The Spanish Journal of Psychology
2009, Vol. 12, No. 2, 823-832

ISSN 1138-7416

Factors Affecting Visual Inference in Single-Case Designs

Verônica M. Ximenes[1], Rumen Manolov[2], Antonio Solanas[2] and Vicenç Quera[2]
[1]Universitadade Federal do Ceará (Brazil)
[2]Universitat de Barcelona (Spain)

Visual inspection remains the most frequently applied method for detecting treatment effects in single-case designs. The advantages and limitations of visual inference are here discussed in relation to other procedures for assessing intervention effectiveness. The first part of the paper reviews previous research on visual analysis, paying special attention to the validation of visual analysts' decisions, inter-judge agreement, and false alarm and omission rates. The most relevant factors affecting visual inspection (i.e., effect size, autocorrelation, data variability, and analysts' expertise) are highlighted and incorporated into an empirical simulation study with the aim of providing further evidence about the reliability of visual analysis. Our results concur with previous studies that have reported the relationship between serial dependence and increased Type I rates. Participants with greater experience appeared to be more conservative and used more consistent criteria when assessing graphed data. Nonetheless, the decisions made by both professionals and students did not match sufficiently the simulated data features, and we also found low intra-judge agreement, thus suggesting that visual inspection should be complemented by other methods when assessing treatment effectiveness.
Keywords: visual inspection, single-case designs, autocorrelation, expertise.

La inspección visual sigue siendo el método más utilizado para detectar tratamientos efectivos en diseños de caso único. El presente trabajo comenta las ventajas y limitaciones de la inferencia visual en relación con otros procedimientos empleados para evaluar la efectividad de las intervenciones. La primera parte del manuscrito revisa investigaciones previas sobre el análisis visual, enfocando la validación de las decisiones de los analistas visuales, la concordancia entre jueces y las tasas de falsas alarmas y omisión. Se hace énfasis en los factores que más afectan a la inspección visual (i.e., tamaño del efecto, autocorrelación, variabilidad en los datos y experiencia de los analistas) y éstos se incluyen en un estudio de simulación que pretende aportar evidencias sobre la calidad del análisis visual. Nuestros resultados coinciden con estudios previos sobre la relación entre la dependencia serial y un incremento en las tasas de error Tipo I. Los participantes con mayor experiencia parecen ser más conservadores y utilizan criterios más consistentes al evaluar datos gráficos. No obstante, tanto las decisiones de los profesionales y como las de los estudiantes no se corresponden lo suficiente con los datos simulados. Además, se encontró una baja consistencia intra-jueces, sugiriendo que la inspección visual se debería complementar por otros métodos a la hora de evaluar la efectividad de los tratamientos.
Palabras clave: inspección visual, diseños de caso único, autocorrelación, experiencia.

This research was supported by the Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa of the Generalitat de Catalunya, the European Social Fund, the Ministerio de Educación y Ciencia grant SEJ2005-07310-C02-01/PSIC, and the Generalitat de Catalunya grant 2005SGR-00098.

The authors would like to thank the anonymous reviewers for their useful comments and suggestions, which contributed to improving the manuscript.

Correspondence concerning this article should be addressed to Rumen Manolov, Departament de Metodologia de les Ciències del Comportament, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, 08035-Barcelona (Spain). Phone: +34-93-3125844. E-mail: rrumenov13@ub.edu

The data obtained from single-case studies have been and still are mostly analysed by means of visual inspection, which is generally employed alone and seldom complemented by some sort of statistical analysis (Kratochwill & Brody, 1978; Parker & Brossart, 2003; Parker, Hagan-Burke, & Vannest, 2007). The more frequent application of visual analysis in comparison to statistical tests can be explained by the advantages of the former and the limitations of the latter. Graphic displays have a compact and detailed data-reporting format, allow decisions to be made over the course of a study, and enable different analysts to judge for themselves whether an intervention is effective (Richards, Taylor, & Ramasamy, 1997). It has been claimed that no existing statistical technique can simultaneously handle data variability, trend magnitude and direction, cycles, delayed responses, and level changes (Parker, Cryer, & Byrns, 2006; Parker & Hagan-Burke, 2007c). Moreover, there is evidence for the deficient performance, in terms of Type I and/or Type II errors, of statistical techniques as diverse as ANOVA (Toothaker, Banz, Noble, Camp, & Davis, 1983), time-series analysis (Greenwood & Matyas, 1990), the split-middle method (Crosbie, 1987), and randomization tests (Ferron & Ware, 1995; Sierra, Solanas, & Quera, 2005). Additionally, statistical significance tests have been said to impose limitations on researchers (Michael, 1974) and assign only secondary importance to the more relevant clinical significance (Hugdahl & Ŏst, 1981). The use of effect sizes instead of *p*-values is not trouble-free either, as there are currently no established guidelines for their interpretation in single-case designs (Parker & Brossart, 2006).

However, statistical techniques may be needed when a stable baseline does not exist and when results must serve as objective documentation (Kazdin, 1982). Regarding this latter aspect, Busk and Serlin (1992) point out that the use of visual analysis alone does not enable the quantitative integration of results from different studies, as meta-analyses require the computation of effect sizes. The calculation of these summary measures from $N = 1$ data has been strongly influenced by visual inspection. Recent research on single-case designs has centred on developing effect size indices related to the criteria used by visual analysts, such as the amount of data overlap (Ma, 2006; Parker & Hagan-Burke, 2007a; 2007b; Parker et al., 2007).

The objective of the present study is to provide an overview of visual inspection as a means of assessing treatment effectiveness in single-case data, emphasising the factors that appear to influence the performance of visual analysts. Correspondingly, the first part of the paper consists of a review of previous research on the topic. Since an additional goal is to extend the available evidence, an empirical study including the most important factors affecting visual analysis is also presented.

Previous research on visual inference

Prior to discussing the evidence obtained by previous research it should be noted that the variety of studies included in this review also implies a miscellany of methodologies (e.g., real data vs. simulated data, different validation criteria used), which may contribute to discrepant results and conclusions.

Type I and Type II errors. The proportion of cases in which visual analysts detect a nonexistent effect (i.e., commit a Type I error) or miss an existing effect (i.e., a Type II error) can be used as an indicator of their performance. Visual inspection has been argued to be an adequate tool for identifying strong interventions, as it is assumed to detect only large changes in the behaviour measured (Baer, 1977; Barlow & Hersen, 2008). This conservativeness has been seen as an advantage over statistical techniques (Parsonson & Baer, 1986) and has been supported by evidence for high Type II error (i.e., miss) rates (Jones, Weinrott, & Vaught, 1978; Ottenbacher, 1990b). However, more recent studies suggest that Type I errors (i.e., false alarms) can also be excessively frequent (Fisch, 2001; Normand & Bailey, 2006). Moreover, Type I error rates can be inflated according to the number of times a decision is made in response-guided experimentation (Allison, Franklin, & Heshka, 1992).

Inter-rater agreement. When testing the consistency between different visual analysts who are presented with the same graphs, several studies concur that the average concordance tends to be low (e.g., Brossart, Parker, Olsson, & Mahadevan, 2006; DeProspero & Cohen, 1979; Ottenbacher, 1990b; Park, Marascuilo, & Gaylord-Ross, 1990). This poor performance has been attributed to the differences between real-life practice and the settings used in validity studies (Brossart et al., 2006; Parsonson & Baer, 1992). Intra-judge consistency also needs to be explored as it seems to have been overshadowed by the desire to study inter-judge agreement.

The validation of visual inspection. Attempts to validate visual analysis have compared techniques such as the split-middle method (Ottenbacher, 1990b; Richards et al., 1997), time-series analysis (Jones et al., 1978), and randomization tests (Park et al., 1990), and generally report low agreement between the techniques. However, there is clearly no single process called statistical analysis that might provide a definitive validation criterion (Brossart et al., 2006). Moreover, there seems to be a consensus that visual and statistical analyses should be complementary rather than competing methods (e.g., Barlow & Hersen, 2008; Busk & Marascuilo, 1992; DeProspero & Cohen, 1979; Jones, Vaught, & Weinrott, 1977; Morley & Adams, 1991). Accordingly, recent research has focused on exploring techniques that can enhance visual analysis (Parker & Brossart, 2003), and attempts have also been made to combine it with quantitative methods in order to control Type I error rates (Ferron & Jones, 2006).

Apart from using some statistical test to assess the quality of visual analysis a different and probably more appropriate approach consists in constructing data sets and graphs, so that the actual data features are known (Parsonson & Baer, 1992). This strategy implies comparing visual analysts' answers with the correct answers as determined by data characteristics (i.e., presence or absence of intervention effects and extraneous variables). A study of this kind carried out by Normand and Bailey (2006) found low decision accuracy. Additionally, and as we shall see below, knowing the underlying data characteristics enables investigators to explore their relevance in the decision-making process taking place in visual inspection.

Effect size. As expected, greater agreement between visual analysts has been found for larger changes in level between baseline and treatment phases (Gibson & Ottenbacher, 1988; James, Smith, & Milne, 1996; Knapp, 1983; Morales Ortiz, 1992). These findings are consistent with the idea that visual analysts tend to omit small intervention effects (Parsonson & Baer, 1986).

Autocorrelation. For the precise estimation of autocorrelation from real data, long data series are required (Huitema & McKean, 1991) and an alternative is to generate data with known values of the autoregressive parameter (Fisher, Kelley, & Lomas, 2003). There have been divergent reports about the importance (Matyas & Greenwood, 1990a; 1990b) or lack of importance of serial dependence (Ottenbacher, 1986; 1990a) for visual inference, although both original and replication studies seem to support the former (James et al., 1996; Morales Ortiz, 1992; Rojahn & Schulze, 1985).

Data variability. The influence of data variability on visual inference has been studied from different perspectives. On the one hand, evidence regarding within-phase variability suggests that lower degrees of dispersion favour visual analysts' decisions (DeProspero & Cohen, 1979). On the other hand, changes in variability between phases have been found to correlate only minimally with inter-rater agreement (Gibson & Ottenbacher, 1988; Ottenbacher, 1986). Finally, data with greater variability across the whole data series are related to higher Type I error rates (Matyas & Greenwood, 1990b; Morales Ortiz, 1992).

Expertise. The area of expertise of the applied psychologist inspecting graphed data has also been shown to be relevant. Experts in visual analysis tend to use the magnitude of effect as a criterion, whereas the type of effect and the amount of variability in the data are usually taken into account by trainees in statistical analysis (Furlong & Wampold, 1982; Wampold & Furlong, 1981). Greater inter-rater reliability has been found among professionals with a statistical background (regardless of whether they have experience in single-case designs) in comparison to single-case design analysts (Harbst, Ottenbacher, & Harris, 1991). On the other hand, having more experience of visual analysis does not necessarily imply a better performance (Knapp, 1983; Richards et al., 1997). Training may lead to more conservative judgments (James et al., 1996), or encourage analysts to rely solely on the criterion in which they have been trained (Skiba, Deno, Marston, & Casey, 1989).

An empirical study of factors affecting visual inference

Estimating the magnitude of effect, the degree of autocorrelation and variability from real data can be troublesome and depends on the procedure employed. Thus, in the present study, data with known parameters were generated by means of Monte Carlo methods in order to assess the accuracy of visual analysts' decisions; in other words, the correct answers (i.e., existence and magnitude of effect) are determined by these parameters. This makes it unnecessary to use any statistical technique as a gold standard for assessing visual inspection, as comparison between tools from different domains is not advisable, especially when there is no "known truth" (Parsonson & Baer, 1992).

Method

Data generation. AB designs with 20 observation points (10 in each phase) were generated according to the following model: $y_t = \mu + \delta + \varphi_1 \cdot y_{t-1} + \varepsilon_t$, where y_t is the data point at moment t, μ is the mean of the process (set to 25), and δ is the level change (set to zero for phase A). The aim was to generate increments in the response rate in phase B, represented by four values of d, the standardised mean difference (.0, .6, 1.2, and 1.8); the last three are referred to as "small", "intermediate" and "large" throughout this article. These labels have no relation to Cohen's (1988) benchmarks, whose appropriateness for single-case designs has been questioned due to the lack of correspondence with the effect sizes typically found in this kind of data (Parker et al., 2005). The expression for calculating the effect size d was presented by Cohen (1988) as $d = (\mu_B - \mu_A) / \sigma$. In the current study, this expression can be rewritten as $d = (\mu_A + \delta - \mu_A) / \sigma$, as phase B has the phase A mean μ_A increased by δ, and so $d = \delta / \sigma$. Hence, the level change to incorporate in the data generation formula depended on the desired effect size and on the standard deviation of the series: $\delta = d \cdot \sigma$. Since we intended to study the importance of level change while controlling data variability, autocorrelation also had to be controlled as it too affects data variability. In summary, the δ needed for a specific data series was obtained as a function of the effect size chosen, data variability, and the degree of serial dependence, the objective being to enable comparability of experimental conditions. The relationship between the factors is represented by the following expression, adapted from Kendall and Ord (1990):

$$\delta = d \cdot \sqrt{\frac{2\sigma_\varepsilon^2}{n}\left[1 + \frac{2\sum_{k=1}^{n}(n-k)\varphi_1^k}{n}\right]},$$

where k is the number of lags, n is the phase length (10, in the present study), φ_1 is the lag-one autocorrelation coefficient,

y_{t-1} is the data point at the moment previous to moment t, and ε_t is the error term generated from a normal distribution with mean equal to zero and standard deviation σ_ε.

The degrees of autocorrelation studied included both negative (−.8 and −.4) and positive values (.4 and .8), as well as independent series (φ_1 = .0). The level of serial dependence simulated was the same for the entire series (i.e., no distinction between phases was made). It should be noted that specifying negative autocorrelation is likely to produce data sets with greater variability, such as an alternation of high and low data points, while positive serial dependence may lead to upward or downward trends. Both of these data features are supposed to have an impact on the decisions made by visual analysts. It is likely that both the lack of stability and the presence of trends (i.e., $\varphi_1 = |.8|$) lead to constructing clinically undesirable baselines. However, such cases represent a small subset of the large number of samples generated.

In the present study the influence of whole-series variability was explored. The levels of variability studied were represented by a coefficient of variation (CV) of 50% and 150%, respectively. As CV = 100 (σ_ε / μ), the two values of the standard deviation of the error term were σ_ε = (CV · μ)/100 = (50 · 25) / 100 = 12.5 and σ_ε = (CV · μ)/100 = (150 · 25) / 100 = 37.5.

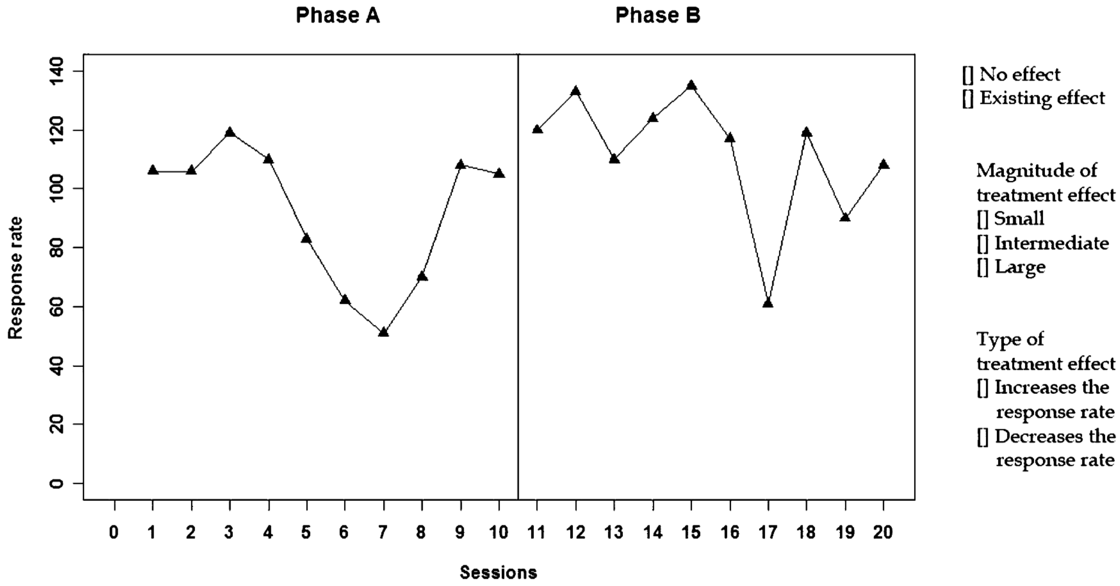
The simultaneous manipulation of the variables relevant to visual analysis may lead to confounding the specific effect of individual factors on the performance of visual analysts (Parsonson & Baer, 1992). In contrast, due to the simulation procedure followed in the present study, it was possible to study the influence of effect size, autocorrelation, and data variability separately. This is an innovative feature of the study, as previous simulation studies (e.g., Fisher et al., 2003; Matyas & Greenwood, 1990b) have, generally, manipulated level changes, variability, and autocorrelation simultaneously.

Participants

The 57 participants were divided into two sub-groups in order to study the influence of the level of expertise. An intentional sample of 24 psychologists with professional experience in visual analysis and single-case designs in academic and clinical contexts was selected by contacting all available experts. The other sub-group comprised a relatively similar number (33) of psychology undergraduates who had already passed courses such as *Applied and experimental designs* and *Behaviour therapy and modification*, which include visual analysis as part of their contents.

Instrument

Two questionnaires consisting of 60 graphs each were used in the present study. The first 40 graphs were all different from each other, while the last 20 were replications of half of the previously presented graphs and were incorporated to allow the calculation of an intra-judge agreement index. The two types of questionnaires differed only in terms of the effect size used to construct the graphs. Questionnaire type A included null and small effects, and type B intermediate and large effects. The order of the graphs was the same for all exemplars of the same type of questionnaire. This order was random with respect to the values of d, φ_1, and CV. The statement accompanying each questionnaire was as follows: "For each graph, indicate whether the treatment has an effect on the response rate or not. If you answer affirmatively, mark the magnitude and type of effect". This statement was intended to avoid dichotomous (presence/absence of treatment effect) answers and to permit participants to use their own criteria when



[] No effect
[] Existing effect

Magnitude of treatment effect
[] Small
[] Intermediate
[] Large

Type of treatment effect
[] Increases the response rate
[] Decreases the response rate

Figure 1. A graph belonging to a type B questionnaire.

making decisions, as suggested by Brossart et al. (2006). Figure 1 illustrates the format common to all graphs. The graphs constructed followed the guidelines for graphic representation summarised by Franklin, Gorman, Beasley, and Allison (1997). The questionnaires also included questions about the criteria used by the analysts in their decision-making processes.

Graphs were not contextualised, despite recommendations (Parsonson & Baer, 1992; Brossart et al., 2006), as we did not want the professionals' psychological area of specialisation to interfere with their decisions. For instance, if some of the clinicians treat people presenting auto-aggressive behaviour and others work with children with attention deficit hyperactivity disorder and the context of the graph is the former, then different judgements might be obtained due merely to previous experience (or lack of it) in the specific field.

Procedure

The students answered the questionnaire during class in the presence of one of the authors so that any doubts they might have could be addressed. There was no time limit imposed and the students needed an average of 19 minutes to respond to the instrument (17 to type A and 16 to type B). Each of the professionals was contacted individually and they received the same explanations as those given to students. Thirteen people responded to questionnaire type A and 11 to questionnaire type B, and these were returned by the established deadline. Each sub-group of participants answered only one type of questionnaire in order to avoid fatigue, since the total number of graphs was 120, which could jeopardise the answers to the last graphs.

Data analyses

The following indices were computed for each participant using the first 40 graphs of each questionnaire: a) Magnitude – calculated as the proportion of correct detections of the magnitude of the effect regardless of its direction; b) Type – calculated as the proportion of correct detections of effect direction regardless of its magnitude; c) Sensitivity – calculated as the proportion of graphs with $d \neq 0$ to which an "effect exists" answer is given by a participant; d) Specificity – calculated as the proportion of graphs with $d = 0$ to which an "effect does not exist" answer is given by a participant; e) Accuracy – calculated once through Cohen's (1960) κ with respect to the detection of presence versus absence of effect (using the same data as the Sensitivity and Specificity indices) and then once more with respect to the detection of the magnitude of the effect. In contrast with the previous indices, kappa is able to control for agreement due to chance.

Intra-judge consistency was also assessed using the last 20 items of the questionnaires and the ones they replicated. The calculation was as follows: 1) assign ranks 1, 2, 3, and 4 to participant answers "no effect", "small effect", "intermediate effect", and "large effect", respectively; 2) repeat the ranking procedure for both copies of each graph; 3) compute Spearman's rank order correlation, for each participant, between the ranks corresponding to the two copies of the same graph. Hence, the Consistency index ranged from −1 to 1.

Results

Performance of visual analysts

Participants' accuracy in distinguishing between presence and absence of treatment effect, as computed from Table 1, was low for both sub-groups (with kappa values of .101 for students and .104 for professionals, respectively). As expected, the accuracy in discerning the magnitude of effect, as computed from Table 2, was even lower: κ = .034 for students and κ = .095 for professionals.

The Mann-Whitney U test was used to explore differences in the indices (Consistency, Magnitude, and Type) between the two types of questionnaires and between

Table 1
Presence vs. absence of treatment effect: agreement between participants' responses and known truth

Participants' response	Data characteristics	
	Presence of effect	Absence of effect
Students – total responses = 1302		
Effect detected	54.15%	16.05%
No effect detected	20.05%	9.75%
Professionals – total responses = 944		
Effect detected	47.03%	14.30%
No effect detected	25.85%	12.82%

Note. Each cell represents the percentage of graphs for which the particular crossing between participant decisions and the correct answer occurred. The main diagonal should include 75% and 25%, since three-fourths of the graphs were constructed with intervention effect ($d \neq 0$).

Table 2
Magnitude of treatment effect: agreement between participants' responses and known truth.

Participants' response	Data characteristics: effect simulated			
	None	Small	Intermediate	Large
Students – total responses = 1302				
No effect	**9.75%**	8.60%	6.53%	4.92%
Small	6.30%	**6.22%**	5.91%	5.22%
Intermediate	6.22%	6.37%	**5.91%**	7.60%
Large	3.53%	4.45%	5.99%	**6.45%**
Professionals – total responses = 944				
No effect	**12.82%**	10.81%	8.16%	6.89%
Small	4.24%	**5.51%**	6.14%	5.61%
Intermediate	7.42%	6.36%	**4.45%**	5.30%
Large	2.65%	4.24%	4.03%	**5.40%**

Note. Each cell represents the percentage of graphs for which the particular crossing between participant decisions and the correct answer occurred. The main diagonal should include 25%, since each magnitude of effect (d = .0, .6, 1.2, and 1.8) was used to construct one-fourth of the graphs.

the two sub-groups. Intra-judge consistency was greater when assessing graphs with nonexistent or small effects for both students (mean rank for type A questionnaire 25.00 vs. mean rank for type B questionnaire 8.50, p < .001) and professionals (mean rank for type A questionnaire 18.00 vs. rank sum for type B questionnaire 6, p < .001). For the type A questionnaire professionals were more consistent in their decisions (mean rank for professionals 20.85 vs. mean rank for students 11.41, p = .0036). For the sub-group of professionals the correct detection of the magnitude of effect was also easier when null- and small-effect graphs were evaluated (mean rank for type A questionnaire 15.85 vs. mean rank for type B questionnaire 8.55, p = .0114). In contrast, the type (i.e., direction) of effect was better detected for type B questionnaires, which represent data sets with greater intervention effects. This finding was common to both students (mean rank for type B questionnaire 24.41 vs. mean rank for type A questionnaire 10.03, p < .001) and professionals (mean rank for type B questionnaire 16.05 vs. mean rank for type A questionnaire 9.50, p = .0232). For the type B questionnaire professionals were the better performing sub-group in terms of the Type index (mean rank for professionals 24.27 vs. mean rank for students 6.94, p = .033).

Influence of the factors studied

For both levels of experience we found that positive autocorrelation was associated with the overestimation of treatment effects. Figures 2 and 3 illustrate the association between positive serial dependence and lower accuracy in detecting the absence of treatment effect. The figures also show that negative serial dependence was associated with more conservative and, in the case of effect absence, more

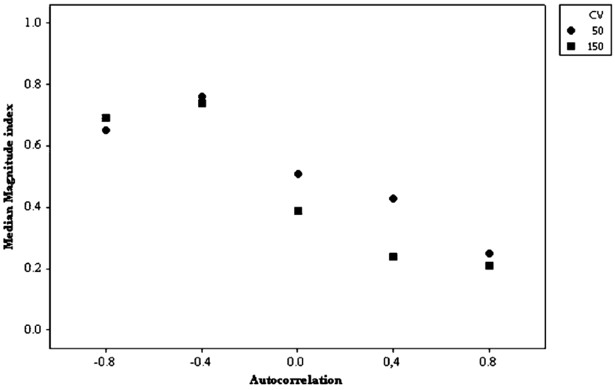


Figure 2. Detecting the magnitude of the effect simulated, in relation to autocorrelation and data variability (CV), for data series with d = .0 (i.e. no treatment effect simulated).

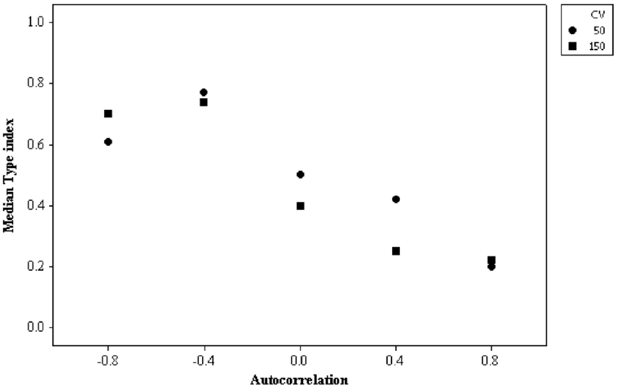


Figure 3. Detecting the direction of the effect simulated, in relation to autocorrelation and data variability (CV), for data series with d = .0 (i.e. no treatment effect simulated).

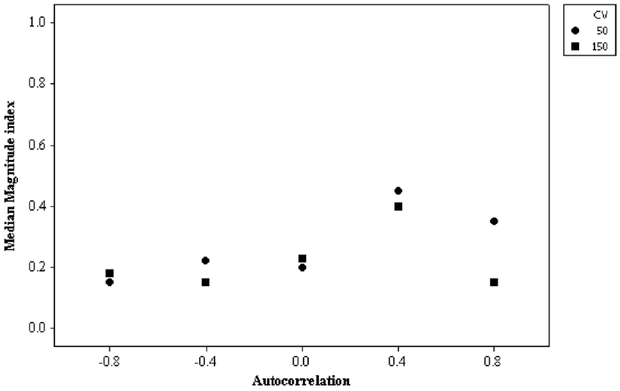


Figure 4. Detecting the magnitude of the effect simulated, in relation to autocorrelation and data variability (CV), for data series with *d* = 0.6 ("small" treatment effect simulated).

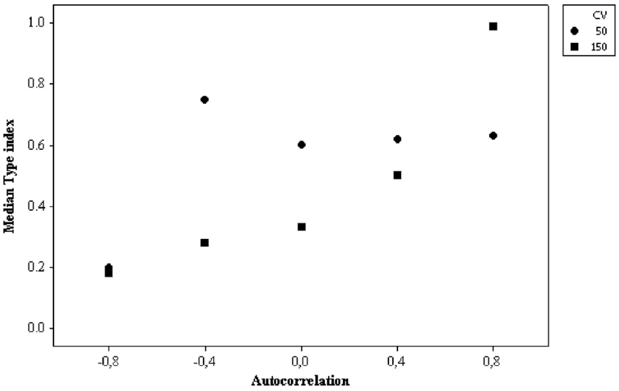


Figure 6. Detecting the direction of the effect simulated, in relation to autocorrelation and data variability (CV), for data series with *d* = 0.6 ("small" treatment effect simulated).

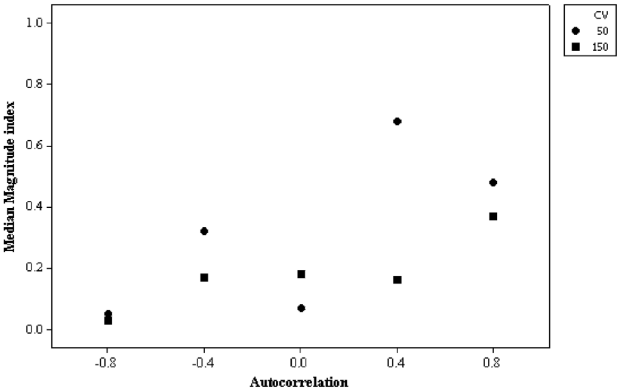


Figure 5. Detecting the magnitude of the effect simulated, in relation to autocorrelation and data variability (CV), for data series with *d* = 1.8 ("intermediate" treatment effect simulated).

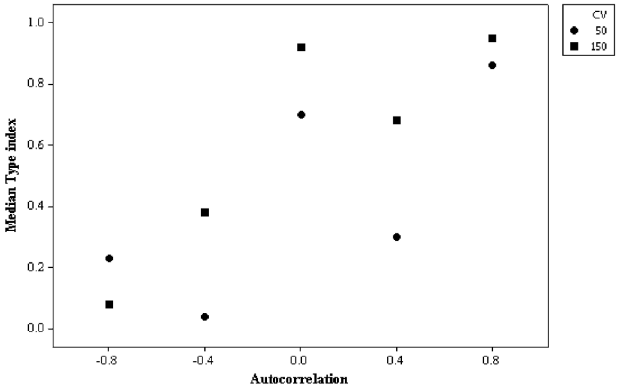


Figure 7. Detecting the direction of the effect simulated, in relation to autocorrelation and data variability (CV), for data series with *d* = 1.2 ("intermediate" treatment effect simulated).

accurate judgments. However, when a treatment effect exists, positive autocorrelation generally helps in detecting it, especially in the presence of less data variability (see Figures 4 and 5 for different effect sizes). Complementarily, negative autocorrelation undermines the correct detection of magnitude and also the direction of the existing treatment effect (see Figures 6 and 7 for different effect sizes).

The figures also show the influence of data variability on the accuracy of decisions made by participants. The lower values of the Magnitude index for CV = 150 suggest that greater overall data variability affects negatively the decision-making process. Detecting the direction of treatment effects is also troublesome in more disperse data series, although greater CV does not always imply a lower Type index.

As regards expertise, the main difference between students and professionals was the greater conservativeness of the latter, expressed in greater Specificity (.473 vs. .376) and lower Sensitivity (.645 vs. .730). Additionally, the professionals were more consistent in the criteria they used when making decisions, according to their reports.

Discussion

Visual analysts' performance

The Accuracy indices indicate scarce correspondence between the answers given by participants and the correct answers according to simulation parameters. The same conclusions can be drawn from inspection of the Magnitude and Type indices. Moreover, intra-judge consistency was found to be low, except in the case of professionals assessing graphs with small or nonexistent effects.

Type I and Type II errors

Some poor values of the Magnitude index were obtained for both graphs with and without a simulated effect. That is, visual inspection was not unequivocally related to liberality or conservativeness, but rather false alarms and omissions could prevail as a function of factors such as autocorrelation and expertise.

Effect size

The fact that the Consistency index was low in the case of d = 1.2 and 1.8 may be attributable to the difficulty of the task, which required distinguishing between intermediate and large treatment effects. It appears that participants found it easier to differentiate between no effect and small effects, although the concordance indices were not excessively high for the type A questionnaire either. These results do not, however, mean that greater effects are more difficult to detect, but may be attributable to the method followed.

Autocorrelation

Autocorrelation was shown to influence the accuracy of the decision-making process: positive serial dependence highlighted the simulated effects and was associated with more frequent Type I errors, while negative autocorrelation made it more difficult to detect effects.

Data variability

This factor requires further investigation as its pattern of influence is not clear enough, although there is some evidence that greater data variability hinders visual inference.

Expertise

Professionals tended to use more consistent criteria than did students when making decisions about intervention effectiveness. Greater consistency, however, does not imply greater accuracy. Students were more sensitive to treatment effects (committing fewer Type II errors), but they also tended to commit more Type I errors. Hence, more experience was associated with greater conservativeness, as pointed out by James et al. (1996).

Previous and current results seem to indicate that the performance of visual analysts depends on several factors. It is a technique that can lead to different decisions according to the person applying it (in relation to their experience and analytical criteria) and depending on how data are displayed (in relation to the presence or absence of visual aids). Data characteristics are also important, as visual analysis has been shown to be more reliable (and especially useful) for greater treatment effects, represented by larger effect sizes. Nevertheless, the conservativeness of visual inspection is not warranted when data are serially dependent, a likely situation in single-case designs (Parker, 2006). In addition, the low inter-judge agreement reported by previous research and the intra-judge inconsistencies detected in the present study suggest that visual analysis needs to be complemented by some analytical technique that provides quantitative results in order to enhance the decision-making process of applied psychologists. One possibility is to incorporate visual aids, since there is some evidence that regression or trend lines can increase inter-rater agreement, reliability and decision accuracy (Bailey, 1984; Fisher et al., 2003; Hagopian et al., 1997; Hojem & Ottenbacher, 1988; Skiba et al., 1989). Thus, it would be possible to combine the consistency of a systematic procedure (e.g., a statistical test, an effect size index or a visual aid) with the flexibility of visual analysis in the context of a specific area and in accordance with the aims and particular features of each individual case.

The resemblance between computer-generated and real behavioural data is always an issue in simulation studies. However, in the case of visual analysis there does not seem to be an optimal basis for assessing its reliability, since statistical techniques cannot unequivocally reveal the truth. A specific limitation of the present study is that only one design structure (AB) was simulated, and it would be necessary to investigate the decision-making process of visual analysts exposed to designs with more phase changes (e.g., ABAB designs) as these are likely to make it easier to assess results. In the event that all visual analysts participating in a study are specialists in the same psychological field, an example belonging to that area could be used to contextualise the graphical representations. In this way, an experimental situation resembling real-life settings could be achieved.

Future studies might focus on factors not studied here but which may also influence the decision-making process of visual analysts; for instance, the presence of overall trend or changes in trend seem to have an impact on decision accuracy (Gibson & Ottenbacher, 1988; Matyas & Greenwood, 1990b). Simulation studies may also specify delayed and temporary changes in response rate, as these are likely in behavioural data.

References

Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and Type I error rate in single-case designs. *The Journal of Experimental Education, 6*, 45-51.

Baer, D. (1977). "Perhaps it would be better not to know everything". *Journal of Applied Behavior Analysis, 10*, 167-172.

Bailey, D. B. (1984). Effects of lines of progress and semilogarithmic charts of ratings of charted data. *Journal of Applied Behavior Analysis, 17*, 359-365.

Barlow, D. H., & Hersen, M. (2008). *Single case experimental designs. Strategies for studying behavior change* (3rd ed.). Boston: Allyn & Bacon.

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 designs and analysis: New directions for psychology and education* (pp. 159-185). Hillsdale, NJ: Lawrence Erlbaum.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Lawrence Erlbaum.

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

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

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behavioral Assessment, 9*, 141-150.

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

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education, 75*, 66-81.

Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *The Journal of Experimental Education, 63*, 167-178.

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes, 54*, 137-154.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*, 387-406.

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1997). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119-158). Mahwah, NJ: Lawrence Erlbaum.

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inference. *Journal of Applied Behavior Analysis, 15*, 415-421.

Gibson, G., & Ottenbacher, K. (1988). Characteristics influencing the visual analysis of single-subject data: An empirical analysis. *The Journal of Applied Behavioral Science, 24*, 298-314.

Greenwood, K. M. & Matyas, T. A. (1990). Problems with application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12*, 355-370.

Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis, 30*, 313-326.

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

Hojem, M. A., & Ottenbacher, K. J. (1988). Empirical investigation of visual-inspection versus trend-line analysis of single-subject data. *Journal of the American Physical Association, 68*, 983-988.

Hugdahl, K., & Ŏst, L.-G. (1981). On the difference between statistical and clinical significance. *Behavioral Assessment, 3*, 289-295.

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

James, I. A., Smith, P. S., & Milne, D. (1996). Teaching visual analysis of time series data. *Behavioural and Cognitive Psychotherapy, 24*, 247-262.

Jones, R. J., Vaught, R. S., & Weinrott, M. R. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis, 10*, 151-166.

Jones, R. J., Weinrott, M. R., & Vaught, R. S., (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11*, 277-283.

Kazdin, A. E. (1982). *Single-case research design: Methods for clinical and applied settings.* New York: Oxford University Press.

Kendall, M., & Ord, J. K. (1990). *Time series.* London: Addison-Wesley Publishing Company.

Knapp. T. J. (1983). Behavioral analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*, 155-164.

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.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598-617.

Matyas, T. A., & Greenwood, K. M. (1990a). The effect of serial dependence on visual judgment in single-case charts: An addendum. *The Occupational Therapy Journal of Research, 10*, 208-220.

Matyas, T. A., & Greenwood, K. M. (1990b). Visual analysis for single-case time series: effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7*, 647-653.

Morales Ortiz, M. (1992). Análisis cualitativo: Concepto y posibilidades mediante lenguaje gráfico. Unpublished doctoral dissertation, University of Seville, Spain.

Morley, S., & Adams, M. (1991). Graphical analysis of single-case time series data. *British Journal of Clinical Psychology, 30*, 97-115.

Normand, M. P., & Bailey, J. S. (2006). The effects of celeration lines on visual data analysis. *Behavior Modification, 30*, 295-314.

Ottenbacher, K. J. (1986). Reliability and accuracy of visually analyzing graphed data from single-subject designs. *American Journal of Occupational Therapy, 40*, 464-469.

Ottenbacher, K. J. (1990a). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation, 28*, 283-290.

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

Parker, R. I. (2006). Increased reliability for single-case research results: Is 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. (2006). Phase contrasts for multiphase single case intervention designs. *School Psychology Quarterly, 21*, 46-61.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., & Sullivan, J. R. (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. (2007a). Median-based overlap analysis for single case data: A second study. *Behavior Modification, 31*, 919-936.

Parker, R. I., & Hagan-Burke, S. (2007b). Single case research results as clinical outcomes. *Journal of School Psychology, 45*, 637-653.

Parker, R. I., & Hagan-Burke, S. (2007c). Useful effect size interpretations for single case research. *Behavior Therapy, 38*, 95-105.

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

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157-186). New York: Plenum Press.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 15-40). Hillsdale, NJ: Lawrence Erlbaum.

Richards, S. B., Taylor, R. L., & Ramasamy, R. (1997). Effects of subject and rater characteristics on the accuracy of visual analysis of single subject data. *Psychology in the Schools, 34*, 355-362.

Rojahn, J., & Schulze, H. H. (1985). The linear regression line as a judgmental aid in visual analysis of serially dependent A-B time-series data. *Journal of Psychopathology and Behavioral Assessment, 7*, 191-206.

Skiba, R., Deno, S., Marston, D., & Casey, A. (1989). Influence of trend estimation and subject familiarity on practitioners judgements of intervention effectiveness. *Journal of Special Education, 22*, 433-446.

Sierra, V., Solanas, A., & Quera, V. (2005). Randomization tests for systematic single-case designs are not always appropriate. *The Journal of Experimental Education, 73*, 140-160.

Toothaker, L. E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). N = 1 designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 4*, 289-309.

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inference. *Behavioral Assessment, 3*, 79-82.

Received June 23, 2008
Revision received October 22, 2008
Accepted December 20, 2008