Intervention research: appraising study designs, interpreting findings and creating research in clinical practice

Susan H. Ebbels[1,2]

[1]Moor House School & College, Surrey, UK

[2]Division of Psychology and Language Sciences, University College London, UK

Correspondence concerning this article should be addressed to Susan Ebbels, Moor House School & College, Mill Lane, Hurst Green, Oxted, Surrey, RH8 9AQ, UK.

ebbelss@moorhouseschool.co.uk

Abstract

Speech-language pathologists (SLPs) are increasingly required to read, interpret and create evidence regarding the effectiveness of interventions. This requires a good understanding of the strengths and weaknesses of different intervention study designs. This paper aims to take readers through a range of designs commonly used in speech-language pathology, working from those with the least to most experimental control, with a particular focus on how the more robust designs avoid some of the limitations of weaker designs. It then discusses the factors other than research design which need to be considered when deciding whether or not to implement an intervention in clinical practice. The final section offers some tips and advice on carrying out research in clinical practice, with the hope that more SLPs will become actively involved in creating intervention research.

Introduction

Evidence-based practice is key to providing the best possible service for our clients. In order to deliver evidence-based practice, clinicians need to integrate individual clinical expertise and their clients' values with the best available clinical evidence (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996). Therefore, it is crucial that clinicians are able to identify the best available research evidence by reading the literature and applying a sound knowledge of the strengths and limitations of different intervention study designs.

In some areas of speech-language pathology practice, however, the intervention evidence is very limited. Thus, speech-language pathologists (SLPs) may need to use evidence that is only partially related to their clinical situation and to place more reliance on their clinical expertise while waiting for more relevant evidence to emerge. An alternative solution is for SLPs to create their own evidence. SLPs who investigate the effectiveness of interventions delivered in their particular setting and with their particular client group create evidence which is highly relevant for that situation and client group, while also increasing their own ability and confidence in making evidence-based decisions. This can lead to more effective intervention and hence improved outcomes for their clients.

Practising SLPs may be anxious about carrying out research and feel this is best left to those working in universities who have more research skills and time to devote to research. While this may be the case, intervention studies can be very time-consuming and costly due to the labour-intensive process of administering repeated assessments and providing intervention. Thus, limited numbers of intervention studies are likely to be funded. However, practising SLPs are already carrying out assessments and intervention, so collaborations between practising SLPs and universities could significantly reduce the costs of intervention studies, as the intervention is already being provided,

funded from elsewhere. Such collaborations therefore have the advantage of creating intervention research which is highly clinically relevant and in a cost effective manner, while drawing on the research expertise of university-employed staff.

Combining theoretical and research experience with clinical experience can benefit intervention studies as well as increasing the skills and knowledge of those involved. Snowling and Hulme (2011) argue for a "virtuous circle" linking theory with practice, whereby theory leads to the formulation of possible interventions, which are then evaluated in intervention studies with strong designs, the results of which are used to inform and refine theory. I would add that clinical experience also has a role to play and can contribute to the formulation of theoretically well-founded interventions. Clinicians will often have insights into the practicalities of delivering interventions that could help improve the effectiveness of those interventions, for example, how long and frequent sessions should be, how often the focus of activities needs to change to keep clients' attention and other tips for motivating clients and potentially boosting learning. When the intervention has been evaluated, the results can extend the SLP's clinical experience. Thus, a double "virtuous circle" could be created where both theory and clinical experience help to formulate interventions and the results of those interventions inform and improve both theory and the clinical experience of those involved.

Given the value of SLPs being involved in intervention research, both as consumers (reading and understanding the literature and applying relevant findings to their clinical work) and increasingly as (co-)creators of intervention research, it is important they have sufficient knowledge of intervention research design. This paper aims to provide SLPs with some of that knowledge by discussing the strengths and limitations of intervention study designs commonly used in speech-language pathology with the aim that SLPs will be better able to critically appraise studies they read and also that some will use this information to help them design and carry out research studies within their clinical practice which are as robust as possible.

4

My intervention research experience and knowledge is primarily with children with Developmental Language Disorder (DLD) and therefore, many of the examples of studies I provide will relate to this client group. However, this paper aims to be relevant to those working in a range of client groups and settings.

Intervention study design

The design of an intervention study is fundamental to its robustness and reliability and needs to be planned carefully in advance. When carrying out intervention studies in clinical settings, many factors are at play, only some of which relate directly to the intervention itself. Thus, in order to separate the effects of the intervention from the effects of other non-specific factors, we need studies which control for as many of these as possible. Some designs are much more robust than others as they control for more of the spurious factors which could influence outcomes. Involving larger numbers of participants also increases reliability and the ability to generalize the findings to other people, but the size and degree of experimental control of a study interact to improve reliability, with experimental controls being the more crucial element.

INSERT FIGURE 1 ABOUT HERE

Figure 1 shows a schematic view of this: increasing numbers of participants are shown on the x axis and designs with increasing levels of experimental control on the y axis. Also marked on Figure 1 are four hypothetical studies: Studies A and B have good experimental control, but Study B has many more participants than Study A; Studies C and D on the other hand have poor experimental control but D has more participants than C. The most reliable of these four studies is Study B with a good experimental control and large numbers of participants and the least reliable is Study C, with a weak experimental control and few participants. Studies A and D however, have different strengths and weaknesses and following positive preliminary results in studies with these designs, further research would need to be conducted to increase confidence in the results. A positive finding in Study A

would need to be replicated with more participants. Similarly, a positive finding in Study D would need to be replicated with greater experimental control. However, a clinician may need to make clinical decisions based on evidence from studies such as A and D before more reliable studies have been carried out. In this case, a small study with good experimental control is likely to be more reliable than a large study with weak control, but both need to be treated with caution. In terms of carrying out studies, it could be argued that Study D would waste resources (by involving a large number of participants, but in an experimental design likely to produce unreliable results) and that Study A, which would be cheaper, may therefore be the better option.

For SLPs who are designing intervention studies, it is important to try to maximize both the number of participants and the experimental control. If only a fixed number of participants are available, it is particularly important to try to maximise the degree of experimental control. Conversely, if a particular design has to be used (maybe due to practical restrictions), maximising the number of participants is important. Later, I discuss different experimental designs and the level of control they provide, starting with the least robust. For each, I first discuss the design in terms of timings and types of assessments relative to the intervention and then what factors each design can and cannot control for.

In addition to the overall design of an intervention study in terms of timings of assessments and interventions, other features are also important and should be considered by SLPs who are appraising a study carried out by others or planning to carry out a study themselves. These features include: how representative the participants are and how outcomes are assessed. In general, findings can only be generalised to participants who are similar to those in the original study. In order to investigate the effectiveness of the intervention in other groups, further studies will need to be carried out. Assessment of outcomes is complex. The tests need to be appropriate to the research question and the participants and sensitive to the intervention. For example, if an intervention is hypothesised to cause a change in a very specific area of language, but the outcome

measure is a standardised test which only includes one question relating to the specific area, change on that measure is unlikely, even if the intervention has caused large changes in the specific area of language targeted. Thus, it is often necessary to create tests specifically for an intervention study.

Generalisation of new skills may also be important to assess. This may include generalisation to standardised tests, but may also be to other areas of language and/or educational performance, or to other situations such as general conversation or performance in the classroom. It is important to consider in advance how much change you would expect or desire in these areas, again this comes back to the research question. If the main aim of an intervention is to improve performance in the classroom, this would be the primary outcome and crucial to measure. However, if the aim is to improve a small area of functioning with a very short intervention, a change in classroom performance may not be expected as this may require many more hours of intervention, and thus may not be relevant to measure.

Assessment should ideally be carried out "blind". This means that the assessor does not know how individual participants fit into the design of the study. Thus, they may not know which participants have versus have not had intervention, or they may know the participant has had intervention, but not which items in the test battery have been targeted versus not targeted. Having blind assessors reduces the chance of bias, both during the assessment and scoring process. In our research, we have sourced blind assessors from various places: student SLPs who are on placement, or who come on a voluntary basis in order to gain experience of research (Ebbels, Nicoll, Clark, Eachus, Gallagher, Horniman et al., 2012), SLP assistants within the team who have been kept blind the content of the participants' intervention sessions (Ebbels, Maric, Murphy, & Turner, 2014) or SLPs swapping with other SLPs in the same team who again are unaware of the precise nature of the intervention each participant has received (Ebbels, Wright, Brockbank, Godfrey, Harris, Leniston et al., in press).

At a minimum, assessments should be carried out before and after intervention (methods of increasing experimental control are discussed below). However, it might also be important to test

whether new skills are maintained after a period of time. Intervention studies often have a hypothesis that intervention will improve skills, but what happens after intervention ceases is also of interest; new skills may diminish (i.e., the intervention has only a short-term effect), or they may remain stable (i.e., the gains from intervention are maintained), or they may even improve further (i.e., the intervention has triggered a change which continues after the intervention has ceased).

Degree of experimental control

In sections 1-10 below, I discuss in turn each experimental design shown in Figure 1 and their strengths and limitations, working from those with the least to most experimental control.

1. Limitations of anecdotes and clinical experience

SLPs' clinical experience together with information and anecdotes from colleagues are used more frequently than other sources of information for guiding their intervention decisions (Nail-Chiwetalu & Ratner, 2007; Zipoli & Kennedy, 2005). However, while such information may provide a useful starting point in considering whether to use an intervention, anecdotes and clinical experience alone are subject to considerable bias. We are all liable to fall for the "therapeutic illusion" (Casarett, 2016; Thomas, 1978), whereby everyone involved in an intervention (both clinicians and patients), believes the intervention is more effective than it actually is. We may interpret a change on our measures as an intervention effect, when it may actually be random variation, a "placebo effect", natural history, other factors unrelated to the intervention, or "regression to the mean". Regression to the mean is a phenomenon in which extreme test scores tend to become less extreme (regress to the mean) when the test is repeated. This is a problem when participants or targets have been chosen for intervention due to low levels of performance on a measure which is subsequently used to evaluate progress. Imagine a scenario where child A has a 'true' standard score on a standardized test of 90, but happens to score 83 on a particular day. If intervention is provided for all children with scores below 85, child A would receive intervention. However, at the next test, child A's score would be more likely to be near their true score of 90. This would appear to be an improvement, when in

actual fact it is merely due to random variations in their scores. Conversely, consider child B whose true score is 80 but who happens to score 87. Child B's subsequent score would be expected to be more similar to their true score of 80 (i.e., decrease) at the next test point. When evaluating the performance of a group, normally child B's spurious decrease would cancel out child A's spurious increase, but not if child B has been excluded from intervention due to a pre-intervention score above the cut-off. Now, imagine a study which includes several (or many!) children whose pre-intervention scores fall on the opposite side of the cut-off to their true scores. If this study gives intervention only to the half with artificially low pre-intervention scores and not to the half with artificially high pre-intervention scores, the intervention group is likely to have on average higher scores post-intervention, but this spurious increase in scores is purely due to random variation and regression to the mean of extreme scores; it is not an effect of intervention.

Thus, clinicians need to recognize that clinical practice which relies on just anecdotes and experience could be flawed and lead to clinical experience which consists merely of "making the same mistakes with increasing confidence" (Isaacs & Fitzgerald, 1999; O'Donnell & Bunker, 1997). In order to avoid this, we need to look to studies which aim to reduce some of the biases to which we are all susceptible.

2. Change in raw scores

A first step to reducing bias when evaluating an intervention, is to measure performance before and after intervention on a measure which is relevant to the intervention. In order to reduce bias, this should be carried out in the same way on both occasions (e.g., same test items, scoring and rating procedure, situation and tester) and ideally by a tester who is "blind". Asking those involved with the client (including the SLP) if they think there is improvement can give some measure of functional improvement, but this is again subject to the "therapeutic illusion", especially when they have been closely involved in the intervention.

Interpreting changes in raw scores

Assuming that two raw scores have been obtained, one pre- and one post-intervention, the next question is: what do these results mean? Do they show good progress? The post-intervention value being higher than the pre-intervention value may or may not mean good progress has been made. This depends on the degree of difference between the two scores, what the two scores represent and whether this difference is important. For example, a difference of five between two scores might be important if this represents a change on a test of understanding classroom instructions from 3/8 to 8/8 or a change in life expectancy from 50 to 55 years. However, if the change is from 50% to 55% on correct production of a target phoneme in words, this may not be important to the client and also may just be random variation in performance from one testing point to the next. Statistical tests are available for measuring whether a change on a test which is carried out twice is significant. For an introduction to suitable tests aimed at SLPs see Pring (2005).

Let us assume that our pre- and post-intervention test raw scores differ significantly. In these circumstances, can we infer that the intervention has been effective? No. It may be that the intervention was effective, but it is also possible that an array of other factors unrelated to the intervention have led to the increase in score.

Limitations of raw scores: other factors could be responsible for 'progress'

For children, maturation and general development are likely explanations for many changes in performance. As children develop cognitively, physically and emotionally and gain in experience of the world, merely by being alive in the world, we would expect performance to improve in most areas. In addition, most children are receiving education, both formal (in schools and nurseries) and informally at home and elsewhere. Thus, it is important to know what you would expect in terms of change for a child in a similar situation of a similar age not receiving the intervention. To interpret an intervention as being effective, the progress with intervention needs to be greater than that which would be expected without the intervention. Natural history is also important in more medical

situations, where some spontaneous recovery might be expected, so successful interventions would need to show that they have accelerated that recovery. For clients with degenerative conditions, a successful intervention may slow the rate of decline. Thus, in all client groups, it is crucial to be able to compare changes with intervention to changes which would have been expected if the intervention had not been provided.

Another factor which is important to consider with repeated measurements is regression to the mean and practice effects. To reduce regression to the mean, studies should avoid selecting items or participants based on particularly low scores on the first assessment or use different measures for identification of participants from the measure(s) used to evaluate progress (Zhang & Tomblin, 2003). If the same assessment needs to be used for identification and evaluation of progress, studies could include a baseline period, so regression to the mean occurs before intervention starts (see sections 4, 6, 7, 9). Alternatively, studies could use or control areas which have similar pre-intervention scores to the target area, or control items, which are selected using the same criteria, but which are not treated (see section 5, 6, 7, 9). The most common method is to use control participants, who are identified using the same criteria, but do not receive intervention. Thus regression to the mean should be similar in both the intervention and control groups (see sections 8-10). In addition, mere experience with a task could also improve performance on the second occasion due to practice effects, even if underlying skills have not improved. To control for practice effects, a study could test participants on control items the same number of times as target items, so a practice effect would affect both targets and controls (see sections 5-7), or test control participants on the test items without providing them with intervention (see sections 8-10)

Thus, in order to conclude that an intervention has been effective, we need to know whether progress is different from what would be expected without the intervention given other potential factors (natural history, maturation, regression to the mean, practice or placebo effects, other interventions / education they are receiving). The different designs described in sections 3-10

control to a greater or lesser extent for each of these and we will go through these designs from the least to most robust and discuss the degree to which they control for these different factors.

3. Change in standard score

Standard scores on standardised tests can help to control for maturation and general world experience in children. Increasing standard scores would indicate that a child is progressing at a faster rate than the children in the standardisation sample and thus progress is greater than would be expected given general maturation and world experience.

Therefore, if a child has low performance on a standardised language test, for example, their SLP could look to see both whether both the raw and standard scores improve. If their raw scores improve, this indicates progress relative to their own pre-intervention scores, but despite improving raw scores, their standard scores may decrease or remain stable, or indeed they may increase. If their standard scores increase, this shows that they are "catching up" or "closing the gap" with their typically developing peers, if they remain stable, they are making progress at the same rate and if they decrease, the gap is widening.

Standard scores provide information about performance relative to the children in the standardisation sample of the test. It may be, however, that for a particular group of children, different patterns of progress are expected. Again, it is important to know the natural history for particular groups. For example, studies have shown that for children with DLD, with respect to their understanding of vocabulary, the gap tends to widen with age between their performance and that of typically developing children (Rice & Hoffman, 2015). This widening gap is despite increasing raw scores and is probably due to a slower rate of vocabulary learning among this group, relative to the efficient vocabulary learning of typically developing children and teenagers. In other areas of language, such as expressive language, the trajectories of children with DLD parallel those of typically developing children (Conti-Ramsden, St Clair, Pickles, & Durkin, 2012). Thus stable standard

scores are expected. If, in contrast, a study finds increased standard scores, this could indicate that progress in this area has accelerated.

Limitations of standard scores

While standard scores can control for maturation, they do not control for practice effects (although standardised test manuals usually provide a time period after which you would not expect a practice effect) or for other random or predictable factors such as other intervention or teaching which the client may be receiving. Thus, while it may be possible to say that a child or group of children is making faster than expected progress, it is not possible to say what factors underlie this progress. Regression to the mean may be a problem when children have been selected for a study purely on the basis of their low standard scores pre-intervention and progress with intervention is measured on the same test (Tomblin, Zhang, Buckwalter, & O'Brien, 2003). This is less of a problem when they have been selected on a different test or criteria, or if the pre-intervention test is carried out more than once (in which case, regression to the mean would occur before intervention started).

4. **Within participant control (single baseline)**

Some studies control for natural history and regression to the mean by using a baseline period. The design of these studies is shown in Figure 2. These can be used for a single case or for a group of participants.

INSERT FIGURE 2 ABOUT HERE

In this design, the same assessment is carried out at least twice before intervention starts. This provides information about the rate of progress without intervention. This period of no intervention before the intervention starts is known as the baseline period. If the intervention has no effect, we would expect a similar rate of progress during the baseline and the intervention period. If the baseline period is a similar length to the period of intervention, then no effect of intervention would be shown by a similar degree of change between assessments 1 and 2 as between assessments 2 and 3. In contrast, a change of slope in the intervention period compared with the baseline period

13

could be due to the intervention (see Howard for a description of how to analyse this statistically within a single subject). For examples of this design used with a group see Zwitserlood, Wijnen, van Weerdenburg, and Verhoeven (2015), Bolderson, Dosanjh, Milligan, Pring, and Chiat (2011), Falkus, Tilley, Thomas, Hockey, Kennedy, Arnold et al. (2016) and Petersen, Gillam, and Gillam (2008) and for examples of studies with single cases see Riches (2013) and Kambanaros, Michaelides, and Grohmann (2016).

SLPs thinking of using this design need to plan ahead so that they can carry out a least two tests prior to starting intervention. Ideally, if only two pre-intervention assessments are being carried out, the gap between these should be similar to the predicted length of the intervention in order to control for maturation. For SLPs working in schools, school holiday periods can work well as baseline periods. If the first assessment is carried out before the holidays start, the second assessment and the intervention can take place as soon as school resumes.

This design can help control for maturation (as long the rate of change due to maturation is expected to be stable during the time period of the study), regression to the mean and practice effects (unless the practice effect is cumulative such that it is stronger each time you repeat the assessment).

Limitations of single baseline design

Even if the slope during the intervention period is significantly different from during the baseline period, the single baseline design only provides limited control over other random or predictable factors. The change between the baseline and intervention period could be due to a placebo effect (where merely seeing an SLP may lead participants to expect they will make progress, thus changing their motivation and effort, leading to increased scores even though their underlying skills are unhanged) and or may coincide with some other change in the client's life (e.g., motivation, health, home or education situation, changes in other interventions or education being provided) which may be exerting a general effect on their performance in all areas, including the area being

measured. It could be these other non-specific factors which are leading to the change in slope, rather than the content of the intervention per se. In those situations where a withdrawal of intervention is likely to lead to a withdrawal of the effect, a reversal design can be used. In this case, if withdrawal of intervention leads to a reversal of performance trends, greater confidence can be placed in the efficacy of the intervention. However, a reversal of intervention effects after intervention has ceased is virtually never a desired or expected outcome in SLP and thus the withdrawal design is of limited use to the profession and as such, other designs are preferable.

5. Within-participant design with control items/area

In situations where all participants will receive intervention (i.e., there is no control group), a certain degree of experimental control can be gained by comparing progress on areas or items you are targeting versus areas or items you are not targeting and do not expect to improve. This design is shown in Figure 3.

INSERT FIGURE 3 ABOUT HERE

Both the control and targeted items/areas are tested pre- and post-intervention. In this design, the comparison of interest is the difference in the progress made on targets versus controls. Any progress seen on the controls could be due to general maturation, placebo or practice effects and/or other non-specific factors which would be expected to affect both the targets and controls. Any additional progress seen only on the targets is likely to be related to the intervention. It is important that pre-intervention performance on targets and controls is similar as this reduces regression to the mean and aids statistical comparisons and interpretation of the results. This design can be strengthened if the targets and controls are counter-balanced across participants, such that the control areas/items for some participants are the targets for others and vice versa.

For examples of studies which have used this design with single cases see Parsons, Law, and Gascoigne (2005), for group studies which have combined a range of targets see Mecrow, Beckwith,

and Klee (2010) and Ebbels et al. (in press) and for studies with counter-balancing of targets and controls across participants see Wilson, Aldersley, Dobson, Edgar, Harding, Luckins et al. (2015).

Limitations of the within-participant design with control items or areas

This design can control for a wide range of factors. However, the choice of control items / areas is crucial as the design relies on finding a difference in progress between targets and controls. If progress on the targets generalizes to the control items/area, the experimental control may be under threat. If the generalization is relatively limited, such that targets still show more progress than controls, experimental control is maintained. However, if progress generalizes to such an extent that targets and controls show equal progress, experimental control is lost. Equal progress on targets and controls could be due to generalization (which is clinically desirable) or could be due to maturation, placebo or practice effects and/or other non-specific factors. In this situation, even if both targets and controls show good progress, it is impossible to draw conclusions regarding the effectiveness of the intervention. Thus, it is crucial that when choosing control areas/items, generalization is not expected.

If SLPs wish to consider the effects of generalization, additional control needs to be added to this design, such as a control (baseline) period (see sections 6 and 7), or a control group (see sections 8-10).

6. Within-participant design with single baseline and control items/area

This design combines the two previous designs, using both a baseline period and control items/area and is shown in Figure 4. Thus, if targeted items/area improve more with intervention than before intervention and more than control items/area, this controls for maturation, placebo or practice effects, regression to the mean and other factors which would be expected to improve the control as well as the targeted items/area.

This design has advantages over the use of control items with no baseline, as a change in controls with intervention which is greater than the change during the baseline is more likely to be due to

generalisation than to practice effects or general maturation. A example of single case studies or case series using this design are Kulkarni, Pring, and Ebbels (2014) and Best (2005).

INSERT FIGURE 4 ABOUT HERE

Limitations of within-participant designs with single baseline and control items/areas

While this design is stronger than previous designs, as changes seen in the control items during intervention but not during baseline are unlikely to be due to maturation and practice effects, they could still be due to a placebo effect or other factors which could be occurring in the client's life around the time of changing from baseline to intervention. If the changes only occur in the targeted items/areas and not the controls, it is likely that these are due to the intervention, but if they also occur in the control items or areas, this weakens the design as it this could be due to generalisation, or to other factors. Thus, as before, it is crucial to choose control items/areas to which generalisation is not expected, otherwise experimental control can be lost.

7. **Within-participant multiple baseline design**

The key feature of a multiple baseline design is a staggered start to intervention. When used within participants, it may be different items/areas which receive intervention but at different times. This design is essentially the same as the previous design except the control items also receive intervention but at a later date. This is illustrated in Figure 5. Thus, a baseline period is used (with at least two testing points), followed by intervention for Target A, while Target B is held in an extended baseline. Following intervention for Target A, Target B is treated. Maintenance of Target A may also be assessed at the final assessment point. If Target A improves more with the first intervention than during baseline and more than Target B, this design controls for maturation, placebo and practice effects. If Target B also improves more during its intervention period than during its baseline, this provides better control for other factors. This is because, if both Targets A and B improve only when their specific intervention is provided and not before, it is less likely that non-intervention-specific

factors are causing these specific changes. An example of a study using this design with a case series is Culatta and Horn (1982)

INSERT FIGURE 5 ABOUT HERE

Limitations of within-participant multiple baseline design

This design has similar limitations to the previous designs: if Target B improves during intervention for Target A, (but not baseline) this still controls for maturation and practice effects, but a change while still in extended baseline (while Target A is receiving intervention) could either be due to generalization or other factors, including a placebo effect. In order to control for other factors such as activities happening in classroom education, other children in the same class could act as controls, as general classroom activities should affect their performance, but generalization from intervention would not. Such an addition would then include comparisons between participants (see sections 8-10).

8. Between participants comparisons (with non-random assignment)

Including as control participants other clients who have similar profiles and are in similar settings can control for some of the effects of other non-specific factors and allow more reliable investigations of the effects of generalisation. The most common design is to administer a pre- and post-intervention measure to two groups of participants, but only provide intervention to one group. The crucial comparison is between progress made by the intervention group and that made by the control group. This design is shown in Figure 6. If the groups are very similar pre-therapy and the intervention group make more progress than the control group, this controls for maturation, practice effects and other factors which the two groups have in common, as these would be expected to affect the performance of both groups.

INSERT FIGURE 6 ABOUT HERE

In order to make comparisons across groups with small numbers of participants, a between participants multiple-baseline design can be used. This is similar to the within-participant multiple baseline design (see Figure 5), except that it is participants rather than targets which have variable baseline lengths. Thus, a single area may be targeted, but in more than one participant, with staggered starts to intervention. If the slope of performance changes only when intervention is introduced for each participant, with increasing numbers of participants this makes it more likely that the intervention itself is responsible for the change. For an example of a study using this design, see Petersen, Gillam, Spencer, and Gillam (2010).

Limitations of between participants comparisons with non-random assignment

The main limitation of group comparisons of intervention and control participants is that the two groups may differ from each other in ways which are predictable (e.g., different classes, schools, teachers, abilities, backgrounds, other help/support) or unpredictable. Even if all obvious factors are balanced between the groups, they may still differ in ways which have not been considered. Therefore, differences between the groups in the amount of progress made during the intervention period (for the intervention group), may be due to differences between the groups rather than to the intervention. An example of this possible limitation is a study such as Motsch and Riehemann (2008), where the teachers of the experimental group volunteered for an advanced course and those of the control group did not, hence the teachers may have differed in fundamental ways (e.g., motivation) which could have affected the participants' response to intervention more than the nature of the intervention itself.

The best solution to this problem is to randomly assign participants to the groups as, if the numbers are big enough, all potential factors should balance out between the groups (see section 10). Another approach, especially with small numbers, is to combine a between-participant and within-participant multiple baseline design (see section 9). An alternative solution is to provide the control group with intervention after the experimental group has stopped receiving intervention (i.e., the

controls become a *waiting* control group). If they also respond to the intervention when they didn't during their extended baseline it is less likely that other non-specific factors account for the differences in progress between the groups after the first phase of intervention. Adding intervention for a waiting control group, then becomes a similar design to the between-participants multiple baseline designs (see Figure 5) often used for (a series of) single cases, where the waiting controls are in effect held in an extended baseline and have a staggered start to intervention.

This design does not usually control for a placebo effect. However, this can be controlled for by providing non-intervention-specific special attention to the control group instead of just no intervention. This could even be intervention but on a different, unrelated area (which is not expected to generalise to the area under investigation). Indeed, in our research, we frequently use this approach as all children in our setting have to receive intervention, so our (waiting) controls receive intervention in a different area to that being investigated in the study, rather than no intervention. This avoids the ethical dilemma of involving participants in a study who receive no intervention whilst also controlling for possible placebo effects.

9. Combined between and within participant designs

Some group studies (e.g., Smith-Lock, Leitao, Lambert, & Nickels, 2013) add in within-participant control by adding a baseline period for both the intervention and control groups. This follows a similar pattern to Figure 4 but it is participants rather than items/areas which act as controls by receiving either no intervention or, as in Smith-Lock et al. (2013) receiving intervention in a different area, thus controlling for the placebo effect. This study also included a control measure for the experimental intervention group, so placebo and non-specific effects were controlled both between and within participants. Such additions strengthen the design and also allow researchers to look at the performance of individuals within each group.

For studies with small numbers of participants, a multiple baseline design both between and within participants is a strong design (see Figure 7). At least two participants are involved, but increased

numbers improves reliability and generalizability and also introduces the possibility of comparing performance across groups. In this design, all participants undergo a baseline period with at least two assessment points, then the two (groups of) participants receive intervention, but on different targets. After a period of intervention, they both swap to the other target. If progress is seen on each target only when it is targeted, it is likely that it is the intervention which underlies the change rather than other non-specific factors (which would be expected to affect both targets regardless of the focus of intervention). This is even more likely when the targets and participants are randomly assigned to the different periods of intervention and when more participants are included. Ebbels and van der Lely (2001) used this design, albeit without randomisation.

INSERT FIGURE 7 ABOUT HERE

Limitations of combined between and within participant designs

As with previous within-participant designs, it is important that generalization does not occur between the two targets. If intervention on either target improved performance equally on both targets, the design would in effect be reduced to a single baseline design (see section 4), which has less experimental control and where conclusions regarding the effectiveness of the intervention are harder to draw. Thus, it is essential that the target areas are chosen very carefully such that generalization between them is not expected.

10. **Between participant design (randomised control trial)**

The Randomised Control Trial (RCT) is seen as the "gold standard" intervention design, because with sufficiently large numbers and random assignment all potential factors other than the intervention become evenly distributed between the groups and are thus unlikely to affect the results. The design of an RCT at its simplest is represented in Figure 6. This design is feasible within clinical practice, although it is easiest where intervention is 1:1. For example, if a group of clients are all due to receive a period of intervention; they could be randomly assigned to "experimental intervention" versus "control" groups and assessed before and after intervention is provided. A "control group"

can take various forms, which SLPs may view as more or less ethically acceptable. They could receive no intervention (e.g., Fey, Cleave, & Long, 1997), or "treatment as usual" (e.g., Adams, Lockton, Freed, Gaile, Earl, McBean et al., 2012; Boyle, McCartney, O'Hare, & Forbes, 2009; Cohen, Hodson, O'Hare, Boyle, Durrani, McCartney et al., 2005), or intervention in a different area (e.g., Ebbels, van der Lely, & Dockrell, 2007; Mulac & Tomlinson, 1977) or a non-specific intervention (e.g., the "academic enrichment group" in Gillam, Loeb, Hoffman, Bohman, Champlin, Thibodeau et al., 2008) which are not predicted to affect the target area. Alternatively, they could be a "waiting control group", who also receive the experimental intervention after intervention for the "experimental intervention" group has finished. During the "experimental intervention" group phase, these waiting controls could either receive no intervention (e.g., Fey, Cleave, Long, & Hughes, 1993; Fey, Finestack, Gajewski, Popescu, & Lewine, 2010; Fricke, Bowyer-Crane, Haley, Hulme, & Snowling, 2013), or they could receive intervention on a different, unrelated area which is not expected to affect the target area (e.g., Ebbels et al., 2014; 2012). Studies vary in whether they report the progress made by the waiting control group (which delays publication of the study, but strengthens the findings), or only the results after the first stage of intervention for the experimental group. Clinicians often worry about the ethics of control groups. In my view, if there is as yet no evidence an experimental intervention may be effective; it is perfectly acceptable to withhold this intervention for the purposes of a study which could contribute future evidence. Indeed, waiting control groups may get the best deal, particularly if they only receive the experimental intervention if the first phase of the trial indicates it is effective and not if there is doubt about its effectiveness.

This design can also be extended to investigate generalisation by including assessments of items or areas where generalisation is expected. Both groups of participants are tested on both target and generalisation items, but only the intervention group receives intervention. If this group improves on both controls and targets, but the control group do not, it is likely that the progress on the generalisation test is due to the intervention. It could also be due to a placebo effect, but this could be controlled by giving intervention to the control group on another area at the same time. Findings

22

from RCTs can be further strengthened by using a waiting control group, who then go on to receive intervention. If they also make progress after intervention, but not while acting as controls, this strengthens the conclusion that the intervention is effective. We carried out an RCT using this design (Ebbels et al., 2014) and included both a control structure (where we did not expect generalisation) and a generalisation test (where we were specifically looking for generalisation when the target intervention was received). These extensions to the basic design in Figure 6, however, while strengthening the design, do make it much more complex and thus more difficult to carry out. As an example of an extended and more complex design see Figure 8 for the design of the Ebbels et al. (2014) study.

INSERT FIGURE 8 ABOUT HERE

Limitations of RCTs

Randomised control trials are the most robust design. However, it is important that the numbers in the randomisation sample are sufficient that randomisation is likely to have led to a balance of all potential influencing factors between the groups. If a study has too few participants, the design in section 9 may be more appropriate. Ideally, randomisation would be carried out at the level of the individual, but in some studies this is not possible. For example, an intervention involving training of education staff may need to be carried out at a school level. While schools could be randomised to different groups, the students within those schools have not been randomised and thus large numbers of schools would be required for potential factors to be evenly distributed between the groups. This design (known as a cluster randomised control trial) is complex to design and analyse but the majority of such studies do not account for clustering in their design or analysis (Campbell, Elbourne, & Altman, 2004). For example, a trial involving two schools which are randomised one to receive and one not to receive intervention (such as Starling, Munro, Togher, & Arciuli, 2012) is not an RCT as the participants are not randomly allocated to schools, so there is no guarantee that the

two schools, the staff teaching in them and the students attending them do not differ in some important ways (indeed this is very likely).

As with other designs, placebo effects can only be controlled for if the control group receives some kind of "intervention"; this could just be special attention, or intervention on another unrelated topic.

Critical appraisal of studies

The design of a study can be appraised in terms of its robustness without reading the results or discussion. Indeed some suggest (Greenhalgh, 1997) that readers should decide whether or not to read a paper by first reading the method only and if the design is insufficiently robust, to abandon reading the rest of the paper as it probably "belong(s) in the bin" (p.243)! When considering the robustness of the design, the reader needs to consider: the degree of experimental control provided (see above) and the number of participants (generally greater numbers of participants increases reliability). For studies with a robust design and large number of participants, more confidence can be placed in the results (see Figure 1), whether those results are in favour of the intervention studied, or not.

Having decided that a study has a robust design with a sufficient number of participants to produce reasonably reliable results, the reader needs to consider other points before deciding whether or not to use the intervention in clinical practice. The first is whether the results are statistically significant and the degree of significance (lower p-values are more significant). In general, a marginally significant result (e.g., p=0.045) should be considered with more caution than a highly significant result (e.g., p=0.002). The second factor to consider is the magnitude of the effect (see Cohen, 1988 for more information regarding effect sizes) and whether the effect is relevant to the clients (i.e., is it clinically significant?). The third factor interacts with consideration of the size of the effect and this is the amount and cost of the input which is required to obtain that effect. An

intervention which has a small, but clinically relevant effect and costs very little to implement may be as worthwhile to include in clinical practice as an intervention with a very large, clinically very important effect with a high cost. However, interventions with small effects and high costs may not be appropriate to include in clinical practice, even if they have statistically significant results. This is particularly the case if other interventions have similar effects for lower costs, or larger effects for the same cost. The final factor to consider is how similar the participants in the study are compared with to those in the SLP's clinical practice. If the differences are too great, the study may be irrelevant to the SLP's client group. However, if an SLP's clients are similar in some ways to the study participants but different in others, it may be worthwhile trying the intervention. In this case, however, the SLP should evaluate the effectiveness of the intervention with their different client group.

How can I start to be research active and what support do I need?

For an SLP with a regular caseload, only a few tweaks may be needed to turn standard intervention into a research project. All designs can be carried out as part of routine practice if everyone involved is willing to be flexible and committed to the purposes of the project. Measuring indicators of outcomes (what you want to achieve) before and after intervention is good clinical practice and can form the beginnings of research. Thus, there is no definite line between research and good clinical practice, but research generally includes greater controls. Even RCTs are feasible as part of clinical practice and don't need to have huge numbers of participants if you are only interested in large effects. Indeed, in my experience, I have found small-scale RCTs (e.g., Ebbels et al., 2014; 2012) easier to carry out than within-participant designs (Ebbels et al., in press). This is particularly true where generalisation might be expected, as identifying suitable controls areas or items can be very difficult.

The main requirements for carrying out research in clinical practice are time and support. Time is needed for staff to develop research skills, and to design and carry out projects. Planning time needs to be built in and time spent at the planning stage can dramatically improve the usefulness of a project. The research design needs to be carefully thought through to maximise the robustness of the design given practical constraints. Assessment and intervention plans, materials and resources may need to be created specifically for a project. Those carrying out the intervention (and assessments) will need training to ensure they carry these out to the requirements of the project (treatment fidelity). It may be necessary to source "blind" assessors from outside your organisation (SLP students can be a good source of assessors); this will also take time to organise. Inclusion of your research project in your appraisal or progress review may allow for ring-fencing of time and increased motivation to prioritise the project on all sides. In my organisation, half a day a week of dedicated time has proved sufficient for clinicians to plan and coordinate small-scale research projects, while larger scale projects have required more dedicated time. The participants involved in a project will also need to commit more time to a project than to usual intervention. This is mainly due to the increased number of assessments required for more rigorous designs. They may also be required to attend for more intervention. Hopefully, if the study is theoretically and clinically well-motivated, this increase in time on their part will result in better outcomes for them, which is ethically more acceptable.

Carrying out a research project in clinical practice also requires support, particularly from the management in your organisation. This is more likely to be forthcoming if your proposed research is of direct clinical relevance to your service. However, you may also need the support of your colleagues (particularly if they will be providing some of the intervention). Administrative support would also be helpful. A crucial element, however, is to gain support from someone with research expertise who can provide advice prior to the study on research design including how many participants may be required, ethical requirements and options for analysis. On completion of your study they can also advise on dissemination of your findings.

Conclusions

Clinical practice of SLPs will be improved if we all incorporate aspects of evidence-based practice into our work. Whether we are interpreting the research studies of others, or designing our own, we need a good understanding of research design and an ability to recognise weaknesses in intervention studies which may reduce the reliability of study findings. Striving to maximise both the robustness and clinical relevance of intervention studies and ensuring that SLPs have the time, skills and support to read and (co-)create research and apply relevant findings to their clinical practice, should be a priority for the profession.

Adams, C., Lockton, E., Freed, J., Gaile, J., Earl, G., McBean, K., Nash, M., Green, J., Vail, A., & Law, J. (2012). The Social Communication Intervention Project: a randomized controlled trial of the effectiveness of speech and language therapy for school-age children who have pragmatic and social communication problems with or without autism spectrum disorder. *International Journal of Language & Communication Disorders, 47*(3), 233-244. Retrieved from WOS:000302941900001

Best, W. (2005). Investigation of a new intervention for children with word-finding problems. *International Journal of Language & Communication Disorders, 40*(3), 279-318.

Bolderson, S., Dosanjh, C., Milligan, C., Pring, T., & Chiat, S. (2011). Colourful semantics: A clinical investigation. *Child Language Teaching & Therapy, 27*(3), 344-353. Retrieved from WOS:000294864600007

Boyle, J. M., McCartney, E., O'Hare, A., & Forbes, J. (2009). Direct versus indirect and individual versus group modes of language therapy for children with primary language impairment: principal outcomes from a randomized controlled trial and economic evaluation. *International Journal of Language & Communication Disorders, 44*(6), 826-846. Retrieved from WOS:000275345700002

Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). CONSORT statement: extension to cluster randomised trials. *Bmj, 328*(7441), 702-708. doi:10.1136/bmj.328.7441.702

Casarett, D. (2016). The Science of Choosing Wisely — Overcoming the Therapeutic Illusion. *New England Journal of Medicine, 374*(13), 1203-1205. doi:doi:10.1056/NEJMp1516803

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, W., Hodson, A., O'Hare, A., Boyle, J., Durrani, T., McCartney, E., Mattey, M., Naftalin, L., & Watson, J. (2005). Effects of computer-based intervention using acoustically modified speech (Fast ForWord-Language) in severe mixed receptive-expressive language impairment: outcomes from a randomized control trial. *Journal of Speech Language and Hearing Research, 48*(3), 715-729.

Conti-Ramsden, G., St Clair, M. C., Pickles, A., & Durkin, K. (2012). Developmental Trajectories of Verbal and Nonverbal Skills in Individuals With a History of Specific Language Impairment:

From Childhood to Adolescence. *Journal of Speech Language and Hearing Research, 55*(6), 1716-1735. Retrieved from WOS:000314531600010

Culatta, B., & Horn, D. (1982). A Program for Achieving Generalization of Grammatical Rules to Spontaneous Discourse. *Journal of Speech and Hearing Disorders, 47*(2), 174-180. Retrieved from <Go to ISI>://A1982PC03100011

Ebbels, S. H., Maric, N., Murphy, A., & Turner, G. (2014). Improving comprehension in adolescents with severe receptive language impairments: a randomised control trial of intervention for coordinating conjunctions. *International Journal of Language & Communication Disorders, 49*(1), 30-48.

Ebbels, S. H., Nicoll, H., Clark, B., Eachus, B., Gallagher, A. L., Horniman, K., Jennings, M., McEvoy, K., Nimmo, L., & Turner, G. (2012). Effectiveness of semantic therapy for word-finding difficulties in pupils with persistent language impairments: a randomized control trial. *International Journal of Language & Communication Disorders, 47*(1), 35-51. doi:10.1111/j.1460-6984.2011.00073.x

Ebbels, S. H., van der Lely, H. K. J., & Dockrell, J. E. (2007). Intervention for verb argument structure in children with persistent SLI: a randomized control trial. *Journal of Speech Language and Hearing Research, 50*, 1330-1349.

Ebbels, S. H., Wright, L., Brockbank, S., Godfrey, C., Harris, C., Leniston, H., Neary, K., Nicoll, H., Nicoll, L., Scott, J., & Marić, N. (in press). Effectiveness of 1:1 speech and language therapy for older children with (developmental) language disorder. *International Journal of Language & Communication Disorders*. doi:10.1111/1460-6984.12297

Falkus, G., Tilley, C., Thomas, C., Hockey, H., Kennedy, A., Arnold, T., Thorburn, B., Jones, K., Patel, B., Pimenta, C., Shah, R., Tweedie, F., O'Brien, F., Leahy, R., & Pring, T. (2016). Assessing the effectiveness of parent–child interaction therapy with language delayed children: A clinical investigation. *Child Language Teaching and Therapy, 32*(1), 7-17. doi:10.1177/0265659015574918

Fey, M. E., Cleave, P., Long, S. H., & Hughes, D. L. (1993). Two Approaches to the Facilitation of Grammar in Children with Language Impairment: An Experimental Evaluation. *Journal of Speech and Hearing Research, 36*, 141-157.

Fey, M. E., Cleave, P. L., & Long, S. H. (1997). Two models of grammar facilitation in children with language impairments: phase 2. *Journal of Speech Language and Hearing Research, 40*, 5-19.

Fey, M. E., Finestack, L. H., Gajewski, B. J., Popescu, M., & Lewine, J. D. (2010). A Preliminary Evaluation of Fast ForWord-Language as an Adjuvant Treatment in Language Intervention. *Journal of Speech Language and Hearing Research, 53*(2), 430-449. doi:doi:10.1044/1092-4388(2009/08-0225)

Fricke, S., Bowyer-Crane, C. A., Haley, A. J., Hulme, C., & Snowling, M. (2013). Efficacy of language intervention in the early years. *Journal of Child Psychology and Psychiatry, 54*(3), 280-290.

Gillam, R. B., Loeb, D. F., Hoffman, L. M., Bohman, T., Champlin, C. A., Thibodeau, L., Widen, J., Brandel, J., & Friel-Patti, S. (2008). The efficacy of Fast ForWord Language Intervention in school-age children with language impairment: A Randomized controlled trial. *Journal of Speech Language and Hearing Research, 51*(1), 97-119. doi:doi:10.1044/1092-4388(2008/007)

Greenhalgh, T. (1997). How to read a paper. Getting your bearings (deciding what the paper is about). *BMJ: British Medical Journal, 315*(7102), 243.

Isaacs, D., & Fitzgerald, D. (1999). Seven alternatives to evidence based medicine. *Bmj, 319*(7225), 1618. doi:10.1136/bmj.319.7225.1618

Kambanaros, M., Michaelides, M., & Grohmann, K. K. (2016). Cross-linguistic transfer effects after phonologically based cognate therapy in a case of multilingual specific language impairment (SLI). *International Journal of Language & Communication Disorders*, n/a-n/a. doi:10.1111/1460-6984.12270

Kulkarni, A., Pring, T., & Ebbels, S. (2014). Evaluating the effectiveness of Shape Coding therapy to develop the use of regular past tense morphemes in two children with language impairments. *Child Language Teaching and Therapy, 30*(3), 245-254.

Mecrow, C., Beckwith, J., & Klee, T. (2010). An exploratory trial of the effectiveness of an enhanced consultative approach to delivering speech and language intervention in schools. *International Journal of Language & Communication Disorders, 45*(3), 354-367. doi:DOI: 10.3109/13682820903040268

Motsch, H. J., & Riehemann, S. (2008). Effects of 'Context-Optimization' on the acquisition of grammatical case in children with specific language impairment: an experimental evaluation in the classroom. *International Journal of Language & Communication Disorders, 43*(6), 683-698. doi:DOI: 10.1080/13682820701794728

Mulac, A., & Tomlinson, C. N. (1977). Generalization of an operant remediation program for syntax with language delayed children. *Journal of Communication Disorders, 10*, 231-243.

Nail-Chiwetalu, B., & Ratner, N. B. (2007). An assessment of the information-seeking abilities and needs of practicing speech-language pathologists. *Journal of the Medical Library Association, 95*(2), 182.

O'Donnell, M., & Bunker, J. (1997). A sceptic's medical dictionary. *BMJ-British Medical Journal-International Edition, 315*(7119), 1387-1387.

Parsons, S., Law, J., & Gascoigne, M. (2005). Teaching receptive vocabulary to children with specific language impairment: a curriculum-based approach. *Child Language Teaching and Therapy, 21*(1), 39-59.

Petersen, D. B., Gillam, S. L., Spencer, T., & Gillam, R. B. (2010). The Effects of Literate Narrative Intervention on Children With Neurologically Based Language Impairments: An Early Stage Study. *Journal of Speech Language and Hearing Research, 53*(4), 961-981. Retrieved from WOS:000280598100014

Petersen, D. R., Gillam, S. L., & Gillam, R. R. (2008). Emerging procedures in narrative assessment - The index of narrative complexity. *Topics in Language Disorders, 28*(2), 115-130. Retrieved from WOS:000256043300005

Pring, T. (2005). *Research Methods in Communication Disorders*. London: Whurr Publishing for Professionals.

Rice, M. L., & Hoffman, L. (2015). Predicting vocabulary growth in children with and without specific language impairment: a longitudinal study from 2;6 to 21 years of age. *Journal of speech, language, and hearing research : JSLHR, 58*(2), 345-359. Retrieved from MEDLINE:25611623

Riches, N. (2013). Treating the passive in children with specific language impairment: A usage-based approach. *Child Language Teaching and Therapy, 29*(2), 155-169.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *Bmj, 312*(7023), 71-72.

Smith-Lock, K. M., Leitao, S., Lambert, L., & Nickels, L. (2013). Effective intervention for expressive grammar in children with specific language impairment. *International Journal of Language & Communication Disorders, 48*(3), 265-282. Retrieved from WOS:000318572000003

Snowling, M. J., & Hulme, C. (2011). Evidence-based interventions for reading and language difficulties: Creating a virtuous circle. *British Journal of Educational Psychology, 81*(1), 1-23.

Starling, J., Munro, N., Togher, L., & Arciuli, J. (2012). Training secondary school teachers in instructional language modification techniques to support adolescents with language impairment: a randomized controlled trial. *Language, Speech, and Hearing Services in Schools, 43*(4), 474-495. Retrieved from MEDLINE:22826368

Thomas, K. B. (1978). The consultation and the therapeutic illusion. *Bmj, 1*(6123), 1327-1328. doi:10.1136/bmj.1.6123.1327

Tomblin, J. B., Zhang, X., Buckwalter, P., & O'Brien, M. (2003). The stability of primary language disorder: Four years after kindergarten diagnosis. *Journal of Speech Language and Hearing Research, 46*(6), 1283-1296.

Wilson, J., Aldersley, A., Dobson, C., Edgar, S., Harding, C., Luckins, J., Wiseman, F., & Pring, T. (2015). The effectiveness of semantic therapy for the word finding difficulties of children with severe and complex speech, language and communication needs. *Child Language Teaching and Therapy, 31*, 7-17.

Zhang, X., & Tomblin, J. B. (2003). Explaining and controlling regression to the mean in longitudinal research designs. *Journal of Speech, Language, and Hearing Research, 46*(6), 1340-1351.

Zipoli, R. P., & Kennedy, M. (2005). Evidence-Based Practice Among Speech-Language PathologistsAttitudes, Utilization, and Barriers. *American Journal of Speech-Language Pathology, 14*(3), 208-220.

Zwitserlood, R., Wijnen, F., van Weerdenburg, M., & Verhoeven, L. (2015). 'MetaTaal': enhancing complex syntax in children with specific language impairment-a metalinguistic and multimodal approach. *International Journal of Language & Communication Disorders, 50*(3), 273-297. Retrieved from WOS:000353418200001

Figure 1 – contributions of experimental control and numbers of participants to study robustness
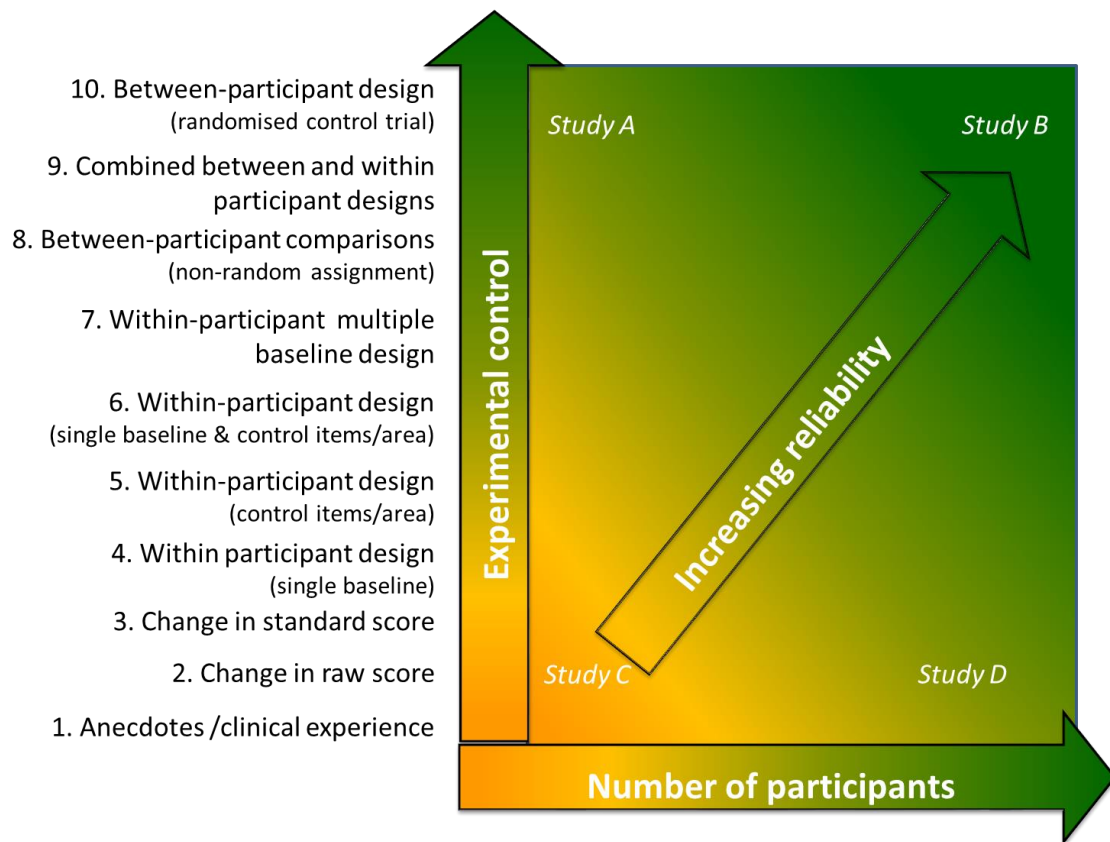

10. Between-participant design
(randomised control trial)
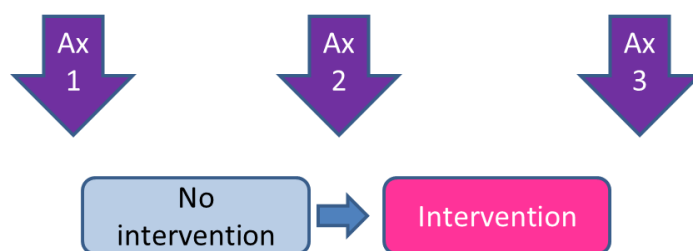
9. Combined between and within
participant designs

8. Between-participant comparisons
(non-random assignment)

7. Within-participant multiple
baseline design

6. Within-participant design
(single baseline & control items/area)

5. Within-participant design
(control items/area)

4. Within participant design
(single baseline)

3. Change in standard score

2. Change in raw score

1. Anecdotes /clinical experience

Experimental control

Increasing reliability

Number of participants

Study A

Study B

Study C

Study D

Figure 2 – within participants single baseline design

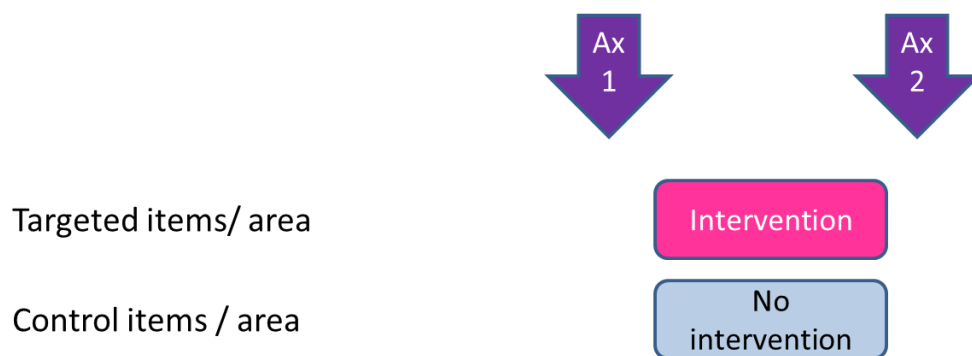Figure 3 – within participants design with control items / area

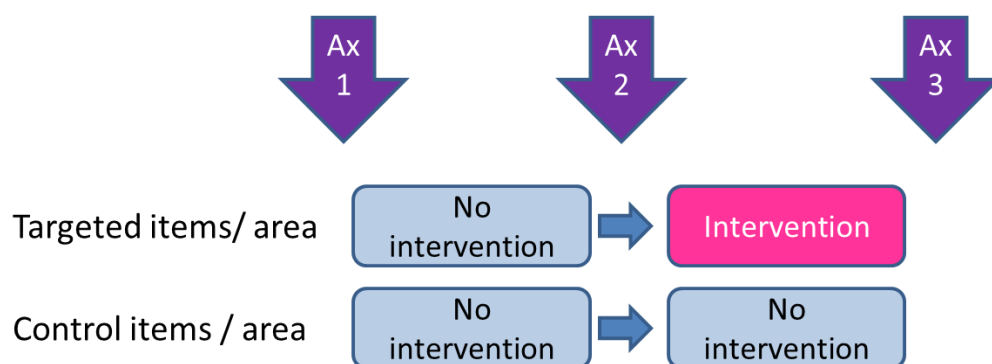Figure 4 – within participants design with single baseline and control items/area

Figure 5 – within participants multiple baseline design across targets
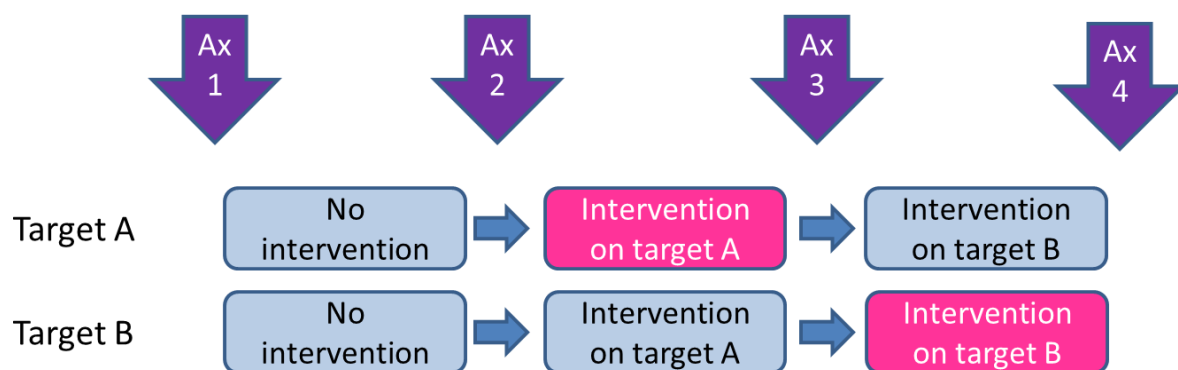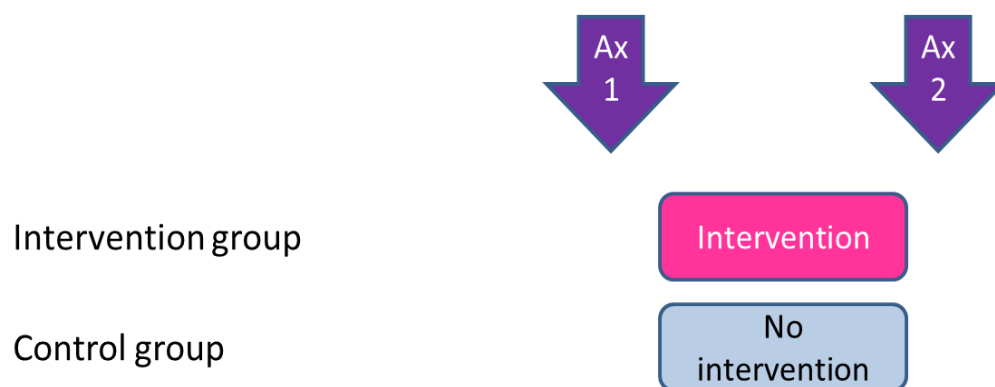

Figure 6 – Between participant comparisons

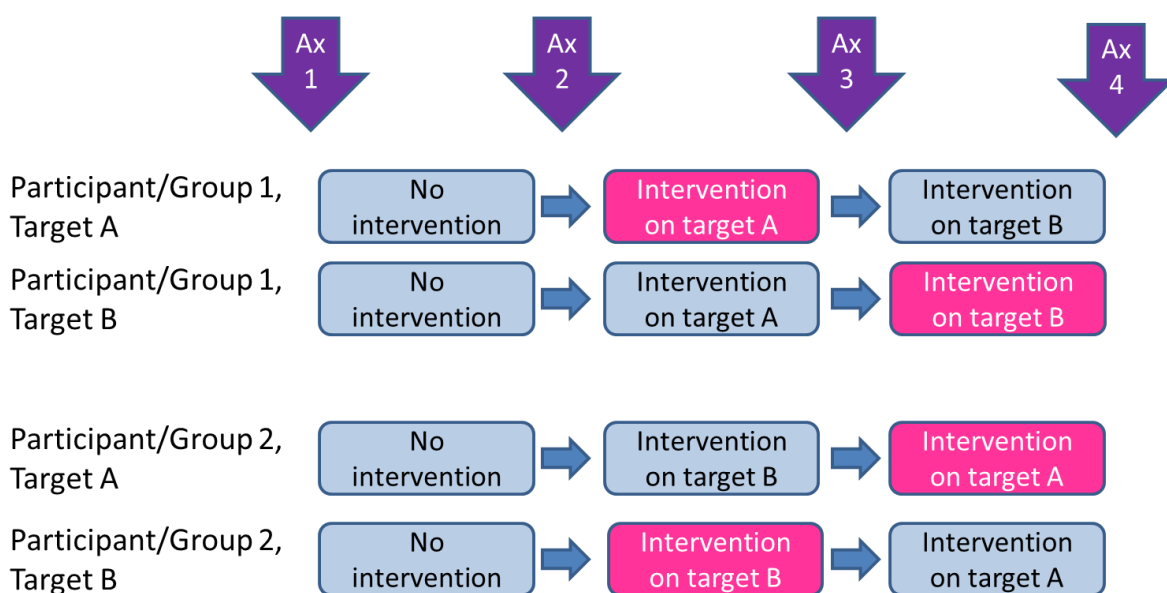Figure 7 – between and within participants multiple baseline design

Figure 8 – randomised control trial with waiting controls, plus control and generalisation tasks as used in Ebbels et al. (2014)