advanced regression methods for single-case designs: Studying propranolol in the treatment for agitation associated with traumatic brain injury. *Rehabilitation Psychology, 53*, 357–369.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531–563.

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 159–185). Hillsdale, NJ: Erlbaum.

Center, B. A., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, *19*, 387–400.

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

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573–579.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245–277). Mahwah, NJ: Erlbaum.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Harbst, K. B., Ottenbacher, K. J., & Harris, S. R. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy, 71*, 107–115.

Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis, 13*, 543–559.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and expploratory data analysis*. New York: Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1985). *Exploring data tables, trends, and shapes*. New York: Wiley.

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin, 110*, 291–304.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate scientist*. Thousand Oaks, CA: Sage.

Kaplan, R. M., & Groessl, E. J. (2002). Applications of cost-effectiveness methodologies in behavioral medicine. *Journal of Consulting and Clinical Psychology, 70*, 482–493.

Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification, 2*, 291–307.

Linden Software Ltd. (1998). i-extractor (Version 1.0) [Computer software]. U.K.: Author.

Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215–243). Mahwah, NJ: Erlbaum.

McHugh, K. R., & Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. *The American Psychologist, 65*, 73–84.

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design. *The American Psychologist, 56*, 119–127.

Newnham, E. A., & Page, A. C. (2010). Bridging the gap between best evidence and best practice in mental health. *Clinical Psychology Review, 30*, 127–142.

Ottenbacher, K. J. (1990). When is a picture worth a thousand p values? A comparison of visual and quantitative methods to

analyze single subject data. *Journal of Special Education, 23*, 436–449.

Park, H., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis of single-case designs. *Journal of Experimental Education, 58*, 311–320.

Parker, R. I. (2006). Increased reliability for single-case research results: Is the bootstrap the answer? *Behavior Therapy, 37*, 326–338.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189–211.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., De-Alba, R. G., Baugh, F. G., et al. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34*, 116–132.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418–443.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education, 40*, 194–204.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Revisedth ed.). Newbury Park, CA: Sage.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*, 871–880.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 188–196.

Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavioral data: An alternative perspective. *Behavioral Assessment, 10*, 243–251.

Shavelson, R., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Suen, H. K., & Ary, D. (1987). Autocorrelation in applied behavior analysis: Myth or reality? *Behavioral Assessment, 9*, 125–130.

TIBCO Software Inc. (2008a). *Spotfire S+ (Version 8.1) [Computer program]*. Seattle, WA: TIBCO.

TIBCO Software Inc. (2008b). *TIBCO Spotfire S+ 8.1 Robust Library user's guide*. Palo Alto, CA: Author.

Tucker, J. A., & Reed, G. M. (2008). Evidentiary pluralism as a strategy for research and evidence-based practice in rehabilitation psychology. *Rehabilitation Psychology, 53*, 279–293.

Tucker, J. A., & Roth, D. L. (2006). Extending the evidence hierarchy to enhance evidence - based practice for substance use disorders. *Addiction, 101*, 918–932.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 448–485). Standford, CA: Stanford University Press.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (1998a). The goals and strategies of robust methods. *The British Journal of Mathematical and Statistical Psychology, 31*, 1–39.

Wilcox, R. R. (1998b). How many discoveries have been lost by ignoring modern statistical methods? *The American Psychologist, 53*, 300–314.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Elsevier.

Wilcox, R. R., & Keselman, H. J. (2004). Robust regression methods: Achieving small standard errors when there is heteroscedasticity. *Understanding Statistics, 3*, 349–364.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis (Vol. 59)*. Beverly Hills, CA: Sage.