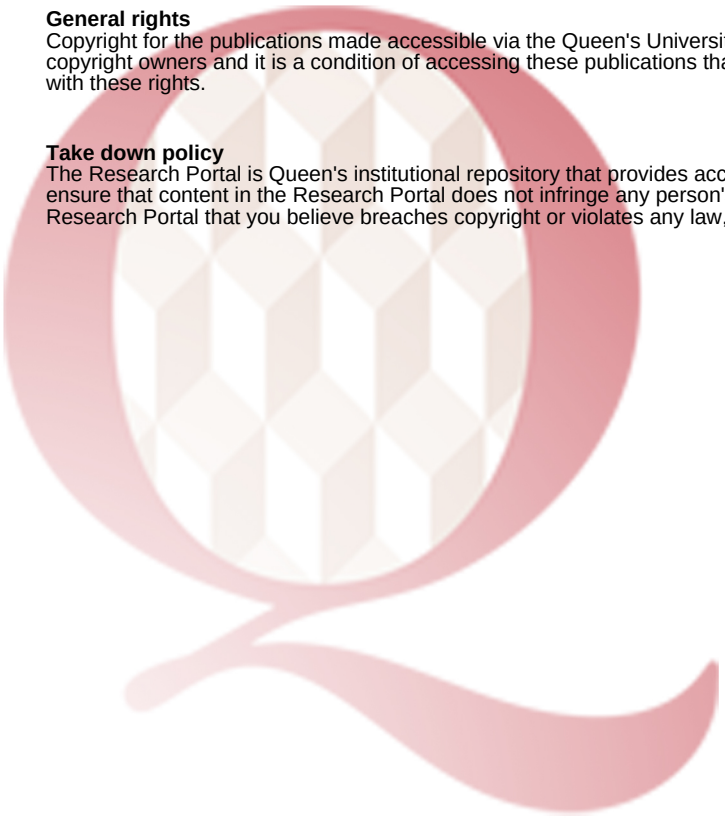
Behaviour analysis and evidence-based education

Dounavi, K., & Dillenburger, K. (2013). Behaviour analysis and evidence-based education. Effective Education, 4(2), 191-207. DOI: 10.1080/19415532.2013.855007

Published in:
Effective Education

Document Version:
Peer reviewed version



Dounavi, K. & Dillenburger, K. (2013). Behaviour analysis and evidence-based education. *Effective Education,* DOI: 10.1080/19415532.2013.855007

Behaviour Analysis and Evidence-based Education

Katerina Dounavi

&

Karola Dillenburger

Centre for Behaviour Analysis

School of Education,

Queen's University Belfast

69-71, University Street,

Belfast, BT7 1HL,

Northern Ireland

The authors would like to thank Professor Paul Connolly (QUB) for his valuable feedback on earlier versions of this manuscript. Correspondence relating to this paper should be addressed to k.dounavi@qub.ac.uk

Behaviour Analysis and Evidence-based Education

Abstract

Education has a powerful and long-term effect on people's lives and therefore should be based on evidence of what works best. This assertion warrants a definition of what constitutes good research evidence. Two research designs that are often thought to come from diametrically opposed fields, single-subject research designs and randomised controlled-trials, are described and common features, such as the use of probabilistic assumptions and the aim of discovering causal relations are delineated. Differences between the two research designs are also highlighted and this is used as the basis to set out how these two research designs might better be used to complement one another. Recommendations for future action are made accordingly.

Keywords: evidence-based education; behaviour analysis; single-subject research designs; randomised controlled-trials

What is Evidence-Based Education

Most of us would agree with Nelson Mandela's (2003, para.13) statement that "[e]ducation is the most powerful weapon we can use to change the world". However, when trying to reach consensus on a definition of what education actually is, what aims it serves, and how it can be improved several factors arise that make this a complex task. In this paper, we look at particular issues related to evidence-based education. First we explore two different experimental research designs that are commonly used to evidence the effectiveness of specific educational procedures, randomised controlled trials (RCTs) and single-subject research designs, then we outline how the natural science of Behaviour Analysis informs education and finally we conclude by delineating some important differences between natural and social sciences. These three areas are examined with the view to demonstrate that there are several variables that play an important role in evidence-based effective education and that these need to be taken into account in order to produce a robust and meaningful body of research.

According to the United Nations (2001, Article 29[1]), the main aim of education is "[t]he development of the child's personality, talents and mental and physical abilities to their fullest potential". This assertion defines education broadly and suggests that education is to be considered the means of achieving overall improvement in an individual's quality of life. This stands in contrast to some formal educational settings, such as schools, where education is frequently conceptualised rather narrowly as the students' performance in relation to curricular goals. The United Nations definition identifies that education is expected to ensure success not only in academic goals but also in other developmental areas. This contention is particularly pertinent for individuals with special needs for whom evidence-based educational interventions are expected to lead to progress in a range of domains, such as in self-help, motor, social and communication skills (Sundberg, 2008). In brief, according to the United Nations definition, education can be understood as any arrangement of environmental contingencies that leads to the development of practical, cognitive and emotional skills that enhance meaningful inclusion of individuals in society and enhances their chances of living a fulfilled life.

With a clear emphasis on standarisation, the No Child Left Behind Act (U.S. Congress, 2001) established a legal framework for the implementation of educational practices that included the need for measurable positive outcomes in students'

performance. Important changes in educational policies have followed the publication of this Act in the United States, where now a child's progress on academic, social, and other areas has to be measured against pre-defined standards. Educational strategies that are found to systematically produce positive outcomes are considered 'evidence-based' and are implemented on a large scale.

Equally, in England, current educational policies mandate for teaching strategies, hierarchisation of curriculum goals and design of teaching materials to be based on sound research results and require researchers and educators to continue conducting rigorous experiments in areas where clear measures of effectiveness and efficiency and the corresponding guidelines are missing (Davies, 1999). Funding priorities focus on research that contributes to or is based on evidence of what works and aims towards creating a robust body of knowledge that can lead to improvements of educational practice. The creation of the Education Endowment Foundation in 2011 in England is a good example of a funding body focussing on research of 'what works' in educational settings, such as schools.

'Reading' is a good example of an academic content area where evidence-based practices have been clearly outlined and are widely promoted. In the report produced by the National Reading Panel (2000), several teaching strategies were identified that reliably produced improved reading performance in children and recommendations for educational practice were made accordingly. For the inclusion of studies in this analysis, one of the criteria was that '[s]tudies had to adopt an experimental or quasi-experimental design with a control group or a multiple baseline method' (National Reading Panel, 2000, p 2-2). This is not a new criterion for research on effectiveness in education. For example some five decades ago, Campbell and Stanley (1963) thought that experiments or quasi-experiments constitute evidence of the effects of an intervention and the latter should be employed when true experiments were not feasible. Somewhat more recently, Davies (1999, p114) asserted that 'evidence consists of the results of randomised controlled trials or other experimental and quasi-experimental studies' and suggested that these research methods respond better than others to questions related to the effectiveness of an intervention in comparison to another intervention or approach. These kinds of statements highlight the importance of the actual research designs that are adopted when claims of causality are to be made (Slavin, 2002). Since experiments are considered to be the benchmark of evidence-

based practice, the question is what experimental or quasi-experimental research designs are.

Experimental Research Designs

Experimentation is defined as 'the scientist's way of discovering nature's rules' (Cooper, Heron, & Heward, 2007, p162). In other words, the aim of experiments is to determine causality. The scientist approaches this by measuring the dependent variable, i.e. the phenomenon under study, such as students' reading performance while manipulating the independent variable, i.e. the phenomenon that is considered responsible for the changes observed in the dependent variable, such as a specific teaching strategy. Depending on how well extraneous variables have been controlled, i.e. variables other than the independent variable, conclusions can be reached with varying levels of certainty in relation to the cause of the observed change, if any, in the dependent variable. Hence, in order to increase the internal validity of the experiment, researchers seek to make the necessary arrangements for the observed changes to be attributable to the independent variable, the intervention, and nothing else. Scientists also seek to ensure external validity, which refers to the ability for the findings to be generalisable into the wide population. Good research includes data on internal and external validity and reliability.

Increasing internal validity and the certainty with which we claim causality is synonymous with researchers seeking to ensure that the observed change in the dependent variable occurred only when the change in the independent variable was put in place, that no change in the dependent variable occurred when the independent variable was not changed and that repeatedly introducing and withdrawing or alternating the independent variable produces systematic changes in the dependent variable, for example in terms of its frequency, duration and intensity. For example, Kelley and Stokes (1982) demonstrated the effectiveness of a behaviour contract (independent variable) to increase classroom productivity (dependent variable), by recording data on students' performance under baseline conditions, when the behaviour contract was in place, when it was withdrawn (second baseline condition), and then when it was again introduced.

Apart from designs that reach internal validity through repeatedly introducing and withdrawing the intervention and monitoring the change in a particular individual's

behaviour (dependent variable), some research designs reach internal validity by alternating the delivery of two different interventions and monitoring the according changes on the dependent variable or by introducing the independent variable across different behaviours, subjects or settings in multiple steps (known as intra-subject or single-system research designs).

When withdrawals or alternations of the independent variable are not possible or desirable, different research designs are employed in order to increase internal validity and reach causality. A type of design that serves this purpose is the multiple baseline research design across behaviours, subjects, or settings. In other cases, internal validity is reached by randomly allocating subjects either to an experimental or to a control group that differ only in that one group receives the intervention while the other does not (known as group research designs). In these designs, if a difference between the mean values of the two groups is found to be statistically significant (independently of whether the change occurred in the same direction or amount for all subjects of the group), then this is attributable to the independent variable as the random allocation of subjects is assumed to have created two groups that are matched in all respects other than one receiving the intervention and the other not.

Ultimately, the social significance of the observed change is determined by factors such as the impact it has had on: the individual's life, important others' lives, or on the ability of the individual to interact with others. In order to determine the social (as opposed to the statistical) significance and validity of the change that was achieved, a number of qualitative methods are used, such as interviewing the individual, interviewing important others, assessing competent peers, and determining if the change generalised to the acquisition of other new skills (Barlow, Noch, & Hersen, 2009).

A wide range of research designs are used to assess what works in education and to ensure internal and external validity and reliability but, for the scope of the present paper, we will focus in more detail on the two most commonly used methods introduced above: randomised controlled trials (RCTs) and single-system research designs (SSR).

Randomised Controlled Trials (RCTs) and Single-Subject Research (SSR) designs

RCTs originated in medical research where they have been viewed as the 'gold standard' for measuring the impact of a given medication or intervention (Slade & Priebe, 2001). Most systematic reviews and meta-analyses rely solely or mainly on

synthesising studies employing RCT designs and specific guidelines aiming to improve their reporting have been suggested (Schulz, Altman, Moher, & the CONSORT Group, 2010). RCTs are based on intergroup comparisons. They typically include an experimental group, the group that receives the intervention, and a control group , the group that typically receives no intervention, although this can vary, e.g. often the control group receives a different intervention. The underlying assumption of RCTs is that if a given intervention is put in place and found to produce a change to the behaviour of a group of people, while at the same time another group of people who received no intervention did not show the same change of behaviour, it is claimed that the intervention is responsible for the observed change and extraneous variables can be ruled out. For example, in order to test for the effect of a specific educational intervention, researchers would administer a reading package to the experimental group and monitor individuals' progress in reading accuracy and fluency while at the same time administer no intervention to the control group and monitor for the progress of reading accuracy and fluency with these individuals. Often, a second control group would be used to control for placebo effects; for example, the teacher's attention alone would be administered and the course of reading accuracy and fluency would again be monitored. If the experimental group were the only group to show significant improvement in their reading skills, then causal inferences on the reading package's effectiveness would be made (e.g., Connolly, 2009; Connolly, O'Hare, & Mitchell, 2012).

While RCTs are now used increasingly in social and health sciences, in the natural sciences (e.g. Physics) an inductive rather than deductive approach guides scientific enquiry. This means that direct experimentation is used to examine the effects of changes in the independent variable on the dependent variable by manipulating, observing, and measuring each phenomenon separately in intra-subject rather inter-group designs (Kirkup, 1994). Internal validity is achieved by observing how individual responses within each system are changed as a result of manipulations of the independent variable rather than by comparing average changes in different systems to changes in independent variables across time as in RCTs and other inter-group designs. External validity and generalisation in these cases of direct experimentation is ensured through replication of intra-subject experimental results.

In SSRs, repeated observations of the dependent variable within the same experimental subject across different experimental conditions or within different

subjects within the same experimental conditions allows for the calculation of the exact effect of the independent variable. In order to further clarify this point, let's look at the behaviour of a physicist who studies gravity. The physicist observes and translates in quantified terms the dependent variable, e.g. the speed with which a given object approaches the ground, and then manipulates the independent variable, e.g. she may change the composition of the medium in which the object is located, the mass/weight of the object, or the height of the drop, to finally reach specific conclusions about the forces that influence the fall. In order to verify the results of her experiment, the physicist would attempt to replicate the experiment by examining a second object under the same or controllably different conditions and would test whether the same results are obtained. She would publish her results in reputable journals to allow other physicists to replicate her experiments. After a number of experiments showing consistent changes in the dependent variable following specific manipulations of the independent variable, conclusions about the natural laws that control or cause the fall of objects would be relatively safe.

While this example may seem simplistic, it is highly relevant when examined more closely. What would happen if physicists examined 100 different falling objects at the same time and, rather than separately controlling for the independent variable that affects each object, they registered the average speed with which objects approach the ground? If these objects were randomly assigned to two groups, one in which the independent variable (e.g. the density of the medium where objects are located) was manipulated and a second group in which no such manipulation was put in place, physicists could conclude that if statistically significant differences were found between the experimental and the control groups in the average speed with which objects approach the ground, this would be attributable to the difference in the density of the medium in which the objects are located. However, the constraints of the research design would still not allow scientists to examine the causes of the observed differences among the objects. If for example we examine if and how the medium in which an object is placed affects gravity and for this purpose we examine objects placed in gas and liquid media, then we may observe that liquids reduce the effect of gravity. However, even if this would allow us to conclude that the medium plays an important role in how gravity affects an object, the variance in the calculation would not allow us to conclude that gravity is a law of nature. Since wooden objects float, this particular research design would potentially detect a statistically significant difference among all

objects let fall in a liquid and those let fall in a gas medium but it would not explain why some objects do not fall and instead float and might even suggest that gravity is not a law of nature. Clearly then, specific research designs can confound the conclusions we draw from research.

Let's assume that an educational RCT is conducted with the aim to test the effect of a specific intervention on students' aggressive behaviour. In this study, one group would receive no treatment (control group) and a second group would receive the intervention (experimental group), for example, the removal of students from the classroom each time they engaged in aggressive behaviour. Let's say that the results would show that aggressive behaviour of both groups were maintained at high levels maybe even slightly increased with a low effect size and no significant differences were observed between the treatment and the control group. Based on these findings, the treatment would be deemed to be ineffective. In this hypothetical case, if we examined the data more closely, thus on an individual level, we would more probably observe that some students showed a great improvement, i.e., their aggressive behaviour decreased significantly, but many other students showed no change or significant increases in the levels or aggressive behaviour, thus closely matched the mean changes observed in the control group. Therefore, the expected conclusion would be that the intervention worked well for some students, but may even be counterproductive for others, in that it increased their aggressive behaviour. Once this conclusion was reached, researchers could then examine why the intervention works well for some individuals, thus seek the cause or the variable that maintains aggressive behaviour and then manipulate it accordingly. However, in RCTs any positive changes in aggressive behaviour that may have been observed for some individuals would be attributable to random variation and would not be further examined. RCTs would not be capable of generating evidence of the differential effects of the treatment on each individual level and therefore would not be able to identify the independent variable ruling the behaviour of all students. When seeking to discover the 'real' underlying cause, we refer to the variable that if manipulated, and given that other factors are also controlled, always produces the expected effect. In cases where these other factors have been controlled and random variation cannot be explained, causality has not been reached. RCTs then clearly are not the appropriate research design in these cases, though in cases they can identify some independent variables mediating the variation (e.g., age, genre), since there is no functional account of why each individual's aggressive behaviour is maintained

meaning that no conclusion can be reached as to why some individuals showed decreased levels of aggression post-test while others did not. An alternative more appropriate way forward in this case would be to conduct a functional analysis and then use the intervention that meets the functional requirements for the behaviour change targets.

The idea of conducting functional analyses for the detection of the variables that maintain behaviour was introduced to behavioural and educational interventions by Iwata, Dorsey, Slifer, et al. (1994). Today, conducting functional behaviour analyses is common practice before any effective new teaching strategy is put in place especially when dealing with challenging behaviours (e.g., Mueller, Nkosi, & Hine, 2011). In brief, functional analyses experimentally analyse human behaviour in relation to environmental contingencies, i.e., the antecedent and consequent conditions of which the behaviour in question is a function. More specifically, a functional analysis uses the systematic manipulation of independent variables to identify the cause/s of a behaviour, thus the contingencies that if manipulated reliably produce changes in the behaviour. Consequently, in cases where change is socially desirable, functional analyses identify the path that leads to the behaviour change targets. As such, functional analyses serve as the key for successful educational interventions or treatment.

Going back to the previous example, we could hypothesize that there were a number of students in both the treatment and the control conditions that showed aggressive behaviour as a function of gaining attention (social reinforcement). Contingent removal of these students from a socially stimulating classroom, translated as removal of social attention after they had engaged in aggressive behaviour, produced a decrease in aggressive behaviour. Plenty of studies published in the scientific literature show the effectiveness of these kinds of time-out from positive reinforcement procedures to decrease inappropriate behaviour (e.g., Donaldson & Vollmer, 2011). However, for some of the students aggressive behaviour was maintained by escape from demands, defined as negative reinforcement, and contingent removal from an academically demanding classroom led to increases in aggressive behaviour; they learned that aggressive behaviour got them out of the academically demanding environment. Of course, this kind of escape-maintained problem behaviour has not only detrimental effects on their own academic and social development, but also disrupts the activities of their class of peers. Other students, whose aggressive behaviour was maintained by automatic reinforcement, defined as the sensory consequences produced

by the behaviour itself, showed hardly any change in the levels of aggressive behaviour. Maybe when these students were removed from classroom to a quiet place that provides less sensory stimulation, they engaged in increased levels of aggressive behaviour to compensate the lack of stimulation or they simply continued engaging in the same levels of aggressive behaviour resulting in the same levels of sensory stimulation. The aggressive behaviour of students in the control group is likely to have served similar functions but the level of aggressive behaviour remained largely unchanged throughout the observation period, due to the fact that no new contingencies were introduced. In an intra-group comparison study, these spurious effects showed that the intervention had no or only minimal differential effects although clearly, this kind of intervention could have been highly effective for some of the students from both the experimental and control groups; those whose behaviour was maintained by positive social reinforcement, namely, attention. Nevertheless, in order to identify the independent variable that causes these changes in the individuals' behaviour, we would necessarily need to examine the dependent variable from a closer level (i.e., the intra-subject level), thus a research design that has this potential would be required. As a second step that would increase external validity, an RCT could then be conducted with students whose aggressive behaviour is maintained by positive social reinforcement to compare the effect of a particular intervention on this versus a control group. This would allow for an immediate replication of the findings with numerous participants and would counter-balance the low generalisability potential that SSRs offer unless replicated numerous times, which may be very time-consuming.

In sum, functional analysis procedures were developed from a behaviour analytic understanding of causality. Behaviour analytic research (e.g., Derby, Wacker, Peck, Sasso, DeRaad, Berg, Asmus, & Ulrich, 1994; Hanley, Iwata, & McCord, 2003) has repeatedly shown that evidence of the topography of a behaviour , i.e., what it looks like, its structure or form is not sufficient for understanding the function of a behaviour. This is because behaviour can look the same but have different functions, or it can have the same function but look quite differently (Dillenburger, 2000). Before causal inferences can be made and effective treatments can be designed, the observation and experimental manipulation of the environmental variables that occur before and after the behaviour in question is warranted. In cases where this has been done, dramatic positive changes have been achieved and an overall improvement

in individuals' life has been reached (e.g., Wacker, Lee, Padilla-Dalmau, Kopelman, Lindgren, Kuhle, Pelzel, & Waldron, 2013).

This is the case for children with Autistic Spectrum Disorders (ASD) that have received behaviour-analytic interventions and as a result have acquired new and important life and academic skills that have led to meaningful social inclusion (e.g., Maglione, Gans, Das, Timbie, & Kasari, 2012). However, these positive effects have not only been recorded with individuals with ASD, they have also been obtained with other diverse populations, such as adults with obesity problems (VanWormer, 2004), children with literacy problems (de Rose, de Souza, & Hanna, 1996) and adult gamblers (Dixon & Holton, 2009). All of these interventions were based on functional analyses and were reported in single-system research designs. It is no surprise that when the contingencies that maintain a behaviour are identified, meaningful change becomes possible, and thousands of research studies have repeatedly shown how behaviour can be increased, decreased, or maintained, often quite rapidly, when correctly identified independent variables are manipulated (Engerman, Austin, & Bailey, 1997). Yet, functional analysis or SSRs are not taught routinely at professional qualifying courses in education, health or social care, and SSRs often are not represented in systematic or other reviews of relevant intervention research (e.g., Kazdin, 1982, pviii).

The wide range of studies reported earlier was based on two of the designs of SSR, namely reversal research designs (also called ABAB design, where A stands for baseline and B for intervention) and alternating treatment designs. In an ABAB design, the behaviour in question is continuously monitored while the independent variable is repeatedly introduced and removed. A causal inference is made if the behaviour returns to baseline levels when the intervention is withdrawn, because then there is reason to believe that the intervention 'caused' the behaviour change. Alternating treatment designs aim to discover which of two or more rapidly alternating treatment conditions is more effective. However, there are other single-system research designs that are used frequently in behaviour-analytic studies that also allow for causal inferences to be made. Multiple baseline designs do not require a return to baseline and are used in cases where the effect of the intervention cannot or should not be reversed. Examples of such cases include situations in which ethical reasons refrain the researcher from removing an effective intervention, e.g. an intervention that reduced self-injurious behaviour, or situations in which learning has permanently influenced a behaviour and this cannot be

reversed, e.g. even if we remove an effective intervention that resulted in a student learning to read, the student will still be able to read.

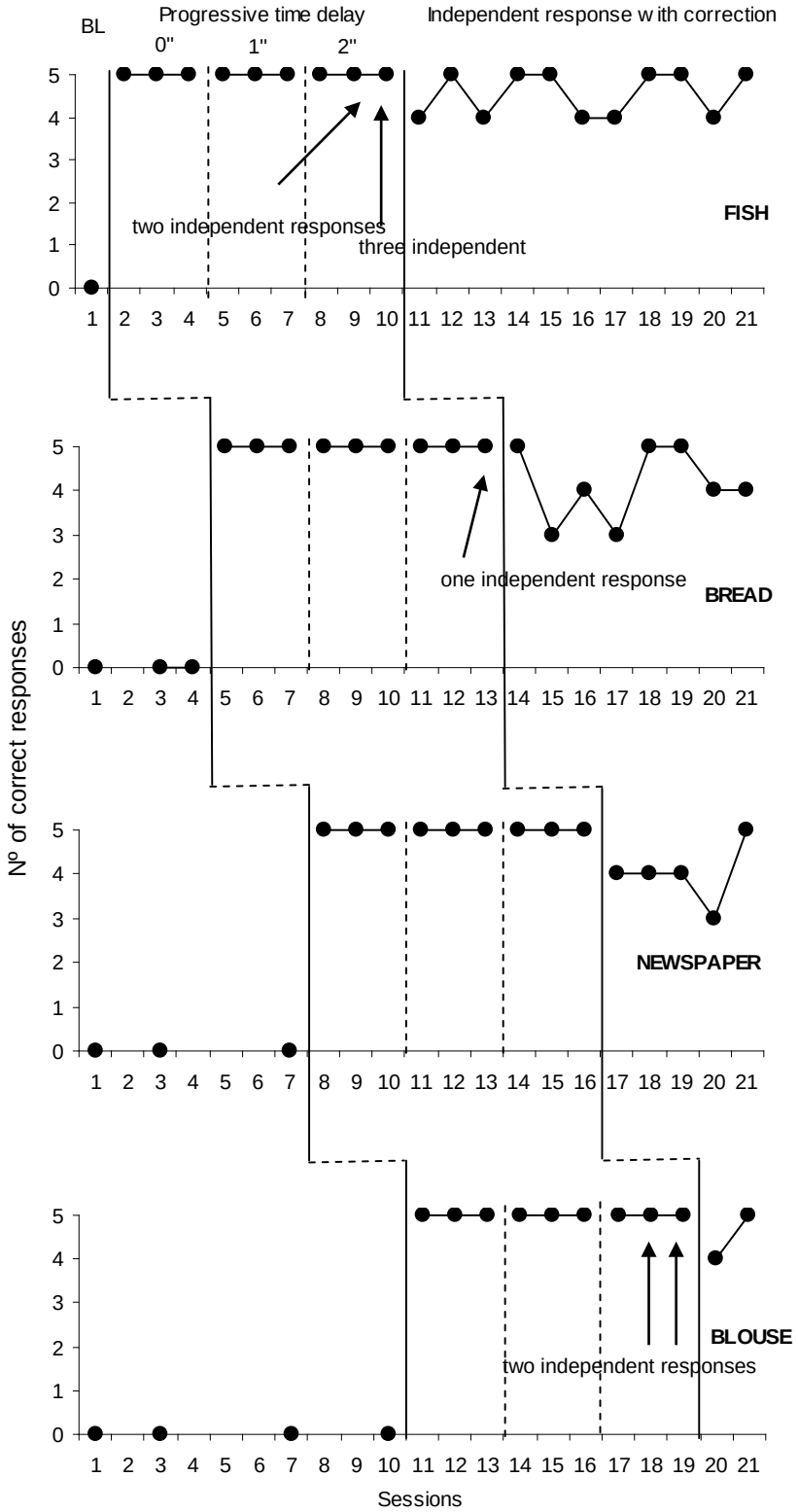
Multiple baseline research designs (Figure; taken from Dounavi, 2013) require the measurement of a number of baselines either of different behaviours for the same participant, the same behaviour in different settings, or behaviour with the same function across different participants. An intervention (change in independent variable) is put in place one at a time, while baselines are continuously recorded for the other dependent variables. Behaviours are randomly allocated to the experimental or control conditions initially. The assumption is that if a change is observed in the behaviour/setting/participant that was targeted by the intervention, without a change in the other behaviour/setting/participant, there is a relatively high degree of certainty that this change can be causally attributed to the change in the independent variable, i.e. the intervention (Kazdin, 1982). If the data for the second and third dependent variables follow the same pattern when the intervention is put in place the certainty about causal relations, and thus internal validity increases. Of course, further replications with the same results add external validity (Cooper, Heron, & Heward, 2007).

In order to understand the logic of single-system research designs, a multiple-baseline research design is shown in the Figure that is taken from a study reported in Dounavi (2013). In this example, a male adult diagnosed with post-stroke global aphasia was taught to name objects when presented with a picture of the object, e.g., he was asked to say “bread” when presented with the picture of bread. In this case, the individual was not able to emit any words independently prior to the treatment, meaning that the number of correct responses when presented with the picture of a fish during the first session was zero, as can be observed in the upper panel of the figure. One of the purposes of the research design was to rule out spontaneous recovery, defined as the possibility that words had been recovered as a result of the time passing after the stroke and not as a result of a treatment. For this purpose, baseline measurements were taken on the number of correct vocal responses of the participant when presented with the picture of an object, with correct responses being defined as vocalizing the name of the object with intelligible articulation (not lacking more than one sound from the original word) and within 3 seconds from the presentation of the picture. As shown in the figure, the participant gave no correct responses before the intervention was put in place for the four words.

Once baseline measurements had been taken on all four words, words were randomly ordered and the treatment was put in place for the first, the word “fish”. After recording stable increases in correct responding and simultaneously monitoring the remaining three behaviours in relation to the other three pictures under baseline conditions to make sure no changes were occurring, the treatment was put in place for the second behaviour which was saying the word “bread”. Once again, the third and fourth behaviours/words were continuously monitored under baseline conditions in order to rule out spontaneous recovery or improvement due to extraneous variables or generalisation. Once the second behaviour had reliably changed as a result of the intervention, the intervention was introduced for the third behaviour and subsequently the fourth behaviour. Following successful intervention in all behaviours/settings/participants, maintenance procedures are put in place, however, in this particular study maintenance data were reported only anecdotally, no direct measurements were taken, thus rendering it difficult to calculate effect sizes. However, the nature of the multiple baseline design itself allows the researcher to reach conclusions on the effectiveness of the treatment, in this case, the progressive time delay method to teach ‘tacts’ (note: ‘tact’ is the technical term for ‘naming objects’; Skinner, 1957) to an adult with global post-stroke aphasia. The progressive time delay procedure consists in delivering a prompt after the presentation of the discriminative stimulus (in this case the picture) and before the subject responds and gradually increasing the time between the presentation of the discriminative stimulus and the delivery of the prompt until the subject’s response becomes independent (i.e., it is emitted under the sole control of the target discriminative stimulus and not the prompt). In the study described in the figure, the experimenter delivered prompts by vocalising the correct word while presenting the picture, 1sec after the presentation of the picture, 2secs after the presentation of the picture and lastly completely faded the prompt. As indicated with arrows, independent responses occurred before the prompts were delivered in the last phase, in which 2secs elapsed between the presentation of the pictures and the delivery of the prompt. Clearly, for this participant the intervention was successful as he learned to say all four words. Of course, in order to achieve external validity this study would further need to be replicated with other individuals, behaviours and settings. The number of replications needed for reaching a satisfactory level of certainty and generalisability could be specified after probabilistic calculations wasa conducted but would probably be required to be elevated. Instead, conducting an RCT

once the intervention had rendered effective with a small number of subjects to exponentially increase external validity and further add certainty to the causal claims would shorten the pathway to fulfilling more research goals. Such a combination of research approaches would presumably constitute an extremely robust evidence for effectiveness of an intervention and thus collaborations between experts of both research designs should be seen more often.

Figure. Multiple baseline design across 'tacts' showing number of correct responses during baseline, progressive time delay and independent with correction conditions.



Dounavi, K. (April, 2013). *Using progressive time-delay to teach tacts to an adult diagnosed with post-stroke aphasia.* Poster session presented at the 7th Annual Conference of the Division of Behaviour Analysis, The Psychological Society of Ireland, Galway, Ireland.

The role of Behaviour Analysis in Evidence-based Education

Evidently, the scientific discipline of behaviour analysis has greatly contributed to the development and refinement of experimental or quasi-experimental research designs, the results of which constitute a robust body of evidence that guides educational and clinical practice and policy (Keenan & Dillenburger, 2011). The basic concepts of the scientific method of enquiry on which behaviour analytic research designs are based are description, prediction, affirmation of the consequent, verification, and replication, e.g., the assumption that when baseline measurements are stable, they remain so if independent variables are held constant by not introducing any intervention. Consequently, the assumption is that behaviour change is due to the manipulation of the independent variable (Cooper, Heron, & Heward, 2007). The inductive (i.e. experimentally tested relations) rather than deductive (i.e. developing a theory and testing hypotheses) logic that underlies single-subject research designs means that observations of concrete and clearly defined phenomena, in this case publicly or privately observable behaviours, lead to the discovery of natural laws or principles (Johnston & Pennypacker, 2009).

Single-subject research designs can be complemented and further strengthened by the use of statistical procedures, such as the calculation of effect sizes used for meta analyses (Koehler & Levin, 2000) or the use of non-regression and regression-based methods in order to summarize the efficacy of interventions (Campbell, 2004). The advantages of combining visual analyses with the use of statistical procedures in single-subject research designs can be important (Beeson & Robey, 2006; Ma, 2006).

Once enough single-system design experiments have been conducted to establish causal relations between a specific procedure and a specific behavioural change, it has been argued that valuable knowledge and external validity could be added by conducting a RCT (Smith, 2013). Once the causal variable has been identified, random variation would be expected to be minimal, and all individuals' behaviours would present similar patterns of change with an RCT further increasing the internal and external validity of the results. This may be possible, given that many teachers already use behavioural techniques in classroom, often without understanding why they work or without being fully prepared to tailor them to address each individual's needs (Dillenburger, 2012). If appropriate training and support were provided, it could be

possible to achieve levels of procedural fidelity that allow for small-scale quasi-experiments or RCTs (e.g., Connolly et al., 2012).

Collaboration between researchers and professionals from applied settings would need to focus on the design, implementation and monitoring of progress of several educational procedures and a number of research designs would be used to establish an evidence base. Obviously, the choice of the most appropriate research design would depend on the research question. For example, for drawing initial inferences on causal relations between phenomena that have not been thoroughly studied before, small-scale experiments in the form of single-subject research designs would be appropriate, while for larger-scale experiments that seek to control for extraneous variables and produce generalisable results both single-subject research designs and RCT could be employed, with multiple baseline research designs better serving these purposes than reversal designs in the case of SSRs. This is because employing an RCT in an area that has not been thoroughly studied may be helpful in identifying some relations between the independent variable and the behaviour change but this may also not hold true if there is much noise in the data, as in the example of random variation in pupils' aggressive behaviour presented previously. Since random variation cannot be controlled or explained with this research design and functional relations cannot be uncovered in all cases, SSRs should be first employed in order to identify them and RCTs would afterwards add.generalisability to the data and of course further increase internal validity by ruling out extraneous variables on a group level.

Is the research design the only critical variable?

After having explained some basic assumptions that underlie experiments and having argued that both RCTs and single-subject research designs seek to pursue a common aim, discovering causal relationships and share some common concepts, such as randomization and control, a number of other variables need to be taken into account when designing and conducting research that aims to produce evidence-based interventions.

Clearly, the disciplines of education, psychology, and behaviour analysis use scientific methods to study the same subject matter, human behaviour (Ledoux, 2002) but the way the go about this remains very distinct. This leads to the assumption that the use of the scientific method is not a standalone criterion when examining the evidence that arises from a certain discipline. For example, education and behaviour analysis both

use scientific methods with the aim to understand the causes of behaviour. The difference lies in the interpretation of the findings. While the former tends to base explanations on concepts that cannot be empirically evidenced, such as mind, talent, or character, behaviour analysis studies human behaviour by focusing on naturally occurring contingencies, i.e., clear if-then relationships between behaviour and environmental events.

An example of the distinction in the interpretation of behavioural observation is the concept of "intelligence". Typically in education, 'intelligence' is used at least in part as an explanatory variable for the reading performance of a student. In contrast, in behaviour analysis the term 'intelligence' is considered at best a convenient summary label for behavioural observations (Grant & Evans, 1994) and any explanatory power ascribed to this term would be avoided as circular reasoning. In this example, a behaviour analyst would search for the explanation of the same behaviour (reading performance) in the contingency than are responsible for the occurrence and maintenance of the behaviour. It is likely that an explanation would be based on the child accessing repeated practice with the reading materials, receiving appropriate prompts, and accessing reinforcers for fluency in reading. This distinction makes behaviour analysis a natural science and categorises education or psychology as social sciences; a characteristic that clarifies the scope and focus with which the phenomenon under study is approached rather than amplifying whether a discipline is scientific or not. This distinction between natural and hypothetical events is critical, since even if it were possible to scientifically prove that hypothetical events have any effect on the dependent variable under study as they are often claimed to do, problems arise when researchers or practitioners are asked to influence these non-natural, hypothetical assumptions or events.

For the shake of clarity, let's consider the following example. A group of students is assessed with the use of a questionnaire and found to show varying levels of cognitive skills. Students are randomly assigned to an experimental and control group and an intervention for the improvement of reading skills is put in place. The results of the study show that the experimental group achieved better outcomes than the control group and that the cognitive level accounted for an important part of the variance of the obtained data within the experimental group, with students that showed higher cognitive skills before the intervention reaching significantly higher scores in the reading assessments than those with lower cognitive scores. Although this claim sheds light to

the effectiveness of the intervention and an RCT would have achieved a high degree of certainty with which to make this claim, if the researchers were asked "What are cognitive skills and how can these be manipulated in order to boost the performance of all students?", they would have to make reference to a non-natural event. 'Cognitive skills' is an umbrella term that describes a number of behaviours that the students did or did not show during the assessment after the presentation of certain stimuli and as such it cannot be manipulated, unless the individual behaviours that are comprised under this category label are manipulated. However, if the focus of a given study relies on the careful observation of clearly defined individual behaviours, i.e., a natural phenomenon rather than the conventional umbrella labels, single-subject research designs would ensure that the researcher would be able to manipulate contingencies and consequently measure the effect on the students' reading skills. In RCTs and other traditional research designs, phenomena are commonly defined in mentalistic terms, thus their examination is rendered more difficult. Mentalistic terms, summary labels, umbrella terms, non-naturally occurring events all make reference to phenomena not observed as such happen in nature and thus concrete, objective and meaningful for behaviour change definitions of these phenomena are difficult to achieve. Nevertheless, measurement of variables can only be conducted if appropriate definitions of phenomena exist, thus this mentalistic terms should be avoided all together in research and applied settings, unless broken down in the individual behaviours that comprise them.

The science of behaviour analysis through both the applied and the experimental branches has repeatedly shown effective ways to produce change in human behaviour, be it increasing, decreasing, generalizing or maintaining existing behaviours or establishing new ones. Through over 100 years of careful experimentation behaviour analysis has discovered that human behaviour, much the same as other natural events, follows natural laws, e.g., operant conditioning and the matching law (Myerson & Hale, 1984). The application of these laws reliably leads to socially relevant, desired behaviour change and thus it is safe to conclude that those charged with having an effect on the behaviour of others, de facto exactly what education is about, would greatly benefit from this knowledge.

In the present paper it is claimed that a consistent framework that aims to study human behaviour as a natural phenomenon and uses natural explanations for this can produce socially significant and evidence-based outcomes, given that the appropriate scientific methods are also used , e.g., appropriate experimentation as described above.

Ultimately the question addressed in this paper relates to the way in which different research designs address causality, and thus can be used to guide large-scale implementation of certain educational interventions. Are causal relations between intervention and outcome data that are observed through SSDs, such as multiple baseline designs, safer than those observed through RCTs? The question relates to the degree of certainty reached with each of these research designs. As we have seen, there are some similarities, such as randomisation, however, the necessity to describe events in natural terms precedes this debate.

Conclusions and Future Recommendations

In this article, we summarised the basic differences between RCTs and single-subject research designs and outlined the advantages and disadvantages of both in relation to evidence-based education. In Table 1 the major advantages, disadvantages and characteristics of SSRs and RCTs have been described. SSRs follow an inductive approach, while RCT a deductive one. SSRs focus on measurable behaviours (e.g., reading performance defined as the number of words read correctly in one minute), while RCTs may well focus on measurable behaviours (the same example of reading performance applies) or latent variables described mentalistically (e.g., cognitive ability).When examining causality, SSRs are strong at individual level, since they can identify the functional relations between the independent and dependent variable, while RCTs are weak at individual level and their effectiveness varies according to the amount of noise in the data at group level. It is important to underline that when no random variation is observed in data at group level, RCTs are very powerful designs for increasing internal and external validity. However this may not always be the case and in these instances, random variation obscures explanations of causality at group and individual level. Finally, as far as generalisability is concerned, SSRs achieve it through numerous replications of the same findings with different participants, in different settings and for different behaviours. Once results have been replicated numerous times, safe conclusions on the principles that govern human behaviour can be gathered and these are supposed to be universally applicable, thus applicable to any human being. RCTs on the other hand achieve generalisability of the results immediately but this is

limited to the sample. It is here assumed that functional relations identified through SSRs and replicated with the use of RCTs would render conclusions universally applicable more readily.

We argued that evidence-based education should rely on definitions of natural, observable events that allow for consistent manipulations of the independent variables and for meaningful observations of changes in the dependent variables that subsequently lead to educational guidelines with applied value. We have shown that single-subject research designs are powerful tools that can be used to discover causal relations between environmental contingencies and individual behaviour with high levels of confidence (Barlow, Noch, & Hersen, 2009; Pechacek, 1978). We suggested that the use of a combination of different scientific methods would add more certainty to the causal inferences but this would have to follow in logical sequence for the results to be sound. Single-subject research designs would have the protagonistic role of defining the variables, affirming the consequent and measuring the exact effects, while RCTs would play a critical role in increasing the external validity of SSR findings and adding a second layer of certainty as far as the control of extraneous variables is concerned.

We recommended that researchers and practitioners are familiar with both methods. Evidently, when knowledge derived from the science of behaviour analysis is used in the design of educational interventions, results are consistently very positive and for this reason, research collaborations should be encouraged. The impact of these collaborations would be enormous, since several research criteria, such as internal and external validity, could immediately be fulfilled if both methods are used in the order specified in the present paper.

Finally, it is important to highlight that there may be ethical or practical reasons that may prohibit real randomization in RCTs (Smith & Pell, 2003). This is especially the case where systematic reviews and meta-analyses have evidenced that certain interventions are more effective than others. This is the case, for example with regard to the existing body or knowledge on effective interventions for individuals with ASD where an abundance of evidence exists in favour of behaviour analytic procedures (American Academy of Pediatrics, 2007; Maine Department of Health and Human Services (DHHS) & Maine Department of Education, 2009; Ministries of Health and Education, 2008; Myers, Johnson, & the Council on Children With Disabilities, 2007; National Standards Report, 2009; National Autism Center, 2009; New Zealand

Guidelines Group, 2008; Scottish Intercollegiate Guidelines Network, 2007; Surgeon General, 1999). Even if in this case, conducting an RCT in a different country for generalisation purposes would probably be worthwhile, the principle of equipoise applies (Fries & Krishnan, 2004) which would prohibit random selection of research participants. The same would apply for randomisation in SSRs, such as in multiple baseline designs across subjects, in which delivering the intervention to some individuals with a delay for research purposes could be devastating (e.g., an intervention aiming to decrease self-injurious behaviour) Ultimately, the selection of research designs has to be guided, not be personal preference of the researcher or pure scientific enquiry, but by scientific, practical, as well as ethical concerns all together.

Table 1. *Main characteristics of SSRs and RCTs.*

	SSRs	RCTs
Approach	inductive	deductive
Focus	measurable behaviours	measurable behaviours or latent variables
Causality	strong at individual level	weak at individual level strong at group level if no noise in the data obscured at group level if random variation exists
Generalisability	achieved through replication	Strong but limited to sample

References

American Academy of Pediatrics. (2007). Management of Children with Autism Spectrum Disorders, *Pediatrics, 120*, 1162-1182.

Barlow, D., Nock, M., & Hersen, M. (2009). Single case experimental designs: strategies for studying behavior for change (3rd edition). Boston, MA: Pearson Education, Inc.

Beeson, P. M. & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review, 16,* 161-169.

Campbell, J. M. (2004). Statistical Comparison of Four Effect Sizes for Single Subject Designs. *Behavior Modification, 28,* 234-246.

Campbell, D. T. & Stanley, J. C. (1963). *Experiments and quasi-experimental designs for research.* Chicago: R. McNally.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis.* 2nd edition. Pearson, Merrill: Prentice Hall.

Connolly, P. (2009): The challenges and prospects for educational effectiveness research. *Effective Education, 1,* 1-12.

Connolly, P., O'Hare, L., & Mitchell, D. (2012). *A Cluster Randomised Controlled Trial Evaluation of Booktime Northern Ireland: A Book Gifting Intervention for Children in Their First Year of Primary School.* Belfast: Centre for Effective Education, Queen's University Belfast.

Davies, P. (1999). What is evidence-based education? British Journal of Educational Studies, 47, 108-121.

de Rose, J. C., de Souza, D. G., & Hanna, E. S. (1996). Teaching reading and spelling: Exclusion and stimulus equivalence. *Journal of Applied Behavior Analysis, 29,* 451-469.

Derby, K. M., Wacker, S. P., Peck, S., Sasso, G., DeRaad, A. Berg, W., Asmus, J., & Ulrich, S. (1994). Functional analysis of separate topographies of aberrant behavior. *Journal of Applied Behavior Analysis*, 27, 267–278.

Dillenburger, K. (2012). Why reinvent the wheel?: A behaviour analyst's reflections on pedagogy for inclusion for students with intellectual and developmental disability. *Journal of Intellectual and Developmental Disability*, 37(2), 169-180.

Dillenburger, K. (2000). Functional assessment and analysis. Entry in M. Davies. *Encyclopaedia of Social Work* (pp. 140-141). Oxford: Blackwell Publishers.

Dixon, M. R. & Holton, B. (2009). Altering the magnitude of delay discounting by pathological gamblers. *Journal of Applied Behavior Analysis, 42,* 269-275.

Donaldson, J. M. & Vollmer, T. R. (2011). An evaluation and comparison of time-out procedures with and without release contingencies. *Journal of Applied Behaviour Analysis, 44,* 693-705.

Engerman, J. A., Austin, J., & Bailey, J. S. (1997). Prompting patron safety belt use at a supermarket. *Journal of Applied Behavior Analysis, 30,* 577-579.

Fries, J. F. & Krishnan, E. (2004). Equipoise, design bias, and randomized controlled trials: the elusive ethics of new drug development. *Arthritis Research & Therapy, 6,* 250-255.

Grant, L. & Evans, A. (1994). *Principles of Behavior Analysis.* 1st edition. New York: HarperCollins College Publishers.

Hanley, G. P., Iwata, B. A., & McCord, B. E. (2003). Functional analysis of problem behavior: A review. *Journal of Applied Behavior Analysis, 36,* 147-185.

Iwata, B.A., Dorsey, M.F., Slifer, K.J., Bauman, K.E., & Richman, G.S. (1994). Towards a functional analysis of self-injurious behaviour. *Journal of Behavior Analysis,* 27, 197-209 (reprinted from *Analysis and Intervention in Developmental Disabilities,* 1982, 2, 3-20).

Johnston, J. M. & Pennypacker, H. S. (2009). *Strategies and tactics of behavioural research.* 3rd edition. New York: Routledge.

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

Keenan, M. & Dillenburger, K. (2011). When all you have is a hammer…: RCTs and hegemony in science. *Research in Autism Spectrum Disorders, 5,* 1-13.

Kelley, M. L. & Stokes, T. F. (1982). Contingency contracting with disadvantaged youths: Improving classroom performance. *Journal of Applied Behavior Analysis, 15,* 447-454.

Kirkup, L. (1994). *Experimental methods: An introduction to the analysis and presentation of data.* Brisbane; Chichester: Wiley.

Koehler, M. J. & Levin, J. R. (2000). RegRand: Statistical software for the multiple-baseline design. *Behavior Research Methods, Instruments, & Computers, 32,* 367-371.

Ledoux, S. F. (2002). Defining Natural Sciences. *Behaviorology Today, 5,* 34-36.

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

Maglione, M. A., Gans, D., Das, L., Timbie, J., & Kasari, C. (2012). Non-medical interventions for children with ASD: Recommended guidelines and further research needs. *Pediatrics*, 130, 169-179.

Maine Department of Health and Human Services (DHHS) and Maine Department of Education. (2009). Interventions for Autism Spectrum Disorders: State of the Evidence. Retrieved on June 23, 2013 at www.maine.gov/dhhs/ocfs/cbhs/ebpac/asd-report.doc

Mandela, N. (July, 2003). Lighting your way to a better future. Speech addressed at Planetarium, University of the Witwatersrand, Johannesburg South Africa. Retrieved June 14, 2013 at http://db.nelsonmandela.org/speeches/pub_view.asp?pg=item&ItemID=NMS909&txtstr=education%20is%20the%20most%20powerful

Ministries of Health and Education. (2008). *New Zealand Autism Spectrum Disorder Guideline.* Wellington: Ministry of Health.

Mueller, M. M., Nkosi, A. & Hine, J. F. (2011). Functional analysis in public schools: A summary of 90 functional analyses. *Journal of Applied Behavior Analysis, 44*, 807-818.

Myers, S. M., Johnson, C. P. & the Council on Children with Disabilities. (2007). Management of Children with Autism Spectrum Disorders. *Pediatrics, 120,* 1162-1182. Retrieved on June 23, 2013 at http://www.feathouston.org/Article1.pdf

Myerson, J. & Hale, S. (1984). Practical implications of the matching law. *Journal of Applied Behavior Analysis, 17*, 367–380.

National Autism Center. (2009). *Evidence-Based Practice and Autism in the Schools: A Guide to Providing Appropriate Interventions to Students with Autism Spectrum Disorders.* Randolph, MA. Retrieved June 23, 2013 at http://www.nationalautismcenter.org/pdf/NAC%20Ed%20Manual_FINAL.pdf

National Reading Panel. (2000). Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction. National Institute of Child Health and Human Development: U.S.A. Retrieved June 14, 2013 at http://www.nichd.nih.gov/publications/pubs/nrp/Documents/report.pdf

National Standards Report. (2009). The National Standards-Addressing the Need for Evidence-based Practice Guidelines for Autism Spectrum Disorders. National Autism Center, Randolph, MA. Retrieved June 26, 2013 at http://www.nationalautismcenter.org/pdf/NAC%20Standards%20Report.pdf

New Zealand Guidelines Group. The effectiveness of applied behaviour analysis interventions for people with autism spectrum disorder. *Systematic Review.* Wellington; 2008.

Pechacek, T. F. (1978). A probabilistic model of intensive designs. *Journal of Applied Behavior Analysis, 11,* 357-362.

Scottish Intercollegiate Guidelines Network (2007). *Assessment, diagnosis and clinical interventions for children and young people with autism spectrum disorders: A national clinical guideline.* Edinburgh (Scotland): Scottish Intercollegiate Guidelines Network (SIGN).

Schulzl, K. F., Altman, D. G., Moher, D., & the CONSORT Group. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Trials, 11, 1-8.

Skinner, B. F. (1957). *Verbal Behavior.* New York: Appleton-Century-Crofts.

Slade, M. & Priebe, S. (2001). Are randomised controlled trials the only gold that glitters? *British Journal of Psychiatry, 179,* 286-287.

Slavin, R. E. (2002). Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher, 31,* 15-21.

Smith, G. C. D & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *British Medical Journal*. 327, 1459-1461.

Smith, T. (2013). What is evidence-based behavior analysis? *The Behavior Analyst*, 36, 7–33.

Sundberg, M. L. (2008). Verbal behavior milestones assessment and placement program: The VB-MAPP. Concord, CA: AVB Press.

Surgeon General. (1999). Mental health: A report of the Surgeon General. U.S. Public Health Service: U.S.A. Retrieved on 21/11/2012 from http://profiles.nlm.nih.gov/ps/retrieve/ResourceMetadata/NNBBJC.

United Nations. (2001). *Committee on the rights of the child. General comment No. 1* (2001): The Aims of Education (UN/CRC/GC/2001/1). Geneva: United Nations.

Retrieved June 26, 2013 at http://www.unhchr.ch/tbs/doc.nsf/(symbol)/CRC.GC.2001.1.En

U.S. Congress. (2001). No Child Left Behind Act of 2001. Washington, DC: Author. Retrieved June 14, 2013 at http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf

VanWormer, J. J. (2004). Pedometers and brief e-counseling: Increasing physical activity for overweight adults. *Journal of Applied Behavior Analysis, 37,* 421-425.

Wacker, D. P., Lee, J. F., Padilla-Dalmau, Y. C., Kopelman, T. G., Lindgren, S. D., Kuhle, J., Pelzel, K. E., & Waldron, D. B. (2013). Conducting functional analysis of problem behavior via telehealth. *Journal of Applied Behavior Analysis, 46,* 31-46.