Article

Behavior Modification
2022, Vol. 46(1) 147–177
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0145445520969188
journals.sagepub.com/home/bmo
SAGE

Toward an Automation of Functional Analysis Interpretation: A Proof of Concept

Alison Cox[1] and Jonathan E. Friedel[2]

Abstract

The advent of functional analysis (FA) methodology paved the way for improved function-based behavioral interventions and ultimately client outcomes. Behavior analysts primarily rely on visual inspection to interpret FA results. However, the literature suggests interpretations may vary across raters resulting in poor interobserver agreement (IOA). To increase interpretation objectivity and address IOA issues, Hagopian et al. created visual-inspection criteria. They reported improved IOA, alongside criteria limitations. Following this, Roane et al. modified these criteria. The current project describes the first steps toward developing a decision support system to assist in FA interpretation. Specifically, we created a computer script, written in R, designed to evaluate FA data and produce an outcome (assign function) based on the Roane et al. criteria. Average agreement between experienced human raters and the computer script outcomes was 81%. We discuss criteria limitations (e.g., vague rules), study implications, and the significance of further research on this topic.

Keywords

decision support systems, functional analysis, visual inspection

[1]Brock University, St. Catharines, Ontario, Canada
[2]National Institute for Occupational Safety and Health, Morgantown, WV, USA

Corresponding Author:
Alison Cox, Department of Applied Disability Studies, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, Ontario L2S 3A1, Canada.
Email: acox@brocku.ca

Visual inspection represents one of the most common methods of data analysis in the field of behavior analysis (cf. Johnston & Pennypacker, 2009) and is generally the primary method for the applied analysis of behavior. With visual inspection, one creates a graphical display of the data of interest and then interprets that data. A more precise definition of visual inspection can be difficult to provide because the nature of visual inspection depends on the question of interest. For example, one could choose to employ different visual inspection strategies to determine the presence or absence of a functional relation (e.g., if response rates are markedly changed [up or down] over successive sessions, if a behavior is under adequate stimulus control, or if a behavior is caused by one set of contingencies as opposed to another set (e.g., functional analysis [FA]; Iwata et al., 1982/1994). Often, visual inspection relies on the "interocular trauma test" (Edwards et al., 1963, p. 217) which is defined as a person knowing what the data mean when the interpretation hits them between the eyes. The common position within psychology is that the *rasion d'être* of data is for statistical analysis, but nuanced positions highlight the importance of visual inspection as a part of statistical analyses (Tufte, 1969).

One of visual inspection's main strengths is that it facilitates the interpretation of clinically meaningful results; in part, because the rater must be able to see marked changes in responding across baseline and treatment conditions with the "naked" eye (e.g., interocular trauma test; cf. Kazdin, 1982). Despite the utility of visual inspection in the applied analysis of behavior, authors have questioned the accuracy, validity, and reliability of visual inspection (also called interobserver agreement; IOA) for several reasons (Ottenbacher, 1990). First, interpretation accuracy may be impacted because human observers can be influenced by extragraphical contextual variables like participant characteristics (Grigg et al., 1989), treatment acceptability (Spirrison & Mauney, 1994), or the social significance of effects (DeProspero & Cohen, 1979). Interpretation accuracy may also be affected by graphical features—such as axes parameters—and the data itself. For example, Bengali and Ottenbacher (1998) explored whether the degree of autocorrelation (e.g., correlation between measures of the same behavior across sessions) influenced raters' visual interpretation of graphs. Raters were more likely to be influenced when the data were significantly autocorrelated, often resulting in incorrect conclusions. More recently, Levin et al. (2012) suggest that different degrees of autocorrelation may have a significant impact on false positives—when the clinician concludes a functional relationship exists even though one does not (i.e., Type I error). Levin et al. (2012) concluded that decision rules for visual inspection should be established to help clinicians interpret single-case research and, therefore, decrease human error impacting client outcomes.

Functional analyses are an area where visual inspection interpretations are particularly important. Functional analysis methodology is used to identify the function (or reason) a targeted problem behavior continues to occur by systematically altering environmental events and providing predetermined consequences immediately after the target behavior occurs. Standard FA protocol typically includes three test conditions, referred to as *alone, attention*, and *demand*; and one *control* condition. Assessment protocol descriptions can be found across the literature in introductory texts (e.g., Cooper et al., 2020), as well as a plethora of peer-reviewed research—starting with Iwata et al. (1982/1994).

Clinicians and researchers typically rely on multielement experimental designs to implement FAs; a feature that can make FAs difficult to interpret visually in contrast to other single-subject experimental designs that rely on extended, successive exposures to experimental conditions (e.g., ABAB—reversal) (Perone, 1991). The inherent complexity of the multielement experimental design leads to many overlapping data paths making it unsurprising that the reliability (Roane et al., 2013) and validity (Diller et al., 2016) of FA interpretations may be questionable at times. Ongoing issues with IOA are problematic because it supports the notion that "the process of visual inspection, would seem to permit, if not actively encourage, subjectivity and inconsistency" (Kazdin, 1982, p. 239). This potentially undermines the field's attempts in demonstrating the effectiveness of behavioral assessment and treatment, as well as increase the difficulty of substantiating its added value in the eyes of other multidisciplinary clinical team members. Within the field, the clinical implications of disagreement regarding visual analysis interpretation may prolong the assessment period (e.g., require repeating assessment conditions, introducing new, superfluous, assessment conditions) delaying treatment development and ultimately client improvement.

In response to some short comings of visual interpretation, researchers have developed formal decision rules for a variety of single-subject research designs including AB (Jones et al., 1978; Ottenbacher, 1990), ABAB (DeProspero & Cohen, 1979), and multielement designs (Bartlett et al., 2011; Danov & Symons, 2008; Hagopian et al., 1997; Lanovaz et al., 2019; Roane et al., 2013). We will focus on decision rules that apply specifically to standard FA methodology (usually multielement design). Specifically, Hagopian et al. (1997) developed and evaluated visual-inspection criteria designed to guide FA outcome interpretation and improve IOA. The criteria consist of a set of rules that operationalized some of the steps expert raters might take in evaluating an FA. One example is their general rule for assigning differentiation (i.e., identifying a causal mechanism for a behavior based on diverse levels of observed response rate across test conditions): "Differentiation is said to occur if at least five or more data points from a test condition fall above the upper criterion line than fall below the lower criterion line" (Hagopian et al., 1997, p. 324). The test condition is one of the possible conditions (e.g., tangible, demand, etc.) from the FA under consideration and the upper criterion and lower criterion lines are defined by the features of the data from the control condition. Their visual-inspection criteria led to improved IOA.

Roane et al. (2013) expanded the decision rules provided by Hagopian et al. (1997) to address some limitations of the original criteria. For example, the Hagopian et al. (1997) criteria were developed based on FAs with 10 data points per condition and the rules could only provide FA interpretations of a single outcome. Therefore, the Hagopian et al. (1997) criteria would not apply to FAs with fewer than 10 data points per condition and could not be used to correctly interpret that there are multiple maintaining variables for a problem behavior. Roane et al. (2013) modified the Hagopian et al. (1997) criteria so that there were rules that could be applied to FAs with varying lengths and for multiple maintaining variables. Roane et al. evaluated their modified decision rules by: (a) investigating IOA with and without the decision rules across two postdoctoral raters, (b) investigating IOA across reviewers without advanced experience in visual inspection, and (c) applying the visual inspection criteria to the categorization of FA outcomes. The authors reported improved IOA coefficients between experienced raters when they had access to the modified visual-inspection criteria compared to when they did not. They also reported improved agreement coefficients between inexperienced raters when they had access to the criteria as well as after participating in directed training on the guidelines.

Although researchers have demonstrated impressive improvements in IOA, these visual-inspection criteria may not adequately combat the influence that graphical and extragraphical contextual variables can have on human raters. Moreover, there may be opportunities to better capitalize on technology (i.e., computer programs). This may enable researchers to refine the criteria provided by Roane et al. (2013) by addressing some existing subjectivity or potential imprecision contained within those criteria. For example, one of the Roane et al. (2013) decision rules is "alone is the highest condition and is significantly higher than toy play" (p. 145). The authors are not referring to statistical significance but rather an interpretation of "significance" that is left to the rater. Thus, what is significantly higher to one rater may not be significantly higher to another. Assigning specific values for these sorts of cases may make the decision rules more precise and less subjective.

One benefit of enhancing decision rule precision, to accommodate automation, is that computers can then be used to aid human raters visual interpretation. With the advent of, and ongoing technological advancement of the personal computer, it has become increasingly easy to integrate the power of computer programming into all manner of human activities, including *supporting* human decision-making processes. Since the 1990's, many professions have been developing and refining specific programs to aid humans in decision making. These mathematical programs are designed to offset human error in situations that require considering many qualitative and quantitative variables (Ghodsypour & O'Brien, 1998). Clinical decision making in healthcare is an example of a field that has embraced decision support systems. Hospitals are increasingly incorporating decision support systems designed to *capitalize* on clinician expertise without replacing that expertise (Bowen et al., 2018). Specifically, these decision support systems provide clinicians with patient-specific assessment outcomes and recommendations that aid the clinician in their clinical-decision making, and ultimately can prevent adverse patient experiences (including clinical-decisions made leading to preventable patient deaths; Kawamoto et al., 2005).

Decision Support System Development

There are numerous steps that go into designing and building a decision support system. For example, the Tropos system is often used when developing decision support systems and it has many recursive steps (see Giorgini et al., 2004). A comprehensive overview on this topic goes beyond the scope of the current paper. Interested readers may consider reviewing Chai and Jiang (2011) or Sutton et al. (2020) for brief introductions and reviews of decision support systems. Essentially each phase of the decision support system development lifecycle repeats three stages: (a) inputs, (b) activities, and (c) outputs. For the current project, we provided an input (i.e., data from an FA), carried out the desired activity (i.e., run the computer script), and measured the output (i.e., determine whether the script output matched the experienced human raters' interpretation). If the script produced the right output (agreement with experienced human rater), then we carried on; if it produced the wrong output (disagreement with experienced human rater), we examined the disagreements and adjusted the script, where applicable. Generally, during initial program development decision support systems are created with input from experienced human clinicians so that the final product is a program aligning with best practice that balances objective, quantitative input (computer) with qualitative input (experienced clinician). In some cases, a series of experienced human clinicians teach the program how to respond and the outcome is then broadly used across relevant sites (hospitals, individual clinics, etc.). In other cases, these systems may incorporate human clinical expertise, as well as other criteria gleaned from large databases of existing patient characteristics, treatments, and corresponding outcomes. Ultimately, the computer output can be overridden by the clinical expert using the program. The intent is for the program to collate important quantitative information and provide an output that the clinician can use to inform decision-making.

In the field of behavior analysis, utilizing decision support systems in FA interpretation may be beneficial because a computerized script requires objective criteria (e.g., 10% higher vs. "substantially" higher) to make decisions, thus has the potential to minimize guideline subjectivity. Further, a script provides a fully replicable FA interpretation because the objective criteria are based only on features of the data. Given that a computer script is a deterministic process, if the data have not changed and the script has not changed then the script's outcome will always be the same. Additionally, the script can minimize extragraphical and graphical characteristics that influence human raters. Strictly, a script is not conducting a visual inspection but is evaluating the data, therefore the extragraphical and graphical features that influence a human rater are not considered by the script.

Creating a finalized version of a decision support system can be complex. Therefore, the current study's objective was to take the first steps toward creating a decision support system for FA interpretations using specific structured criteria (e.g., Roane et al., 2013). We explored whether it was even possible to create a computer script that would produce outcomes that matched experienced human raters' interpretation.

Method

Dataset

For this study, we used a dataset of 84 graphs depicting FA outcomes. Half of the dataset (42 graphs) was from published articles randomly selected from the *Journal of Applied Behavior Analysis* or *Behavior Modification* (a list of the articles can be made available upon request by contacting the first author). To minimize potential bias influencing article selection, we used a randomizer function in excel to select articles targeting problem behavior in persons with intellectual and developmental disabilities featuring an FA. After the random generator selected the articles, we screened each article to ensure it included at least one multielement FA with at least two conditions (i.e., control, test). The other half were collected by the first author for previous work evaluating the effects of psychotropic medication on problem behavior. We included these 42 datasets because they were readily available

Table 1. Sample Dataset Characteristics.

Characteristic	Percentage of cases
FA with four FA conditions	83%
FA with two, three or five FA conditions	17%
Single topography assessed	64%
Multiple topographies assessed	36%
Undifferentiated outcomes	38%
Automatic reinforcement outcomes	5%
Social positive reinforcement outcomes (attention)	6%
Social positive reinforcement outcomes (tangible)	10%
Social negative reinforcement outcomes	13%
Multiply-controlled outcomes (two functions)	14%
Multiply-controlled outcomes (three functions)	7%

and representative of what a clinician might observe in practice (e.g., behavior function was not always clear). All 84 graphs are available from the first author upon reasonable request, and the full anonymized dataset is hosted online at https://osf.io/86nz9/.

Collectively, the 84 FAs had heterogenous features, such as durations, conditions, and the order of presentation of those conditions. For example, the shortest FA had only eight sessions (two data points per FA condition) while the longest FA had 40 sessions (10 data points per FA condition). All FAs used a multielement design, and most employed procedures similar to those described by Iwata et al. (1982/1994). Thus, possible test conditions included alone, attention, demand, tangible, and control/toy play. Table 1 describes sample dataset characteristics, including: (1) percentage of cases with four FA conditions versus two three or five FA conditions, (2) percentage of cases wherein a single problem behavior topography was assessed versus multiple problem behavior topographies (e.g., problem behavior) assessed during an FA, and (3) percentage of cases depicting undifferentiated versus differentiated outcomes.

Procedure

First, two experienced raters individually applied Roane et al. (2013) structured criteria to determine the *clinical interpretations* of each graph. As such, there were 12 possible outcomes, including: social positive (access to attention); social positive (access to tangible); social negative (escape from demand); automatic; maintained by attention and escape; maintained by attention and tangible reinforcement; maintained by tangible and escape;

maintained by automatic and escape; maintained by automatic and attention; maintained by automatic and tangible; maintained by attention, tangible and escape; and undifferentiated. After interpreting graphs, raters reviewed datasets together to confirm agreement and resolve disagreements. The experienced raters were the first author, and her colleague (who was naïve to the current study's purpose). They had 8 and 11 years of experience conducting FAs and visually inspecting FA outcomes, respectively. Both experienced raters had practiced at the postdoctoral level for a minimum of 4 years and are Board Certified Behavior Analysts—Doctoral. We treated these clinical interpretations as the *final* interpretation of the FA data, thus we adjusted the script (described below) to match the clinical interpretations as closely as possible, where applicable. It is important to note that, prior to interpreting the graphs, a research assistant naïve to the study's purpose recreated each dataset in Prism GraphPad 6 and exported each graph to its own powerpoint slide in a randomized order. So, although the first author was not blinded to the original sources, the datasets were presented in a way that could not be linked back to the respective study participants or published articles. Following this, we divided the complete dataset evenly ($N = 84$) into two sample datasets. We used the first 42 graphs (here on referred to as the training dataset) to build and evaluate the script (described below). The latter 42 graphs (here on referred to as the testing dataset) were never used during the script-building process. We only used the testing dataset to evaluate the effectiveness of the script (described below) on a *novel* set of FAs.

Script Development

We developed the script using a process similar to how machine learning algorithms are developed. Specifically, we wrote the script and evaluated it on the training dataset. Once the script was completed, we tested the effectiveness of the script on the testing dataset. We designed the script to follow the modified visual-inspection criteria proposed by Roane et al. (2013). The script was custom written in R (R Core Team, 2019). R is a free, open-source statistical and analytic software package. The script also relied on the "tidyverse" package (Wickham, 2017). The "tidyverse" package adds many tools to the base R functionality that aid in data handling. For example, one of the key components of "tidyverse" is to extend the basic R functionality of data tables (i.e., base R data frames vs. "tidyverse" tibbles) to allow for faster analyses.

Figure 1 depicts a decision tree that could be used to describe the Roane et al. (2013) structured criteria and is an analogue of the process the script follows. However, it is important to note that the computer script does not strictly follow this decision tree (Figure 1). Instead, the script evaluates all

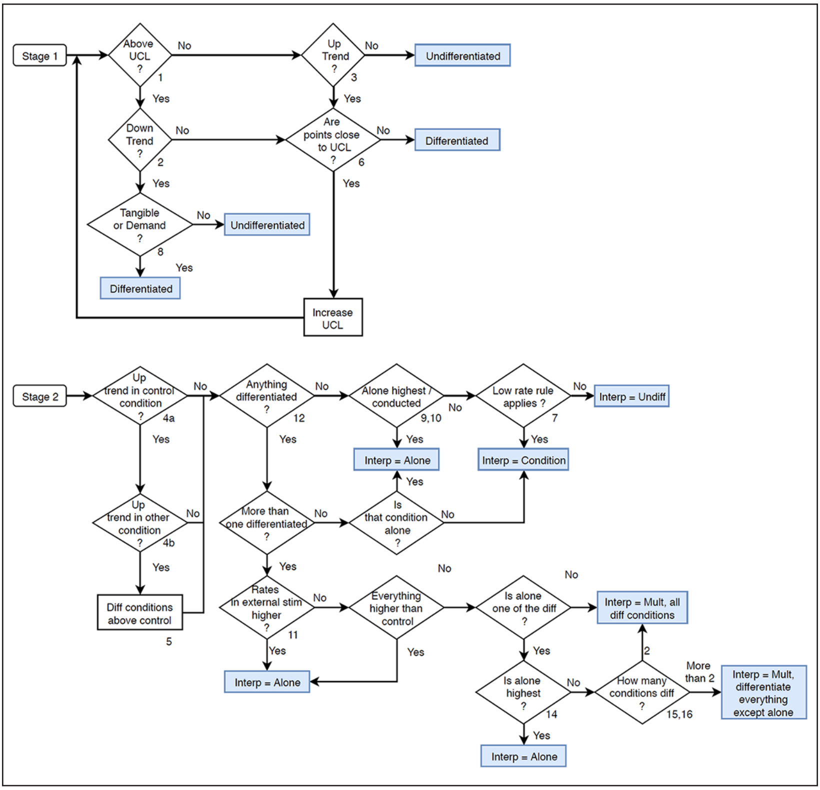
Stage 1
Above UCL ?
1
No
Up Trend ?
3
No
Undifferentiated
Yes
Yes
Down Trend ?
2
No
Are points close to UCL ?
6
No
Differentiated
Yes
Yes
Tangible or Demand ?
8
No
Undifferentiated
Yes
Differentiated
Increase UCL
Stage 2
Up trend in control condition ?
4a
No
Anything differentiated ?
12
No
Alone highest / conducted
9,10
No
Low rate rule applies ?
7
No
Interp = Undiff
Yes
Yes
Yes
Interp = Alone
Yes
Interp = Condition
Yes
Up trend in other condition ?
4b
No
More than one differentiated ?
No
Is that condition alone ?
No
Yes
No
No
Diff conditions above control
5
Rates in external stim higher ?
11
No
Everything higher than control
Is alone one of the diff ?
Interp = Mult, all diff conditions
Yes
Yes
Yes
2
Interp = Alone
Is alone highest ?
14
No
How many conditions diff ?
15,16
More than 2
Interp = Mult, differentiate everything except alone
Yes
Interp = Alone

Figure 1. A decision tree that is an analogue of the process the computer script followed. Open diamonds and squares represent decision points and blue rectangles represent a general outcome (differentiated or undifferentiated) or a final interpretation (function or undifferentiated). Stage 1 is completed for each condition. Stage 2 is completed once Stage 1 has been completed for each condition. The numerals listed near decision points correspond to Table 2 Decision Tree Companion. UCL is the upper criterion line, LCL is the lower criterion line.

possible states for each condition (e.g., it runs through every cell in Figure 1). To conceptualize this process we will provide an example. The first step depicted in Figure 1 is to determine for each condition if there are enough data points above the upper criterion line (UCL) without too many data points being below the lower criteria line (LCL). The criterion lines are calculated as defined in Roane et al. (2013) as the mean ± standard deviation of responding in the control condition. The next step relies on the answer to this first question. If, on the one hand, there are enough data points above the UCL, the next step is to determine if there is a downward trend. If, on the other hand, there are not enough data points above the UCL, then the next step is to determine if there is an upward trend. The script checks how many datapoints are above the UCL, if there is a downward trend, as well as if there is an upward trend even though upward and downward trends are incompatible. We also created Table 2, described below, which functions as a companion to Figure 1 because it provides detailed information about each cell in Figure 1.

From a computational perspective, a modern computer requires very little additional processing to determine every state of every condition of an FA outcome. For this reason, we chose to have the script evaluate every condition for every possible state. One benefit of having the script evaluate every condition for every possible state is that this helped flag *bugs*, as well as identify why the script sometimes disagreed with the clinical ratings. The full anonymized dataset and the version of the script that was used on the testing dataset is hosted online at https://osf.io/86nz9/.

Table 2 displays the natural language translations for our modified visual-inspection criteria for evaluating FA outcomes as they were coded into the script. As mentioned above, the table is a companion to Figure 1 because it identifies which cell in the decision tree applies to which rule and displays where we had to assign estimated values and rules for the modified visual-inspection criteria. For example, the UCL was explicitly defined as one standard deviation above the mean responding during the control/toy play condition (Roane et al., 2013). Such clearly defined rules and criteria could be coded directly into the script. Alternatively, rule 1 for determining automatic reinforcement was more ambiguous. The rule is that responding in the alone condition must be significantly higher than responding the control/toy play condition (Roane et al., 2013, p. 145). In this case, we defined significantly higher as the mean response rate for the alone condition must be 1.2 times greater than the control/toy play condition (e.g., at least 12 RPM for toy play if the alone condition was 10 RPM; see Table 2). We adjusted these estimates as necessary through the iterative training process until the script outcomes matched the experienced raters' interpretations. For clarity, we will refer to these adjustments as *estimated rules*. Finally, a small number of *additional rules* were added to the script for items that may be obvious to a human rater but must be explicitly written for a computer system (Table 2). For example, if a condition within an FA only had three data points and all of them are above the UCL then it should clearly be considered differentiated. However, if there is only one data point in the second half of the FA then the

Table 2. Visual-Inspection Decision Tree Companion.

Item	Roane et al. (2013) rule description	Computer script description (assigned objective criteria or programmed action)
Stage 1		
1	50% or more of the data points from a test condition fall above the UCL than fall below the LCL	Same as Roane et al. (2013).
2	≥40% of data points above UCL must occur in the second half of the assessment	(a) Same as Roane et al. (2013), *AND (b) must have at least one high point (anything above UCL) anywhere AND (c) do not check for downward trend if 100% of data points are above UCL*
3	50% of all data points fall above the UCL in the last half of the assessment, and last data point is above UCL	(a) Same as Roane et al. (2013) *AND (b) don't check if 100% of data points above UCL*
4	"Trend in toy play and most of the other conditions, any condition that is consistently higher than toy play over the assessment course"	(a) There is an upward trend in control condition, AND (b) 75% of test conditions within the assessment meet Item 2, AND (c) median response rate for each condition is greater than or equal to mean response rate for control condition (i.e., is the test condition in question higher than toy play)
5	All conditions high and relatively stable, no overall trends	Stability is operationalized as a range within 0.33 * maximum of rate of response for toy play condition. (a) Only apply relative stability rule if stability obtained (see above), AND (b) responses in >50% of all sessions in that condition HAVE >0 responding, AND (c) none of the trend rules (see Item 1, 2, 3) are met

(continued)

Table 2. (continued)

Item	Roane et al. (2013) rule description	Computer script description (assigned objective criteria or programmed action)
6	Condition meets differentiation criteria (Item 1) but $>$10% data points above UCL by only small amount. Raise UCL by 20% for that condition. Exception: If, at least 50% of data points above UCL in last half of FA, do not adjust UCL and apply upward trends rule	(a) Same as Roane et al. (2013) criteria, AND (b) if there is one data point above UCL by only 20% raise UCL by 20% for that condition only, AND *(c) do not test adjusted UCL for upward trend*
7	Most data points at near-zero levels, condition in which highest rate of behavior occurs is differentiated. One of the high data points must occur in last half of assessment	(a) 50% of control conditions must be near zero (near zero is anything less than or equal to 20% of the UCL) to trigger this rule, AND (b) 40% of all data points must be below UCL * 0.2 (across all conditions), AND (c) one high data point must occur in last half of assessment If (a), (b) and (c) are met, AND (d) $>$50% of high-rate sessions occur in same test condition, AND (e) $\geq$50% of total number of behavior in higher-rate sessions occur in that same condition AND (f) at least one session must be above the UCL in the last half of FA AND (g) condition does not have a downward trend
8	Tangible and demand exclusion from downward trend: Do not apply the downward trend rule if there is a decreasing trend to an efficient rate of responding	(a) All the response rates in second half of FA, must be greater than 2 RPM, AND (b) response rate in last session must be $<$2.5 but $>$2, AND (c) each session must have lower response rate than preceding session, AND (d) do not consider sessions prior to median point of functional analysis

(continued)

Table 2. (continued)

Item	Roane et al. (2013) rule description	Computer script description (assigned objective criteria or programmed action)
Stage 2		
9	No Roane et al. equivalent	*Check that Alone condition was conducted*
10	Alone is highest and significantly higher than toy play	*(a) Any test condition does not need to meet differentiation to check for Alone rules*, AND (b) alone has the highest mean rate of response (<0.01), AND (c) mean rate of response in alone is 20% greater than mean responding in control condition, AND (d) there is no downward trend in alone (Item # 3)
11	Rates of behavior tend to be higher in conditions with less external stimulation than in condition with higher external stimulation	(a) 90% of the low stimulation data points must be higher than the highest high stimulation data point, AND (b) *the response rate in low stimulation condition must be greater than 0, AND (d) assessment must include alone condition for this rule to be triggered*
12	All conditions high and relatively stable, no overall trends	Stability is defined as a range within 0.33 * maximum of rate of response for toy play condition (a) Only apply relative stability rule if stability obtained (see above), AND (b) responses in >50% of all sessions in that condition HAVE >0 responding, AND (c) none of the trend rules (see Item 1, 2, 3) are met (same as Item # 5)
13	Highest condition is alone	Highest mean rates of response out of the differentiated test (<0.01)
14	Four differentiated conditions, and alone is differentiated but not the highest	At least one other test condition mean responding > mean responding in alone condition (by <0.01)

(continued)

Table 2. (continued)

Item	Roane et al. (2013) rule description	Computer script description (assigned objective criteria or programmed action)
15	Three differentiated conditions, and alone is differentiated by not the highest	At least one other test condition mean responding > mean responding in alone condition (by <0.01)
16	Two differentiated conditions, but alone is the lower of the two	Mean responding in other test condition > mean responding in alone condition (by 0.01)

Note. Text in italicized font describe script "bug checks." UCL refers to the upper criterion line. LCL refers to the lower criterion line.

condition also technically meets the requirement for a downward trend because less than 40% of the data points above the UCL occur in the second half of the assessment (Roane et al., 2013, p. 145). In this case, an additional rule was added to the script so that downward trends are ignored if all the data points were above the UCL and no data points fell below the LCL.

We adjusted the script data handling, previously defined rules, estimated rules, and additional rules for the computer based on an iterative process. For each FA in the training dataset, we used the script to evaluate the data. If the script outcome matched the experienced raters' interpretation, we did not make any changes to the script. If the script outcome did not match the experienced raters' interpretation, we first identified the cause of the disagreement then adjusted the script as necessary. Some reasons for disagreements included: (1) data handling bugs in which the proper data was not being used, (2) mathematical errors, and (3) inaccurate estimated rules. Bugs and mathematical errors that were not directly related to the modified visual-inspection criteria (e.g., accidentally including the wrong mathematical operator, incorrectly specifying if. . .then conditionals) were corrected. If the script was working correctly but providing an outcome that did not match the experienced human raters' interpretation because the visual-inspection criteria were being applied incorrectly or an estimated rule was inaccurate then we discussed the cause of the error, possible solutions, and corrected the script. If the structured criteria were being applied incorrectly, we fixed the error. If the estimated rules were producing the error, we adjusted the value of the estimated rule.

We corrected the script until we were satisfied with the outcome it produced for a given FA. There were three reasons we would be satisfied with

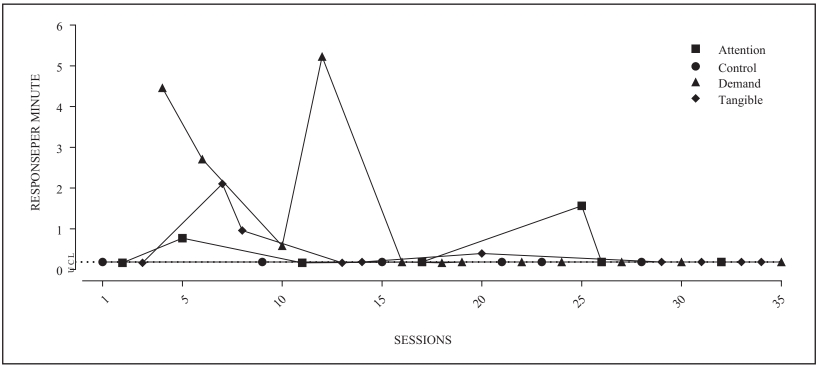

Figure 2. Upper criteria line (UCL) is near zero and represented by a dotted line. UCL label is vertical along the y-axis. For this graph (FA 44 in the full dataset), disagreement was unanticipated because there was a feature of the data that was not explicitly covered by the modified visual-inspection criteria (a control condition with non-zero rates of behavior and no variance), which would not likely influence a human rater's interpretation but could not be handled by the script.

the outcome the script provided. First, and most common, was if it matched the experienced raters' interpretation of the FA data. The second reason we accepted the script's outcome was if there was a disagreement between the script's outcome and the experienced raters' interpretation caused by a difference in how the modified visual-inspection criteria were applied. For example, in one FA a data point was extremely close to the UCL as plotted on a figure, so it was counted by the experienced raters as above the UCL. However, the data point was in fact slightly below the UCL. Therefore, the script did not count that data point as above the UCL. In this case, that one data point was enough to lead to a disagreement the script outcome and the experienced raters' interpretation. The final reason we would accept the script's outcome (over the experienced raters' interpretation) was if we determined the disagreement was due to a special case in the data that was not explicitly addressed by the modified visual-inspection criteria. For example, in one graph, as seen in Figure 2, the LCL and UCL were both 0.2. The Roane et al. (2013) decision rules provide a recommendation of what to do when the UCL and LCL are both zero but not what to do when they are equal and non-zero. If we extended the Roane et al. rule for establishing UCL and LCL lines when they are both equal to zero, the data points that are 0.2 would be considered below the LCL[1]. However, without extending the Roane et al. rules there were mathematical issues interfering with the correct application of the criteria. When we accepted that the script outcome did not match the clinical interpretation because of a special case in the data then the outcome was considered the “final answer” from the script.

We repeated the process of using the script to evaluate each FA in the training dataset several times. In the first few passes through the data, changes to the script were made on demand to ensure that rewriting or adding more code produced the desired changes in the script outcome. That is, when we detected a problem with data from an FA, we adjusted the script until that problem was eliminated for that specific FA without determining whether this adjustment would affect the script’s outcome for other FAs. With later adjustments to the script, we were more deliberate in how changes were made. Specifically, we assessed how one solution impacted the script’s outcome for the other FAs. For example, changing one of the estimated rules to get an outcome that matched the clinical interpretation for FA A might inadvertently change the outcome for FA B that also relied on that estimated rule. For this reason, we repeated the process of adjusting the script and assessing the script’s outcomes until agreement between script outcomes and clinical interpretations were optimal (78% agreement, see Results section below). When we had optimized agreement between script outcomes and clinical interpretations, in accordance with the procedures described above (e.g., adjusting estimated rules to obtain agreement; accepting script outcome over clinical interpretation due to special case in the data), we proceeded to test the script.

Testing the script

After completing the script with the training dataset, we used the script to evaluate all 42 FAs in the testing dataset. During the testing phase, we did not make any adjustments to the script based on the script’s outcomes, except to fix bugs that surfaced (see below). If the script provided an outcome that did not match the experienced raters’ interpretation, then we investigated the cause of the disagreement and categorized it into one of three groups. The script outcome and experienced raters’ interpretation were categorized as a *disagreement* if the rules were followed differently by the script and the experienced raters’, or if an estimated rule lead to a disagreement. Figure 3 illustrates an FA outcome where the experienced raters interpreted the FA as automatic while the script outcome was undifferentiated due to an estimated rule wherein 90% of the low stimulation conditions (e.g., alone, attention, etc.) within the FA need to be above the maximum response rate across all high stimulation conditions (see Table 2). Response rates did not meet this criterion—so the computer assigned an undifferentiated outcome.

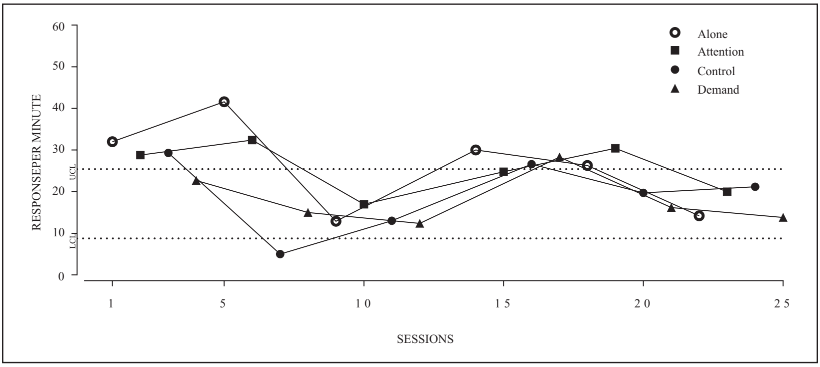

Figure 3. Upper criterion line (UCL) and Lower criterion line (LCL) are represented by upper and lower dotted lines. Both are labelled in vertical text along the y-axis. For this graph (FA 1 in the full dataset), there was disagreement between experienced raters' interpretation and script outcome due to an inaccurate estimated rule. The experienced raters' interpretation was that the behavior was automatically maintained, the script's outcome was that the FA data was undifferentiated.

A disagreement was classified as *unanticipated* if there was a feature of the data that was not explicitly covered by the modified visual-inspection criteria that would not interfere with an experienced rater generating an interpretation but could not be handled by the script (e.g., Figure 2). In Figure 2, an experienced rater would likely have observed no response variability in the control condition and that responding was at the floor. They could have proceeded to treat the UCL and LCL as being zero and assign function accordingly. However, the script could not generate an outcome because there was not way to handle the fact that the UCL and LCL were equal but not equal at 0 as per Roane et al. criteria (see Results section for fulsome description of this scenario). Finally, a disagreement was classified as a bug if we identified the difference was caused by a data handling error in the script and not the result of an application of the decision rules by the script.

Results

As mentioned above, there were data from 42 FAs in each of the training and testing datasets. For the final assessment of the FAs in the training dataset we did not make any changes to the script and obtained 78% (33/42) agreement between the script outcome and experienced raters' interpretation. Eight of the nine disagreements between the script outcome and experienced raters'

Table 3. Training Dataset Disagreements and Rationale.

Graph	Corresponding figure	Experienced raters'	Script response	Rationale
1	Figure 3	Automatic	Undifferentiated	Estimated values may be incorrect
2	—	Escape/Tangible	Tangible/Auto	Script stringently applied inspection criteria
4	—	Escape/Tangible	Escape	Script stringently applied inspection criteria
6	Figure 4	Escape/Tangible	Escape	Script stringently applied inspection criteria
7	—	Undifferentiated	Automatic	Script stringently applied inspection criteria
21	—	Undifferentiated	Escape	Script stringently applied inspection criteria
25	—	Escape	Automatic	Script stringently applied inspection criteria
26	—	Atten/Escape/Tang	Atten/Tang	Script stringently applied inspection criteria
38	—	Escape/Tangible	Escape	Script stringently applied inspection criteria

Note. Atten = Attention; Auto = Automatic; Tang = Tangible.

interpretation of the training dataset resulted from differences in how stringently the script followed the modified visual-inspection criteria relative to the experienced raters. For the ninth disagreement (see Figure 3), we were unable to adjust our estimated rule so that it would produce an outcome matching the experienced raters' interpretation. Table 3 provides a disagreement summary for the training dataset and an abbreviated explanation for discrepancies.

For the testing dataset, we ran the script and identified two bugs and one dataset that the structured criteria could not account for. The bugs were

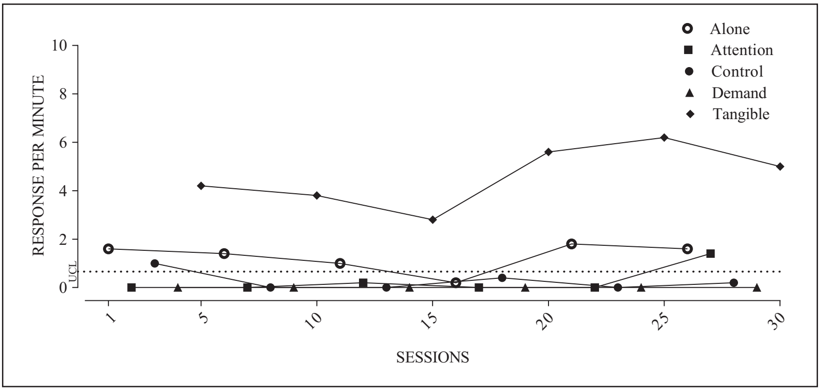

Figure 4. Upper criterion line (UCL) is represented by vertical dotted lines, labelled in vertical text along the y-axis. Lower criterion line is 0 and is not included as a graphical feature. This graph (FA 3) is an example of an agreement between experienced raters' interpretation and script's outcome obtained by bending rule 2d, in the section Rules for Multiple Maintaining Variables (Roane et al., 2013, p. 146). The experienced raters' interpretation was that the behavior was multiply controlled by access to tangible and automatic reinforcement and the original script outcome indicated access to tangible.

caused by data handling errors in the script, one was the result of unexpected data pattern and the other was the result of one of the rules being incorrectly coded. For the unexpected data pattern, a test condition only occurred twice in the first half of the FA and did not occur again; presumably because there were other clear conditions in which responding was occurring and the researchers terminated further testing of that condition. In the criteria, to determine if there is an upward or downward trend in the data you must look at the number of data points in the second half of the FA. Under such conditions, a human rater would ignore the trend rules for that condition because they clearly don't apply. The script always evaluates all the rules for every condition, but because there were no data points in the second half, this led the script to a "divide-by-zero error"; resulting in no output. We added code to handle these special cases, wherein we told the script to automatically score the trends as 0 if there are no data points in a condition during the second half of the FA. The second error occurred because the code that we wrote to account for multiple maintaining variables and four conditions are differentiated (rule 2[b]; Roane et al., 2013, p. 146) did not implement the rule correctly. The rule states that in such cases, automatic reinforcement should be excluded from the final interpretation as one of the multiple controlling variables; the script was not excluding that condition when the rule was evaluated. None of the FAs in the training dataset met the conditions for this rule to be the final determination of an output, therefore we did not detect the coding error until the testing dataset.

The dataset that structured criteria could not account was something that a human rater could easily ignore and apply the rules but caused an error for the computer script. For this FA data (Figure 2) there was a floor of responding at 0.2 RPM. Additionally, every session of the control condition was at this floor; therefore, the mean was 0.2 and the standard deviation was 0. As the UCL and LCL are defined as the mean ± the standard deviation, both the UCL and LCL were 0.2. In the other conditions there were other sessions at this artificial floor of 0.2 and those sessions were included as above the UCL and below the LCL because the rules are data points greater/less than or equal to the criterion line. The structured criteria have a rule to account for the UCL and LCL both being 0, but the rule was not general enough to account for UCL being equal to the LCL and both being non-zero. An experienced rater would have looked at a control condition having no variability and being at the floor on a figure and treated it as the UCL and LCL being zero and followed the structured criterion as such. During our investigation to explain this error, we manually altered the UCL so that it was slightly above the LCL ($0.2 + 10^{-9}$) and the script output matched the experienced raters' interpretation.

After the bugs were corrected, the resulting script outputs matched the experienced raters' interpretation. So, when the script was no longer being terminated prematurely (i.e., it could run through all rules) we observed agreement between the output and experienced raters' interpretations. Reclassifying the two FAs in which bugs in the script were identified resulted in an agreement level of 83%. If we altered the structured criteria to account for the oversight of equal but non-zero criteria lines the agreement rate was 85%, although we will focus on the more conservative 83% agreement rate. Table 4 displays the number of disagreements and accompanying abbreviated rationale for the testing dataset.

Examining the disagreements in the testing dataset more closely, two of the disagreements were caused by differences in how stringently the script followed the structured criteria relative to the experienced raters' (Figure 5) and four of the disagreements were caused at least partially by estimated rules for the script leading to a script outcome that did not match clinical interpretation (e.g., Figure 3). Regarding the strictness of following the rules, Figure 6 displays the data from FA 54 as an example. The experienced raters' interpretation of this graph was that the behavior was multiply controlled by

Table 4. Testing Dataset Disagreement Types and Rationale.

Graph	Corresponding figure	Experienced raters'	Script response	Rationale
44	Figure 2	Undiff	Atten/Esc/Tang	Existing criteria rules may be problematic
53	Figure 6	Undiff	Automatic	Script stringently applied inspection criteria
54	Figure 5	Atten/Tang	Atten/Escape/ Tang	Existing criteria rules may be problematic
59	—	Undiff	Escape	Script stringently applied inspection criteria
73	—	Automatic	Undifferentiated	Script stringently applied inspection criteria
81	—	Undiff	Attention	Script stringently applied inspection criteria
83	—	Undiff	Automatic	Existing criteria rules may be problematic

Note. Atten = Attention; Esc = Escape; Tang = Tangible; Undiff = Undifferentiated.

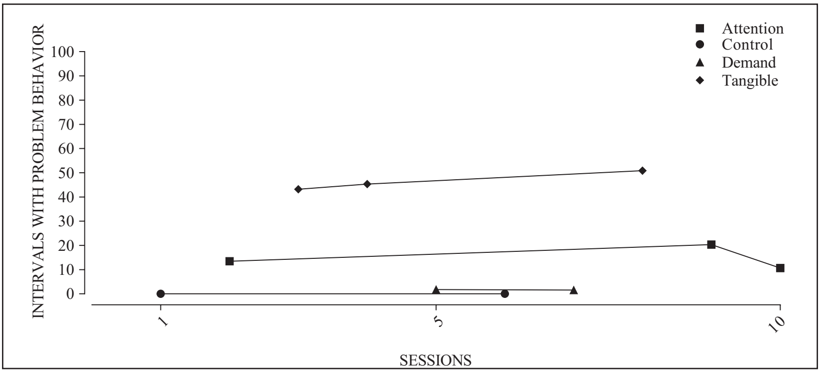

Figure 5. Responding in control condition is zero, therefore upper and lower criterion lines are 0. For this graph (FA 54 in the full dataset), there was disagreement between the experienced raters' interpretation and script outcome resulting from the script's stringent application of the criteria. The experienced raters' interpretation was that the behavior was multiply controlled by access to tangible and attention, the script's outcome was that the behavior was multiply controlled by access to tangible, access to attention, and escape from demand.

access to attention and access to tangible and the script's outcome was that the behavior was multiple controlled by access to attention, access to tangible, and escape from demand. The experienced raters' interpretation did not include escape from demand as a maintaining variable because it is a low-magnitude effect. The modified visual-inspection criteria defined a low-magnitude effect occurring when a condition becomes undifferentiated if the UCL is increased by 20% (cf. rules for low magnitude effects; Roane et al., 2013, p. 146). For example, if 100% of the points are above the UCL for a condition but only 25% of the points are above an adjusted UCL then there would be a low magnitude effect. In the case Figure 5, there was no target behavior in the control/toy play condition therefore the UCL was set to 0 (cf. placing the criterion lines; Roane et al., 2013, p. 145). As a result, the adjusted UCL was 1.2 times higher than 0 RPM. According to the strict interpretation of the Roane et al. (2013) modified visual criterion that the script uses, if there is no target behavior in the control condition then a low magnitude effect is not possible.

Figure 6 displays data from FA 53 and depicts a disagreement between the script outcome and the experienced raters' interpretation caused by an estimated rule. The experienced raters' interpretation of the target behavior was undifferentiated and the script's outcome was automatic reinforcement. For the experienced raters, target responding appeared highly variable (i.e., no stable trends) with data paths that consistently overlapped across conditions indicating an undifferentiated FA. The script's outcome of automatic reinforcement was in response to an estimated value for the automatic reinforcement rule 1 in the Roane et al. (2013) modified visual criteria. The defined rule is: "alone is the highest condition and is significantly higher than toy play. Define highest by calculating the mean of each condition" (Roane et al., 2013, p. 145). To write this rule into the script we had to estimate a value for "significantly higher than toy play/control". Our estimated rule was that for responding in the alone condition to be considered significantly higher than responding in the control condition, responding in the alone condition must be greater than or equal to 120% of the responding during control/toy play condition. This estimated value (120%) lead to the script providing an outcome that did not match the experienced raters. For this FA, the mean rate of response for the alone condition was 16.0 which was the highest mean rate of response. The mean rate of response for the control condition was 12.8 (which, when multiplied by 1.2 was only 15.36 RPM). Therefore, the script determined that the alone condition had the highest rate of response and it was significantly greater than the rate of response in the control/toy play condition.

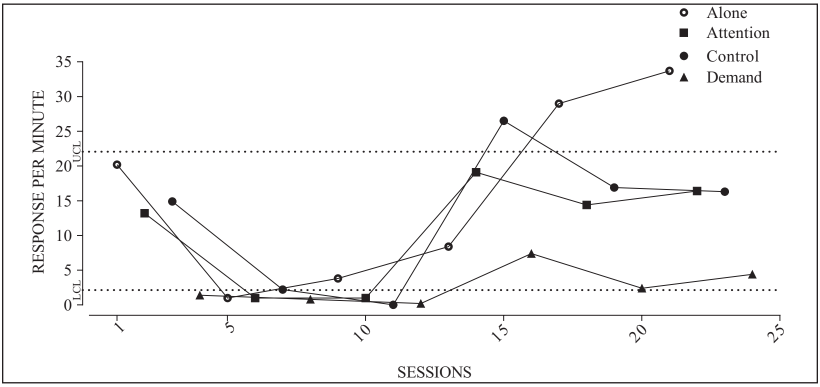

Figure 6. Upper criterion line (UCL) and lower criterion line (LCL) are represented by upper and lower dotted lines. Both are labelled in vertical text along the y-axis. For this graph (FA 53 in the full dataset), there was disagreement between experienced raters' interpretation and script outcome due to an inaccurate estimated rule. The experienced raters' interpretation was that the FA data was undifferentiated, and the script's outcome was that the behavior was maintained by automatic reinforcement.

Discussion

We reported acceptable agreement levels (Cooper et al., 2020) between the computer script outcome and experienced raters' interpretations for the training datasets (78%), for the testing datasets (83%), for an overall average agreement of 81% across training and testing datasets. The results clearly demonstrate that it is possible to create a computer script (based on specific structured criteria) to evaluate FA data and produce an output matching the clinical interpretation of experienced human raters. However, there is ample room for script improvement, as well as some limitations to this specific proof of concept.

It is important to identify future areas to improve the script. One area includes the vague decision rules in the Roane et al. (2013) modified visual-inspection criteria, which required us to create estimated rules in the script and represents one possible explanation for disagreements. Notably, our primary objective was to determine whether it was possible to develop a computer script to evaluate FA data and produce outcomes with an acceptable level of correspondence with clinical interpretations based on a specific structured-criteria (i.e., Roane et al., 2013 modified visual-inspection criteria). Therefore, we made every attempt to avoid adjusting the criteria; only

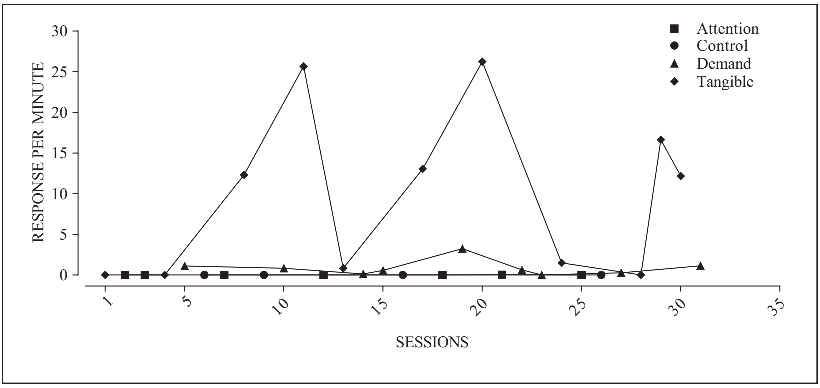

Figure 7. Responding in control condition is zero, therefore upper and lower criterion lines are 0. This graph (FA 6 in the full dataset) depicts disagreement between experienced raters' interpretation and script outcome due to a special case in the data that the modified visual-inspection criteria does not explicitly address. The experienced raters' interpretation was that the behavior was multiply controlled by access to tangible and escape from demand and the script's outcome was escape from demand.

making adjustments (i.e., create estimated rules) when absolutely necessary (see Table 2 for specific examples). We encourage researchers to consider systematically evaluating adjustments and empirically validating existing vague decision rules, as well as their resultant manufactured values to determine which values may produce the most accurate outcomes (i.e., highest agreement between human experienced raters' interpretations and computer script outcomes). This proposed objective went beyond the scope of the current project. However, this type of follow up work could eliminate existing vague rules that could be contributing to Type I and Type II errors (e.g., false positive, false negative) in the literature, in clinical practice and may have been contributing to instances of disagreement in the current project.

Another explanation for observed disagreements is related to the fact that the script stringently followed the rules. One instance where this was clearly depicted was in FA 6 (see Figure 7; the experienced raters' interpretation was that the behavior was multiply controlled by escape from demand and access to tangible and the script's outcome was escape from demand only). In this situation, excluding a function-based intervention component targeting access to tangible would be ill-advised because during the FA there were repeated sessions of extremely high rates of responding during the tangible condition relative to all other conditions. However, the Roane et al. (2013)

modified visual criteria indicate that only escape from demand is the correct interpretation because too many data points in the tangible condition fell below the LCL. Specifically, eight of nine demand data points were above the UCL, and one was below the LCL, resulting in 78% above the UCL. For tangible, responding was highly variable and eight of 11 data points fell above the UCL, while three data points fell below the UCL. Thus, only 45% of data points fell above the UCL (as defined in the Roane et al. criteria). Also, there were only five of 11 data points above the UCL in the second half, so an upward trend was absent. Scenarios like this one are valuable because they clearly highlight existing areas for decision rule improvement and emphasize the value added of a decision support-system wherein, computers augment clinician decision-making rather than supplanting them entirely. Once again, we encourage future research to explore, refine and validate decision rules that produce contradictory outcomes to make improvements to these helpful, important guidelines. Doing so could increase clinical utility and implementation ease (through semi-automation). We also encourage future researchers to develop and evaluate other "standardized" decision rules suited for FA variations (e.g., latency-based, Interview-Informed Synthesized Contingency Analysis).

Although we attempted to retain Roane et al. (2013) rules as much as possible, there were times when we needed to bend the rules slightly (and adjust the script accordingly) to produce agreements between the experienced raters' interpretations and the script outcomes. For example, in Figure 4 (FA 3), the script originally produced an access to tangible outcome. The experienced raters had interpreted this FA as multiply controlled by access to tangible and automatic reinforcement, according to rule 2d, in the section titled "Rules for Multiple Maintaining Variables" (Roane et al., 2013, p. 146). This rule states that when two conditions are differentiated, one of which is alone, and alone has lower responding, the interpretation is multiply controlled with automatic and the other condition. Originally, the script did not consider responding in the alone condition to be differentiated because of a downward trend which Roane et al. (2013) defined as "at least 40% of the data points above the upper CL must occur in the second half of the assessment" (p. 145). To produce agreement, we bent this rule by directing the computer script to flag a downward trend only if less than 40% of the data points fell above the UCL in the latter half of the assessment.

Study Limitations

One possible study limitation is that our experienced raters did not accurately apply the guidelines, which could have impacted average agreement level (81%). However, our two experienced raters were highly qualified and very familiar with applying the Roane et al. (2013) criteria. Therefore, we are confident that this potential error source minimally affected our outcomes, allowing us to create an optimal computer script based on the modified visual-inspection criteria. Additionally, we did not collect IOA data on individual human rater interpretation, rather relied on a consensus process to establish final clinical interpretations. This process is often observed in clinical practice, for example two clinicians meet during case formulation to come to a consensus on behavior function (interpret FA data) prior to developing a treatment plan. Given our objective was to determine whether a computer script could produce outputs matching human clinical interpretations, combined with the fact that our raters were highly skilled, we do not feel this short coming (no IOA between human raters) undermined our objective.

Another possible study limitation is that the computer script, in its current form, is not ready to be deployed for wide use by practicing behavior analysts. This is in part because our agreement coefficient was only 81%, which suggests some room for improvement. However, we are not suggesting perfect experienced rater-computer script agreement (based on a specific structured criteria) is the goal because creating a computer script that has a 100% agreement rate indicates that it is probably over specified. For example, if a computer script is overly precise in that it is "perfect" a behavior analyst who wants to modify the decision rules would drastically affect that perfect computer script. Notably, any script or program to evaluate FA data should always be used as a decision support system and should not be used to supplant human clinical expertise. Finally, using the computer script in its current form requires the user to be fluent in R because it is only a set of mathematical rules applied to a specifically formatted dataset. In the future, this system could be incorporated into a program where a user can input their data and the computer script produces an output to inform the user—who can then use this information to make the final clinical decision. There are other successful demonstrations of creating a freestanding software package for behavior analysts (e.g., Gilroy et al., 2017). The current study is a proof-of-concept demonstrating that it is possible to develop a script based on existing modified structured criteria (Roane et al., 2013) to evaluate single-subject research designs, more specifically FAs.

Study Implications

Behavior analytic researchers have worked diligently over the years to improve assessment efficiency and accuracy by developing, refining and validating FA variations (e.g., brief FA, Derby et al., 1992; latency-based, Thomason-Sassi et al., 2011; FA screening, Querim et al., 2013). This work has invariably impacted the clinical utility of FA technology, minimizing assessment time and maximizing accurate and reliable outcomes produced by using an FA variation. It is curious, however, that the field seems to have largely neglected exploring ways to improve interpretation efficacy through semi-automation; despite behavior analytic literature showcasing attempts to improve the accuracy and reliability of traditional visual inspection through creating systematic guidelines (Bartlett et al., 2011; Danov & Symons, 2008; Hagopian et al., 1997; Lanovaz et al., 2019; Roane et al., 2013). We are not arguing for a single system to replace the expertise of behavior analysts nor are we arguing for this (or any) script to be the final determinant for interpreting single-subject research. In fact, full reliance on a set of rules or script can "force acceptance of problematic data" (Johnston & Pennypacker, 2009, p. 308) such as the script ignoring the importance of tangible reinforcers in Figure 7. These sorts of systems and scripts are designed to assist in the decision making of behavior analysts or potentially to aid in the training of new generations of behavior analysts.

The importance of improving efficiency across all aspects of behavior analytic practice should not be underestimated—especially considering a recent report released by the Behavior Analyst Certification Board Incorporated (2018) illustrated the extensive employment demand for qualified clinicians. Arguably demand exceeds supply, especially in remote areas. Thus, achieving greater efficiencies across all facets of behavior analytic work, including semi-automating FA interpretation, could improve service accessibility.

Semi-automating FA interpretation may also have the added benefit of removing the need for obtaining IOA (i.e., secondary independent observers to corroborate interpretation outcomes) because an automated use of structured criteria is consistently applied each time according to the computer script. With regards to manuscript submissions, interpreting FA outcomes during the peer-review process represents an IOA of sorts. This process could be expedited, or at least standardized, if a consistent set of rules for interpreting single-case designs were established and applied to the results by a computer script as part of the submission process. If there is clear evidence that visual-inspection criteria or a script are providing an incorrect interpretation, researchers and clinicians will provide some justification of their interpretation rather relying on arguments that boil down to "differences of opinion." Finally, it is possible that developing and implementing this type of a process may improve the acceptability of single-case design research outcomes across other disciplines, given one of the primary criticisms is the subjectivity involved in interpretation.

Overall there are many reasons to explore semi-automation in the context of FA interpretation. We strongly encourage future research to explore this relatively understudied area so that the benefits may be fully actualized.

Authors' Note

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention. Jonathan Friedel is now at Georgia Southern University.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Alison Cox https://orcid.org/0000-0001-5396-6122

Jonathan E. Friedel https://orcid.org/0000-0002-1516-330X

Note

[1] Note that we identified this example based on FA data in the testing dataset (FA 44). Therefore, this specific modification to the rule was not implemented in our script.

References

Bartlett, S. M., Rapp, J. T., & Henrickson, M. L. (2011). Detecting false positives in multielement designs: Implications for brief assessments. *Behavior Modification*, *35*(6), 531–552. https://doi.org/10.1177/0145445511415396

Behavior Analyst Certification Board Incorporated. (2018). *US employment demand for behavior analysts: 2010–2017*. https://www.bacb.com/wp-content/uploads/2020/05/US-Employment-Demand-for-Behavior-Analysts_2020_.pdf

Bengali, M. K., & Ottenbacher, K. J. (1998). The effect of autocorrelation on the results of visually analyzing data from single-subject designs. *American Journal of Occupational Therapy*, *52*(8), 650–655.

Bowen, M. E., Bhat, D., Fish, J., Moran, B., Howell-Stampley, T., Kirk, L., Persell, S. D., & Halm, E. A. (2018). Improving performance on preventive health quality measures using clinical decision support to capture care done elsewhere and patient exceptions. *American Journal of Medical Quality*, *33*(3), 237–245. https://doi.org/10.1177/1062860617732830

Chai, Z., & Jiang, H. (2011, December). *A brief review on decision support systems and it's applications*. Paper presented at the 2011 IEEE International Symposium on IT in Medicine and Education, Cuangzhou, China.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2020). *Applied behavioral analysis* (3rd ed.). Pearson.

Danov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification*, *32*(6), 828–839. https://doi.org/10.1177/0145445508318606

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, *12*(4), 573–579. https://doi.org/10.1901/jaba.1979.12-573

Derby, K. M., Wacker, D. P., Sasso, G., Steege, M., Northup, J., Cigrand, K., & Asmus, J. (1992). Brief functional assessment techniques to evaluate aberrant behavior in an outpatient setting: A summary of 79 cases. *Journal of Applied Behavior Analysis*, *25*(3), 713–721. https://doi.org/10.1901/jaba.1992.25-713

Diller, J. W., Barry, R. J., & Gelino, B. W. (2016). Visual analysis of data in a multielement design. *Journal of Applied Behavior Analysis*, *49*(4), 980–985. https://doi.org/10.1002/jaba.325

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. https://doi.org/10.1037/h0044139

Ghodsypour, S. H., & O'Brien, C. (1998). A decision support system for supplier selection using an integrated analytic hierarchy process and linear programming. *International Journal of Production Economics*, *56*, 199–212. https://doi.org/10.1016/S0925-5273(97)00009-1

Gilroy, S. P., Franck, C. T., & Hantula, D. A. (2017). The discounting model selector: Statistical software for delay discounting applications. *Journal of the Experimental Analysis of Behavior*, *107*(3), 388–401. https://doi.org/10.1002/jeab.257

Giorgini, P., Kolp, M., Mylopoulos, J., & Pistore, M. (2004). The tropos methodology: An overview. In Bergenti, F., Gleizes, M., & Zambonelli, F. (Eds.), *Methodologies and software engineering for agent systems*. Kluwer Academic Publishers.

Grigg, N. C., Snell, M. E., & Loyd, B. (1989). Visual analysis of student evaluation data: A qualitative analysis of teacher decision making. *Journal of the Association for Persons with Severe Handicaps*, *14*(1), 23–32. https://doi.org/10.1177/154079698901400104

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*(2), 313–326. https://doi.org/10.1901/jaba.1997.30-313

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1982/1994). Toward a functional analysis of self-injury. *Journal of Applied Behavior Analysis*, *27*(2), 197–209. https://doi.org/10.1016/0270-4684(82)90003-9

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). Routledge.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical interference. *Journal of Applied Behavior Analysis*, *11*(2), 277–283. https://doi.org/10.901/jaba.1978.11-277

Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *British Medical Journal*, *330*(7494), 765. https://doi.org/10.1136/bmj.38398.500764.8F

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford University Press.

Lanovaz, M. J., Cardinal, P., & Francis, M. (2019). Using a visual structured criterion for the analysis of alternating-treatment designs. *Behavior Modification*, *43*(1), 115–131. https://doi.org/10.1177/014544517739278

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB. . . AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology*, *50*(5), 599–624. https://doi.org/10.1016/j.jsp.2012.05.001

Ottenbacher, K. J. (1990). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation*, *28*(5), 283–290.

Perone, M. (1991). Experimental design in the analysis of free-operant behavior. In I. H. Iversen & K. A. Lattal (Eds.), *Experimental Analysis of Behavior, Parts 1 & 2* (pp. 135–171). Elsevier Science.

Querim, A. C., Iwata, B. A., Roscoe, E. M., Schlichenmeyer, K. J., Ortega, J. V., & Hurl, K. E. (2013). Functional analysis screening for problem behavior maintained by automatic reinforcement. *Journal of Applied Behavior Analysis*, *46*(1), 47–60. https://doi.org/10.1002/jaba.26

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roane, H. S., Fisher, W. W., Kelley, M. E., Mevers, J. L., & Bouxsein, K. J. (2013). Using modified visual-inspection criteria to interpret functional analysis outcomes. *Journal of Applied Behavior Analysis*, *46*(1), 130–146. https://doi.org/10.1002/jaba.13

Spirrison, C. L., & Mauney, L. T. (1994). Acceptability bias: The effects of treatment acceptability on visual analysis of graphed data. *Journal of Psychopathology and Behavioral Assessment*, *16*(1), 85–94. https://doi.org/10.1007/BF02229067

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digital Medicine*, *3*, 17. https://doi.org/10.1038/s41746-020-0221-y

Thomason-Sassi, J. L., Iwata, B. A., Neidert, P. L., & Roscoe, E. M. (2011). Response latency as an index of response strength during functional analyses of problem behavior. *Journal of Applied Behavior Analysis*, *44*(1), 51–67. https://doi.org/10.1901/jaba.2011.44-51

Tufte, E. R. (1969). Improving data analysis in political science. *World Politics*, *21*(4), 641–654. https://doi.org/10.2307/2009670

Wickham, H. (2017). Tidyverse: Easily install and load the "tidyverse" (Version R package version 1.2.1). https://CRAN.R-project.org/package=tidyverse

Author Biographies

Alison Cox is an assistant professor in the Department of Applied Disability Studies at Brock University. She completed her doctoral degree in Psychology from the University of Manitoba. Dr. Cox has expertise in assessing and treating severe challenging behaviour, supporting skill acquisition, and supervising early intensive behavioural intervention programs.

Jonathan E. Friedel is an assistant professor in the Psychology Department at Georgia Southern University. He completed his doctoral degree at Utah State University and previously worked at the National Institute for Occupational Safety and Health. His research interests are in behavioral economics, discounting, and data analysis practices for single subject designs.