The effect of SMART Relational Skills training on Intelligence Quotients:

Controlling for Individual Differences in Attentional Skills and Baseline IQ

Bryan Roche1, Jamie Cummins2, Sarah Cassidy1, Anna Dillon3, Lauren Moore1 & Ian Grey4

1 Maynooth University, Co. Kildare. Ireland

2 Ghent University, Ghent. Belgium

3. Emirates College for Advanced Education, Abu Dhabi. UAE

4 United Arab Emirates University, Al Ain. UAE

Corresponding Author: Ian Grey, Department of Cognitive Science, United Arab Emirates University, Al Ain. UAE

Author Notes: Funding for the study was provided by Zayed University, United Arab Emirates. The first author declares a conflict of interest in being the director of a commercial entity that sells the online relational skills training software assessed in the current study. All data and analysis scripts for the study is available on the Open Science Framework (<https://osf.io/hkar5/>).

**Abstract**

The current study served to replicate the finding that extensive derived relational responding fluency training using the SMART (strengthening mental abilities with relational training) method can enhance intelligence quotient scores, while controlling for baseline levels of intelligence and attentional skills. Two separate groups of children based on school enrolment, aged 10-11 years, were assigned to a control or treatment condition *en masse* (i.e., not individual random assignment). All participants were administered the *Test of Everyday Attention* at baseline and completed the *Wechsler Abbreviated Scale of Intelligence* (WASI-II) at baseline and follow-up. Supervised training in derived relational responding fluency was administered at regular intervals over several months for the intervention group only. Results showed that there was a significant increase in IQ scores from baseline to follow-up for the intervention group only. However, controlling for baseline intelligence scores or any of three separate sub-indices of attentional skills eliminated this effect. This outcome serves as an indicator of the importance of controlling for baseline factors in intervention studies and the random allocation of participants to treatment and control groups.

**Highlights**

* We tested the effects on IQ quotients of a *SMART* relational skills intervention using two groups of children aged 10-11 years.
* We controlled for baseline levels of intelligence and attentional skills.
* The SMART effect was replicated, but not after controlling for baseline intelligence or attentional skills.

**Keywords**: Strengthening mental abilities with relational training; Relational Frame Theory, Cognitive training, Derived relational responding, Intelligence.

**The effect of SMART Relational Skills training on Intelligence Quotients:**

**Controlling for Individual Differences in Attentional Skills and Baseline IQ**

The study of intelligence has always been slightly contentious within the field of behavior analysis. However, there is in fact no conceptual or empirical impediment to employing the concept, so long as it is understood in technical terms and without resorting to mentalism (Schlinger, 2003). Indeed, as one of the most well-established concepts in psychology, it may be remiss for behavior analysts not to develop a conceptual and methodological framework for dealing with this concept. In particular, if conceptual ambiguities can be ironed out, behavioral researchers can benefit from the use of tests of intelligence as convenient proxy measures for the fluency of scholastically and professionally important skill sets (e.g., reading, vocabulary, problem-solving, etc.; see Cassidy et al., 2010; McLoughlin, Tyndall, et al, 2020).

The current paper reports on a study that emerged from a Relational Frame Theory (RFT; Hayes, Barnes-Holmes & Roche, 2001) research program that has attempted to disarm and embrace the potentially problematic concept of intelligence. This has been partly achieved through conceptual analysis aimed at reframing the concept of intelligence in behavior-analytic terms (e.g., Cassidy et al., 2010; Colbert et al., 2019; McLoughlin, Tyndall et al, 2020; Moran et al., 2015). Intelligence indices have also been embraced as a potentially important outcome measure of the effects of educational interventions designed to enhance the fluency of underlying skills sets. Employing well-established gold-standard measures of intelligence in our field, therefore, is part of a pragmatic effort to use measures that speak to a wider audience, while at the same time serving some practical purposes in research and within clinical settings.

**Relational Frame Theory**

The ultimate aim of the current research is to develop a brief intervention that might enhance general intelligence, and consequently the specific aptitudes with which the construct is associated. In particular, the program draws upon a core idea from within the RFT literature, which is that most specific aptitudes assessed by gold-standard intelligence tests appear to rely upon fluencies across a relatively small set of easily trained “relational” skills (see Cassidy et al., 2010; Colbert et al., 2019; McLoughlin, Tyndall et al., 2020). This possibility presents the exciting opportunity to generate interventions that can enhance general intelligence with potentially life-changing consequences (see Cassidy, et al., 2011; see also Dixon et al., 2019).

Relational skills are those that involve responding to one stimulus in terms of another as specified by a contextual cue (Hayes et al., 2009). For example, a child may be presented with a toy and then asked to choose a “larger” one from an array of further toys. In this case, the word “larger “functions as the contextual cue which controls the responding to stimuli in the choice array in terms of their *size* relationship with the presented sample.

This type of relational responding can occur in different patterns (e.g., in terms of coordination, opposition, temporality, etc), which are referred to as “frames”. A frame should be understood as a pattern of relational responding to one stimulus in terms of another. For example, a child can be taught to respond to an elephant as larger than a mouse, and also to an elephant as a member of a hierarchical category called “animals”. Thus, a myriad of relational responses are possible for any one stimulus, depending on the contextual cues present. Perhaps most importantly, relational responses may be derived and arbitrarily applicable. That is, verbally-able individuals are almost universally capable of responding to stimuli in terms of their distant abstracted relationship with other stimuli. As a concrete example, a child taught to choose a particular cuddly toy from an array given any one of two words, such as “teddy bear” and “cuddly toy”, may later discriminate that these names are equivalent (coordinate), despite never having been told so explicitly.

RFT has from the outset served the purpose of providing an account for how derived relational responding emerges as a repertoire in the first instance. In particular, it has identified the process of multiple exemplar training (MET). Put simply, this involves caregivers and educators directly reinforcing appropriate relational responses in the context of specific types of relational questions, but doing so using a wide variety of stimuli across a wide variety of contexts. Algebra, for example, requires an extended history of such training sufficient to lead to the emergence of entirely arbitrarily applicable relational responding. When responding correctly to algebraic formulae, the formal properties of the stimuli involved are unimportant, but the subtlety of multiple and simultaneous forms of contextual control over responding is paramount (see Marr, 2015; see also Hayes et al., 2001 for the first book length treatment of RFT).

 Relational responding takes many forms, but the most commonly studied are coordination (Steele & Hayes, 1991), distinction (Roche & Barnes, 1997), opposition (Barnes-Holmes, Barnes-Holmes, & Smeets, 2004; Steele & Hayes, 1991), comparison (Dymond & Barnes, 1995), temporality (O’Hora et al., 2008), analogy (i.e., relations between relations; Stewart, et al., 2004), hierarchy (Slattery & Stewart, 2014) and deixis (i.e., perspective taking; McHugh, et al., 2004).

RFT takes the position that intellectual skills such as problem solving, reading, arithmetic and other scholastic competencies, emerge from the mastery of relevant lower-level relational responding repertoires (Cassidy, et al., 2010; Colbert, et al., 2019; McLoughlin, Tyndall, et al., 2020). For instance, the emergence of an extensive vocabulary relies upon a previously established ability to derive frames of coordination between words in the vernacular, and as such, language ability has long been viewed as reliant upon more fundamental forms of multiple exemplar training occurring naturalistically in educational and social contexts. More specifically, the normal emergence of speech in infants is dependent upon fluency in derived relational responding with equivalence (Barnes, McCullagh, & Keenan, 1990; Dugdale & Lowe, 2000), as is the emergence of reading skills (de Rose, de Souza, Rossito, & de Rose, 1992; Farrington-Flint & Wood, 2007; Mackay, 1985; Sidman, 1971), a sufficient vocabulary (Edwards, Figueras, Mellanby, & Langdon, 2011; Nippold & Sullivan, 1987), grammar (Hock, 2003) and spelling (e.g., Mackay, 1985). Not surprisingly, therefore, RFT researchers have focused in on the possibility that improving the fluency of relational skills should feed upwards to more rapid acquisition of intellectual skills dependent on those forms of relational framing (McLoughlin, Tyndall et al., 2020).

**Promoting Emergence of Advanced Knowledge (PEAK)**

Several studies (e.g., Colbert et al., 2017; 2019; Dixon, et al., 2014; Moran, et al., 2015) have now reported robust correlations between various measures of relational responding competency and objective measures of intelligence. More importantly, interventions designed to enhance relational responding fluency across a range of domains appear to result in increases in standardized intelligence (e.g., Cassidy et al., 2011, 2016, Colbert et al., 2018, McLoughlin et al., 2020, 2022). One such training system, known as the PEAK (Promoting Emergence of Advanced Knowledge; Dunkel-Jackson & Dixon, 2018), combines a Skinnerian approach to establishing verbal behavior with training designed to establish fluency in generalized equivalence responding. This method has been shown to improve intelligence scores on standardized measures of intelligence among Autistic children above and beyond what is achieved using an applied behavior analysis approach alone (Dixon et al., 2019). Furthermore, PEAK’s own verbal behavior assessment scores are reported to correlate well with standardized measures of intelligence (Dixon, et al., 2018).

**Strengthening Mental Abilities with Relational Training (SMART)**

Another system of training, of particular interest in the current study, is known as (SMART) Strengthening Mental Abilities with Relational Training; Cassidy, et al., 2011). This training protocol enhances the fluency of relational responding in domains other than equivalence. Specifically, SMART involves establishing generalized fluency in relational responding in accordance with patterns of coordination, opposition, and comparison (more than and less than) across 55 stages of training and criterion-based testing. Training tasks consist of the presentation of one or more relational premises, using nonsense words as relata (e.g., Paf is the Same as Vek, Cug is opposite to Ler), followed by a relational question and the requirement to make a relational evaluation response (i.e., by choosing either the word YES or NO on the computer screen). The relational question may require a reflexive relational response (e.g., Is Paf the same as Vek?), a relational response based on mutual entailment (e.g., Is Ler opposite to Cug?), or based on combinatorial entailment (e.g., Is Paf Opposite to Ler?). Each stage of training presents tasks within strictly defined complexity parameters (see Cassidy et al., 2016), and continues until the participant satisfies a strict fluency criterion. Each training stage is followed by a testing stage involving tasks at the same level of complexity (but involving novel stimuli which also differ on each trial). Complexity is varied across tasks within the parameters defined for the relevant stage. This high level of stimulus novelty and a potentially infinite number of task types ensures an extremely high standard of multiple exemplar training within and across stages.

In one early study, Cassidy et al. (2011; Experiment 2) administered a precursor to the SMART procedure to a small sample of 10-12 year old school children with learning difficulties, over several months of regular training. The intervention involved training competences in same, opposite, more than and less than relational framing across multiple exemplars (randomly selected nonsense words) and establishing increasingly more fluent responding on successive novel relational problem-solving tasks and stimulus sets over the training period. These authors reported average IQ gains of 13.1 points (*SE*=2.5) for the eight participants.

In a further study employing a more rigorous and extended training protocol, now known as SMART, Cassidy et al. (2016; Experiment 1) delivered an online digitized form of SMART training to 15 typically developing 11-12 year old school children, but without a control group. Large IQ gains were observed for all participants, with an average IQ gain of 23 points recorded for the group. In another study employing the same intervention, Hayes & Stewart (2016), employed standardized measures of memory, literacy, and numeracy as pre and post-intervention measures, and found significant gains in all domains with a sample of 28 11-12 year old children in a randomized controlled trial, with an active control group.

The foregoing general effect has been replicated several times, across several different laboratories (see May et al., 2022 for a systematic review). For example, Colbert et al., (2018) reported on the effect of a SMART intervention that employed the Wechsler Abbreviated Scale of Intelligence (WASI-II) as the pre and post intervention measure in a study employing an IQ matched waiting control group (n=26). These researchers reported significant gains on all three IQ indices; Full-Scale IQ, Verbal IQ and Performance IQ, as well as on each of the WASI-II subtests (Vocabulary, Similarities, Block Design & Matrix Reasoning), compared to an IQ-matched waiting control group. Full scale IQ gains for the control group averaged at less than 1 point, while gains for the experimental group averaged over 18 points (i.e., > 1 SD).

More recently, in an active controlled and stratified trial of the SMART method, employing 6-10 year old children, McLoughlin et al. (2020) found a significant gain in non-verbal IQ (using the Kaufman’s Brief Intelligence Test; Kaufman & Kaufman, 2004) of nearly 9 points for a SMART treatment group, and a drop of over four points for an active chess-training control group, across the intervention period. Importantly, these researchers found that reading ability was predicted by post-intervention Non-verbal IQ (NVIQ), even when baseline NVIQ was controlled for. This effect maintained at one-month follow-up for the treatment group only (see also Thirus et al., 2016 for a further intervention study).

Only one study to date has focused explicitly on the enhancement of scholastic competencies using SMART. In that study (Cassidy et al., 2016; Experiment 2), the SMART intervention was administered to a group of 30 school children (15-17 years) and a widely used standardized scholastic aptitude test (the Differential Aptitudes Test; Bennet et al., 1990) was employed as the pre and post-intervention measure. Significant, but variable gains in overall scholastic ability were reported across the period of the intervention, sufficient to reflect significant educational advancements for the cohort.

**Boundary conditions for SMART**

 A small number of studies have attempted to identify extraneous variables or boundary conditions for the “SMART effect”. This development is important, particularly to contextual behavioral scientists, because an interest in factors that might modulate the effectiveness of a particular intervention is required, given the pragmatic truth criterion. According to Vilardaga et al. (2009), “…determining the ability of a theoretical analysis and subsequent intervention to affect behavior in the contexts in which it would actually be applied is essential.” (p. 124). In effect, the SMART intervention must not merely be shown to be effective under ideal laboratory conditions. Research must eventually begin to identify real-world conditions and difficult-to-control factors they may influence its practical application and effectiveness.

One study that aimed to do just this (Amd & Roche, 2018) tested the robustness of the “SMART effect” in an unstructured educational environment, with poor facilities and educational supports and under less than ideal administration conditions. Specifically, this involved administering SMART training to 35 socially disadvantaged children attending a charitable school in Bangladesh. Adherence to training was low given that attendance at the drop-in school was irregular amongst the largely homeless student population and given that a civil war was ongoing at the time of the study. Explosions were routinely heard on the streets while classes were in progress. Raven’s Matrices (Raven, et al., 1998) was used as the nonverbal measure of fluid intelligence pre and post-intervention. Data analysis focused on the relationship between gains in IQ scores and the amount of training undertaken by the various children in the study, given their various individual personal circumstances. A clear relationship was found between IQ gain and stages completed from the 55 stages provided by the SMART training. More importantly, however, SMART was still acceptably effective even when administered under extremely challenging conditions.

Another study (McLoughlin et al., 2020) examined the effect of baseline intelligence as an attribute factor that might modulate or even compromise the outcomes of SMART interventions. That study found a strong negative relationship between baseline NVIQ and NVIQ gain following intervention, with baseline NVIQ accounting for over 30% of the variance in NVIQ gain across participants. Thus, baseline IQ level may represent a potential confound for SMART effects, with the implication being that amongst those highest in IQ at baseline, the intervention may be ineffective, although this has never been tested.

 Trait personality factors may also intersect with the effects of relational skills interventions, insofar as they may be related to adherence to training, amongst other possible factors. This is important because, previous research (Amd & Roche, 2018; McLoughlin et al., 2022) has indicated that the amount of SMART training undertaken predicts the magnitude of IQ gains observed. McLoughlin, et al. (2022) measured the personality dimensions of the Big Five Aspects Scale (DeYoung, et al., 2007) for all participants. They found a moderate negative relationship between volatility (i.e., one aspect of the Neuroticism trait) and the number of stages of SMART training completed. In addition, Politeness (i.e., an aspect of Agreeableness), was found to correlate strongly with number of training stages completed. Interestingly, however, no other personality traits were found to predict training completion. In effect, personality variables were not significant explanatory concepts for training progress, and so are unlikely to represent boundary conditions for SMART effects.

 Further research is needed to establish additional real-world boundary conditions for the effectiveness of SMART and to further examine potential modulating factors identified so far. That is, we need to know that the effectiveness of SMART will be maintained when using heterogeneous samples of participants with a range of various skill attributes and educational histories. In the interest of expediency, research should focus on measuring the modulating effects of variables that most obviously present themselves on theoretical grounds, and on those most amenable to experimental analysis. One such variable of interest in the current study is level of attentional skill.

**Intelligence and Attentional Skills**

 The concept of attention is difficult to study because it is subject to the same type of definitional ambiguity as is the concept of intelligence. Indeed, as for intelligence, it is likely not useful to consider attention as a singular construct, and it may well prove impossible to do so with an acceptable level of theoretical coherence due to the wide range of processes that it appears to describe (see Hommell et al., 2019). Whatever it is that the concept refers to, however, it is widely agreed that it is a crucial skill set that has been consistently shown over the past decades to be predictive of academic achievement (e.g., Fergusson, et al., 1993; Fergusson, et al., 1997). Attention is also an important predictor of engagement in educational settings and school completion rates (Galera, et al., 2009). However, the relationship between attentional skills and scholastic achievement is a complex one often mediated by other factors, such as teacher-rated classroom performance and parent ratings of homework completion rates and medication use (see Langberg, et al., 2011). Variables such as conduct problems in the classroom also function to mediate the negative effects of poor attentional skills on scholastic attainment. However, attentional skill levels reliably predict academic attainment in their own right (Galera et al., 2009; see also Daley & Birchwood, 2010). Indeed, poor attention may be a better predictor of compromised educational attainment than either diagnosed hyperactivity or impulsivity (Fergusson & Horwood, 1995).

 With regard to measures of intelligence, IQ scores are also predicted by measures of the scope of one's attention combined with measures of attentional control (i.e., focus). Together these account for about one third of the variance in intelligence across individuals (Cowan et al., 2006). This might not be surprising when one considers that even among some of the earliest and most influential intelligence theorists (Spearman 1927; Thurstone, 1938), it had been assumed that foundational attentional skills underlie and load directly on to general intelligence (*g*). Later on, Broadbent (1958) approached attention as a filtering mechanism that allows the individual to maximize their information processing resources and so was a central process in determining one's overall intellectual ability. Carroll’s (1970) model also treated attentional skills as an important prerequisite of performance on intelligence tests.

 Knudsen (2007) provided a component analysis of attention in terms of (a) working memory, (b) competitive selection, (c) top-down sensitivity control, and (d) filtering for stimulus salience. This analysis is relevant here because working memory, in particular, is a crucial component of full-scale intelligence. Indices for working memory are produced by almost all modern IQ tests. Thus, the intersection between the concept of attention and working memory suggests a functional overlap between them. Indeed, one study by Schweizerand Moosbrugger(2004) found both attention and working memory are excellent predictors of intelligence, with attention being more reliable as a predictor, across two different IQ test variants.

 Contemporary definitions of attentional skill identify attention as a multi-dimensional construct involving distinct abilities such as selective, sustained, alternating and divided attention (Gazzaniga & Halpern, 2015). Selective attention refers to the ability to select and focus on particular stimuli while simultaneously suppressing irrelevant or competing distractors while sustained attention relates the maintenance of response persistence and continuous effort over extended periods of time (Ko et al., 2017). Whilst the majority of research in this domain has been concerned with the effects of selective attention, which is conceptually related to inhibition, attentional control appears to correlate well with intelligence measures and this relationship extends to children who, typically have much lower attention spans than adults (Cowan et al., 2006).

**The Current Study**

The current study served two functions: a) to conceptually replicate previous findings regarding the effect on full scale IQ from pre to post SMART interventions with children and; b) to examine whether or not the effect persists when statistically controlling for attentional skills and baseline intelligence levels. Given the emerging complexity and the understanding of the concept of attentional skills, a measure of attentional skills was employed that indexes two separate dimensions of attention (sustained and selective) and combines these to produce a third composite attentional ability score. Each of these will be statistically controlled for separately in post-hoc analysis of the effect of SMART training on the full-scale intelligence scores of a group of 10 – 11 year old children. The study adopted a partially randomized quasi-experimental approach. A wait procedure was employed for the control group who received educational instruction as normal during the intervention period.

The current research is exploratory and inductive in nature. We have not developed or adopted a theoretical position on why precisely attentional skills or baseline intelligence levels should intersect with the effects of our specific relational skills training interventions and no such model or theory is being investigated. The research at this point is pragmatic merely in identifying and quantifying the magnitude of modulating factors that may alter the impact of SMART interventions in applied settings.

**Method**

**Participants**

All students were recruited from two large private international fee paying English speaking schools in Dubai and Abu Dhabi for reasons of convenience and their being amenable to the study. The schools both deliver programmes for learning that are rooted in the National Curriculum for England (NCfE). The schools are characterised by a diversity of cultures, and children are drawn from over 80 countries. Due to the diversity in the ethnicities of the participants, these were not recorded individually.

Information sessions regarding the nature and content of the proposed study were held with the Board of Management of both schools following which both schools agreed to participate. Two classes comprised of children aged 10-11 years were recruited from each school (min. 10 years 6 months, max 11 years 6 months). This age is broadly typical of the ages of participants in SMART studied (e.g., Cassidy et al. 2011; 2016; Hayes & Stewart, 2016; McLoughlin et al., 2022). The different school cohorts were randomly assigned to their roles as experimental and control group by the flip of a coin. Participants could not be matched on baseline assessments across conditions for this reason. Thus, the current procedure does not represent a randomized controlled method but adopts a quasi-experimental design.

 Forty-five participants were recruited in total (21experimentals and 24 controls). Thirteen participants were excluded from analysis (6 from the experimental group, and 7 from the control group) owing to repeated absences from school severely affecting training progress or follow-up testing. No student was omitted based on poor progress alone that was not entirely explained by school absence. The final cohort of participants consisted of 32 children, 15 in the intervention group (7 male, 8 female) and 17 in the control group (7 male and 10 female). The average age for the experimental group was 10.73 years (SD: 0.46) and the average age for the control group was 11.06 years (SD: 0.24). The average overall age for the study sample was 10.91 years (SD: 0.39). No child had been identified as having a learning disability, either specific or general.

## Materials

**Wechsler Abbreviated Scale of Intelligence-2nd Edition**

The Wechsler Abbreviated Scale of Intelligence, 2nd Edition (WASI-II, Wechsler, 2011) was employed as the baseline and post-intervention measure of intelligence. Each child was assessed individually in a quiet room in line with usual practice. The WASI-II provides an approximation of intellectual ability relative to one’s peers in a relatively brief administration time of around 30 minutes. All four IQ subtests (Vocabulary, Similarities, Block Design & Matrix Reasoning) were administered, which allowed for the calculation of a Full Scale (FSIQ) score. Split-half reliability for the subtest scores derived from a population of children aged 6-16 years has been reported as ranging from good (.87) to excellent (.91), while the average reliability coefficients for the VCI, PRI, FSIQ-4, and FSIQ-2 composites are reported as excellent at .94, .92, .96, and .93, respectively (McCrimmon & Smith, 2013). Test-retest stability coefficients over an 88-day period for the same age group range are reported as excellent (.90) for the subtests and good (.87) to excellent (.95) for the composites. With respect to validity for use with children, correlations between the WASI-II and the WISC-IV, are reported as ranging between acceptable (0.81) and excellent (0.92; McCrimmon & Smith, 2013).

**Test of Everyday Attention (TEA-Ch 2)**

The Test of Everyday Attention (second edition; Manly et al., 2016) is a gold standard measure of attention designed for children aged 5 to 15 years. There are two versions of the TEA-Ch 2. Children younger than 8 years old complete the TEA-Ch2 J. Children aged 8 and above complete the TEA-Ch2 A version. In the current study, all children were administered the TEA-Ch 2 A. For children aged 8-15 years the test requires approximately 45 minutes for administration. In the current study the test was administered digitally, with each child being assessed individually in a quiet room under the usual assessment conditions. The software employed automatically generated scaled and indexed scores based on general population norms, stratified by age and sex, and provided indices of selective and sustained attention. When these scores are combined, they provide an overall measure of everyday attention. Individual tasks raw scores are converted to scaled scores (*M*=10, *SD*=3).TEA-Ch2 A for children aged 8 years to 15 years. For the TEA-Ch2 A, internal consistency coefficients ranged from moderate (>.5) to excellent (>.9) with most subtests being good or excellent. Structural Equation Modelling (SEM) of the TEA-Ch2 data has confirmed that the individual subtests contribute to unique variance with two common factors supporting the construct validity of the Selective Attention and Sustained Attention indexes (Manly et al., 2016).

**Relational skills training intervention**

The current study employed online digitized SMART training as the relational skills training intervention. SMART is a commercial tool available at raiseyouriq.com. SMART trains a series of increasingly complex relational problem-solving strategies across a potentially infinite number of training trials (i.e., to criterion) delivered across up to 70 training and testing stages. This approach is characterized as a multiple exemplar approach to training in which a potentially infinite number of similar but not identical tasks, differing within well-defined complexity parameters within each stage of training, and each employing unique nonsense word stimuli as relata, are delivered in sequence until the user produces 16 correct responses sequentially. Each task in a training block involves the presentation of verbal feedback on screen following a response. Mastery of the training criterion results in the presentation of a finite 16-trial test involving novel task types defined within the same complexity parameters but also involving completely novel nonsense words as relata.

Tasks involve the presentation of one or more logical premises followed by a question. For example, a participant may be presented with the premises “DUZ is opposite to JEL. “JEL is the same as LEF”. Is DUZ the same as LEF?”. Two response buttons appear underneath the relational question, a *Yes* button and a *No* button, the left/ right positions of which are randomized across trials (see Figure 1 for a sample task). The nonsense words are always pronounceable in the English vernacular and no word is used twice across any training or testing stage.



**Figure 1.** Sample SMART training trial. The task response window timer is in the top right of the screen. A progress bar is in the top middle of the screen. The egg shape represents one of the gamified features

 Participants proceed through levels of training by mastering both the preceding training block and the subsequent test. If even a single error is made on the finite novel test of mastery following a particular stage, the participant is recycled back to training to criterion once more, followed again by a test involving different but similar tasks and completely unique stimuli to ensure generalization of the skills established during the preceding training block. There is a 30s limit on responding on every trial, after which the response is recorded as incorrect and verbal onscreen feedback is delivered to that effect (along with a warning regarding slow responding).

The first 29 stages of training established fluency and relational skills of same and opposite relational responding simultaneously, whereas the next 26 stages establish fluency in comparative (more than and less than) relational responding. The more complex stages within each of these two relational training “modules” involved a maximum of three premises followed by a relational question about any two relata in the premise array. The final 15 advanced stages consisted of 9 stages of further same/opposite relational skills training and 6 further comparative relational skills training. These tasks involve 4 premises followed by relational question regarding any two of the relata in the premise array (see Cassidy et al., 2016, for a full outline of the procedure).

In previously published studies, participants have been advised to stop their training at stage 55 as is recommended on the training website. However, due to rapid progress in the training by several participants, they were given permission by teachers to continue with the training to stage 70 rather than to be eliminated from the study at stage 55.

**Ethical Approval**

Ethical approval was provided by the Institutional Review Board of the first author’s educational institution (Ref: 14045F). School approval was provided by the Board of Management of each the two participating schools and informed consent was obtained for all participants.

**General Procedure**

All baseline and post-assessments were conducted in the school setting. A testing room was provided by each school for this purpose which met the requirements for appropriate testing conditions (i.e., minimal noise, free from disruption and distraction). Two external qualified psychologists with post-graduate training in psychological assessment administered both the WASI-II and the Teach-2 to all participants following competency checks for both measures. To reduce the threat of experimenter bias, one psychologist was responsible for pre-assessments only and the second was only responsible for all post-assessments only. For both psychologists timepoint and the allocation of particular children to conditions was masked. All training sessions (2 per week for 45 mins approx.) took place in the school setting and in a group setting in the participants’ usual classrooms. Only the experimental group were exposed to SMART, while the control group continued with schooling as normal. The classroom teacher provided supervision for all sessions, provided encouragement and ensured that participants remained on task. However, progress through the training stages was entirely controlled by the SMART software, and the teacher in no way interfered with progress. No training took place outside of normal school hours. The study was discontinued after four months owing to the children in the study completing their semester requirements. A total of 18 weeks elapsed between baseline and post-study assessment. School vacation over this period comprised two weeks leaving a total of 16 weeks for intervention.

Selected for Inclusion

(n = 45)

# Enrollment

Excluded (n = 0)

Randomized (n = 45)

#

# Analysis

Analyzed (n = 17)

Excluded from analysis

(n =0)

Analyzed (n = 15)

Excluded from analysis

(n =0)

Lost to follow up: (n=7)

Withdrew from study (n=4)

Unavailable for Follow-up measures (n=3)

Lost to follow up: (n=6)

Withdrew from study (n=6)

Unavailable for Follow-up measures (n=0)

School 2 Allocated to **Control**

(n = 24)

Baseline WASI-II

Baseline Tea-Ch 2

No intervention.

School 1 Allocated to **SMART** (n = 21)

Baseline WASI-II

Baseline Tea-Ch 2

Intervention = 2 X 45 min (approx.) sessions per week.

Total of 20 hours approx.

**Follow up**

Allocation

**Figure 2**. Consort Diagram illustrating study design

**Results**

Seven participants in the intervention group completed all 70 training stages. Only one participant failed to complete the required minimum 55 core training stages but their data was included in the analysis because they had not terminated their participation in the study and remained engaged with the training throughout. Another participant was not available for follow-up testing, but again due to full engagement with the study and the completion of training their data was retained for analysis. For this participant, a multiple imputation method was employed in all analyses to deal with the missing follow-up IQ data point.

**Analytic approach**

All analyses in this study were conducting using R. Data processing workflows were conducted using the *tidyverse* package (Wickham et a., 2019); Cohen’s d effect sizes for t-tests were computed using the *effsize* package (Torchiano et al., 2020); ANOVAs and ANCOVAs were computed using the *car* package (Fox & Weisberg, 2020); estimated marginal means for ANOVAs and ANCOVAs were computed using the *emmeans* package (Lenth, 2022); and reliable change indices were computed using the *ClinicalSig* package (Ziegler, 2016). In addition to basic descriptive analyses, our confirmatory analytic approach consisted of two phases. As a first phase, we examined our data in four different ways to conceptually replicate previous findings relating to the effectiveness of SMART in increasing intellectual ability test scores. Firstly, we examined whether full-scale IQ scores differed as a function of the interaction between timepoint (baseline vs. follow-up) and intervention condition (SMART vs. control). To do this, we used a 2x2 mixed between-within ANOVA, with FSIQ score as the DV, timepoint as the within-subject IV, and intervention condition as the between-subjects IV. We specifically expected a significant interaction, such that we would find a greater IQ increase in the SMART condition compared to the control condition from baseline to follow-up. We then compared IQ scores at post-intervention across groups, while statistically controlling for baseline IQ (i.e., as a covariate) to account for both pre-existing differences in IQ at baseline and change effects due purely to regression to the mean (i.e., an ANCOVA; see Twisk et al., 2018). Thirdly, we calculated two metrics to quantify individual-level change scores: the “Persons as Effect Sizes” metric (*PAES;* Grice et al., 2020) and Reliable Change Indices (*RCI*; Jacobsen & Truax, 1991). The PAES metric involves computing the proportion of participants in each condition who exhibited a theoretically meaningful increase on the score of interest (in this case, FSIQ scores). Because specifying “theoretically meaningful” change in this case would be difficult, we instead provide a pseudo-specification curve which highlights the proportion of participants whose FSIQ difference from T1 to T2 exceeded various values (ranging from -15, indicating a very strong decrease, to +15, indicating a very strong increase). In general, we expected that the PAES scores would be superior for the SMART condition compared to the control condition. The RCI was computed as a means of determining which individual participants exhibited a statistically significant degree of change. Fourth and finally, we graphically examined the raw IQ scores between the two conditions for each participant which is an approach that has been recommended as an aid to more accurate interpretation of group changes in IQ scores (Burgaleta, et al., 2014).

The second phase of our confirmatory analytic approach involved investigating whether controlling for the three baseline attention scores (selective, sustained, and every day), with and without the addition of baseline IQ, would attenuate any observed improvements in IQ. To do this, we conducted two sets of three ANCOVAs. All ANCOVAs in the first set included FSIQ as the dependent variable, and timepoint and condition as independent variables. All ANCOVAs in the second set included follow-up IQ scores as a dependent variable, experimental condition as a independent variable, and baseline IQ as a covariate. Each ANCOVA in each set then also included one of the measures of attention as a covariate.

**Analyses**

The correlation between baseline IQ and the three attention measures are reported below. Interestingly, baseline intelligence scores correlated significantly only with selective attention scores and did not correlate with either sustained attention scores or the composite everyday attention score.

**Table 1.** Means, standard deviations, and correlations with confidence intervals

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | 1 | 2 | 3 |
|  |  |  |  |
| 1. Baseline IQ |   |   |   |
|   |   |   |   |
| 2. Everyday Attention | .19 |   |   |
|   | [-.17, .51] |   |   |
|   |   |   |   |
| 3. Selective Attention | .51\*\* | .76\*\* |   |
|   | [.19, .73] | [.55, .87] |   |
|   |   |   |   |
| 4. Sustained Attention | -.25 | .54\*\* | -.10 |
|   | [-.55, .11] | [.24, .75] | [-.43, .26] |
|   |   |   |   |

*Note.* Values in square brackets indicate the 95% confidence interval for each correlation.

\* indicates *p* < .05. \*\* indicates *p* < .01.

The average number of training trials required by participants was 3793 with a range from 1676 – 6818 (*95% CI* 2951-4635). Because participants across conditions were not matched for IQ at baseline but rather assigned to conditions based on which school they attended, the baseline IQs of participants in the SMART condition (*M* = 89.67, SD = 8.94) were significantly lower than those in the control condition (*M* = 108.23, SD = 9.43), *t(*30) = 5.694, *p* < .001, Cohen’s *d* = 2.01, *95% CI* [1.13, 2.91]. Selective attention scores were also higher in the control condition (*M* = 132.94, SD = 14.13) compared to the SMART condition (*M* = 102.40, SD = 14.01), *t*(30) = 6.13, *p* < .001, Cohen’s *d* = 2.17, *95% CI* [1.26, 3.08]. Everyday attention scores did not differ significantly across controls (*M* = 117.88, SD = 14.51) and SMART intervention participants (*M* = 108.00, SD = 14.52), t(30) = 1.92, *p* = .064, Cohen’s *d* = 0.68, *95% CI* [-0.06, 1.42]. Conversely, sustained attention scores were in fact higher in the SMART condition (*M* = 109.67, SD = 15.36) compared to the control condition (*M* = 98.00, SD = 15.11), *t*(30) = 2.16, *p* = .039, Cohen’s *d* = 0.77, *95% CI* [0.02, 1.52].



**Figure 3**. Mean IQ scores at baseline and follow-up for the SMART and control conditions. Error bars represent 95% confidence intervals around the estimated marginal means.

We investigated whether IQ scores changed from baseline to follow-up, and whether this change varied as a function of the intervention condition, using a mixed between-within ANOVA. Critically, we found the expected significant interaction between condition and timepoint, *F*(1, 60) = 4.23, *p* = .044, partial eta squared = 0.066; see Figure 3. There was also a main effect of experimental condition, *F*(1, 60) = 35.74, *p* < .001, partial eta squared = 0.405, and no main effect of timepoint, *F*(1, 60) = 0.81, *p* = .372, partial eta squared = 0.011. Post-hoc t-tests revealed that IQ scores increased significantly for participants in the SMART condition from baseline to follow-up (*M* change = 6.33, 95% CI [2.53, 10.13]), *t*(14) = 3.57, *p* = .003, Cohen’s d = 0.70, 95% CI [0.25, 1.16] but not for the control condition, for whom no significant change from baseline to follow-up (*M* change = -2.7, 95% CI [-6.19, 0.78]) was found, *t*(16) = 1.65, *p* = .119, Cohen’s d = 0.31, 95% CI [-0.08, 0.70]. Following this, we conducted a one-way ANCOVA with follow-up FSIQ as a dependent variable, experimental condition as an independent variable, and baseline FSIQ score as a covariate, in order to assess the impact of SMART after controlling for baseline. Results from this model indicated that there was no significant main effect of experimental condition on follow-up scores after controlling for baseline IQ, *F*(1, 29) = 0.52, *p* = .48, partial eta squared = .018. The estimated marginal mean (EMMs) at follow-up was 100, *95% CI* [96.4, 104], for the control group, and 102, *95% CI* [98.3, 106] for the SMART group; see Figure 4.



**Figure 4.** Estimated marginal means for IQ scores at follow-up for the SMART and control conditions after controlling for baseline IQ. Error bars represent 95% confidence intervals around the estimated marginal means.

We then quantified the percentage of participants who exhibited changes in IQ scores across a range of “meaningful differences” using a PAES specification curve. No participants in the SMART condition exhibited a decrease of more than 4 IQ points compared to 41% of participants in the control condition. On the other end, no participant in the control condition exhibited an increase greater than 8 IQ points, compared to 27% in the SMART condition. This specification curve can be examined in detail in Figure 5.



**Figure 5.** Specification curve for the PAES illustrating the proportion of participants in each condition whose IQ change from T1 to T2 was “meaningful”, defined as being greater than *n* points of change, for a range of threshold values from -15 to +15 IQ points (i.e., 1 SD).

We also examined this trend graphically, as can be seen in Figures 6 and 7. When examining the Reliable Change Index (RCI) scores these effects were less pronounced. Only one participant exhibited an RCI greater than 1.96 (SE) in the entire sample (this participant was from the SMART condition). Additionally, one participant exhibited an RCI less than -1.96, indicating a significant decrease from T1 to T2 (this participant was in the control condition).



**Figure 6.** Baseline and follow-up IQ scores for participants in the SMART condition.



**Figure 7.** Baseline and follow-up IQ scores for participants in the control condition.

Finally, we sought to determine the extent to which controlling for attentional abilities would attenuate observed effects of SMART training. To do this, we conducted two sets of three ANCOVAs. The first set of ANCOVAs consisted of identical analyses to our ANOVA above (i.e., with IQ score as DV, and condition and timepoint as IVs) while also adding one of the three attention measures as a covariate. The second set of ANCOVAs consisted of identical analyses to our ANCOVA above (i.e., with follow-up IQ scores as DV, experimental condition as IV, and baseline IQ as a covariate) while also including one of the three attention measures as an additional covariate. We also utilized Bonferroni correction to control for Type I error rates associated with multiple comparisons for each family of tests (i.e., the alpha level for these analyses was set to .0167). Firstly, we detail the first set of tests (i.e., those not adding baseline IQ as an additional covariate). When controlling for sustained attention, there was a significant main effect for condition (*F*(1, 58) = 37.83, *p* < .001, partial eta squared = 0.391), and no significant effect of timepoint (*F*(1, 58) = 0.68, *p* = .415, partial eta squared = 0.011). Most importantly, there was no significant interaction effect after Bonferroni correction between condition and timepoint in predicting IQ scores, *F*(1, 58) = 4.192, *p* = .045, partial eta squared = 0.066 (EMM: SMART baseline = 89.4, 95% CI [84.7, 94], SMART follow-up = 95.7, 95% CI [91.1, 100], control baseline = 108.5, 95% CI [104.1, 113], control follow-up = 105.8, 95% CI [101.4, 110]). For selective attention, the main effect of condition was significant, *F*(1, 58) = 18.18, *p* = .<.001, partial eta squared = 0.236. We did not find a significant effect for either the main effect of timepoint, *F*(1, 58) = 0.671, *p* = .416, partial eta squared = 0.011, or the interaction effect following Bonferonni correction *F*(1, 58) = 4.16, *p* = .046, partial eta squared = 0.066 (EMM: SMART baseline = 89.6, 95% CI [84.3, 94.9], SMART follow-up = 95.9, 95% CI [90.7, 101.2], control baseline = 108.3, 95% CI [103.4, 113.2], control follow-up = 105.6, 95% CI [100.7, 110.5]). The same pattern was also observed for everyday attention: a significant main effect of condition (*F*(1, 58) = 35.00, *p* < .001, partial eta squared = 0.372), no significant main effect of timepoint (, *F*(1, 58) = 0.671, *p* = 416, partial eta squared = 0.011) and no significant interaction effect following Bonferonni correction for multiple comparisons (, *F*(1, 58) = 4.167, *p* = .046, partial eta squared = 0.066; EMM: SMART baseline = 89.8 , 95% CI [85.1, 94.4], SMART follow-up = 96.1, 95% CI [91.4, 100.7], control baseline = 108.2, 95% CI [103.8, 112.5], control follow-up = 105.5, 95% CI [101.1, 109.8]).

Finally, we detail the second set of tests (i.e., those also adding baseline IQ as an additional covariate). After controlling for sustained attention and baseline IQ, there was no significant main effect of condition on follow-up IQ scores, *F*(1, 28) = 0.69, *p* = .684, partial eta squared = .006; this was true also for the analyses with selective attention (*F*(1, 28) = 1.20, *p* = .283, partial eta squared = .041) and everyday attention (*F*(1, 28) = 1.22, *p* = .279, partial eta squared = .042).

**Discussion**

The current study attempted to replicate the widely reported finding that SMART relational skills training has a significant impact on intelligence quotients using established indices. Several such replications have already been conducted across several laboratories (e.g., Colbert et al., 2018, Hayes & Stewart, 2016; Mcloughlin et al., 2020, 2022; see also May et al. 2022 for a systematic review). Nevertheless, the extraordinary claim that improvements in relational skills lead to attendant increases in IQ scores requires extraordinary evidence, including evidence based on replicability. Such evidence can come in the form of large randomized control trials, that are not without their shortcomings (see Hayes et al., in press; Keenan & Dillenberger, 2011). It can also come from within very well-controlled smaller-N studies (particularly when the effect of interest is large), of which the current study is an example. In the case of our study here, the statistical control for baseline intelligence scores when assessing the relative impacts of the two experimental conditions eliminated the relative impact of SMART. The same outcome was arrived at when post-intervention IQ scores across groups were compared while controlling simultaneously for both baseline IQ and attentional skills, and indeed while controlling for attentional skills alone.

Interestingly, selective attention scores were found to correlate significantly with baseline IQ scores, corroborating the theoretical literature outlined in the introduction regarding the relationship between intelligence and attentional skills. In contrast, sustained attention did not correlate with baseline intelligence levels. Nevertheless, the failure to observe an interaction between timepoint and treatment group while controlling for selective attention might suggest that selective attention may in fact represent a boundary condition for the effects of SMART. Of course, this same interpretation could be applied to sustained attention. Thus, it is difficult to know in this case whether or not the attentional variables control for psychologically significant inter-individual variances not attributable to the training itself. While such a possibility could be examined using the current data set (e.g., by examining whether changes in FSIQ in the SMART condition are moderated by baseline attention scores), our small sample size constrains the confidence one would have in any inferences drawn from such an analysis.

Of course, it may not necessarily be the case that these attentional skills represent a boundary condition for the effectiveness of SMART. One alternative explanation can be found in terms of statistical power: our small sample size may be such that the addition of *any* new parameters to our models (e.g., controlling for baseline scores) may reduce the power of our analysis to detect any effect. Importantly, however, this illustrates the importance of *a priori* power analyses and sufficient sample sizes when conducting research using SMART (or indeed, in general). The literature on SMART to date has deservedly been criticized for its generally small sample sizes (May et al., 2022), and our findings here further illustrate the need for such improvements in future work. Importantly, however, our work has uncovered possible boundary conditions for the effectiveness of SMART, and this issue should be pursued more vigorously in future studies.

One further possible explanation for our findings may be found in terms of regression to the mean amongst the treatment group, which had a substantially lower baseline IQ score than the control group. While the IQ gain from pre- to post-treatment was significant for this group only, the covariate analysis indicates a null effect relative to the control group once the disparities in baseline scores are controlled for. While this does not eliminate the possibility that SMART may have impacted IQ, it does suggest that at minimum, any true impact of SMART is necessarily entangled with an instance of regression to the mean. Thus, the current study serves as a salutary lesson in the importance of statistical power in studies of this kind. Several previous small sample studies have not controlled for baseline IQ, albeit having engaged in superior form of random participant assignment and participant matching across groups (e.g., Colbert et al., 2018; Hayes et al., 2016). Indeed, the non-random assignment of participants to experimental groups in this study may have had effects on data distributions that was not unrelated to the small sample size. In effect, this study’s design represents a “perfect storm” of risks for Type 1 error of an intervention effect, and the statistical analyses confirm that this indeed may well have been the case. Indeed, our study might be represented as a sort of case study of inadequate experimental design raising the risk of Type 1 error, precisely as warned against in a recent systematic analysis of SMART by May et al. (2022).

While future studies should of course be more cautious about participant assignment and engage in proper participant matching across experimental conditions, it may be good practice to always consider baseline intelligence as a covariate in the analysis of this type of data going forward (Zhang et al., 2014). Indeed, one recent review of cognitive training interventions (Gobet & Sala, 2022) suggested that most, if not all, reported effects of such training on measures of intelligence could be accounted for by sampling errors, low statistical power and inadequately interrogative statistical methods. Such concerns are not addressed by the current study. Instead, the current results point to the possibility that the *prima facia* (uncontrolled) positive effects of SMART reported here were due to regression of IQ scores to the mean on the post-intervention measures within the treatment group.

To place the foregoing concerns in context, it is worth considering what the literature so far has found regarding the relationship between baseline IQ and IQ gain following SMART. Only four studies have examined this relationship tangentially to their core analysis. Three of these studies found that baseline IQ was not related to the degree of IQ gain resulting from SMART. The first of these (Cassidy et al., 2011, Experiment 2) employed a very small sample size (n=8) which could reasonably be expected to compromise the reliability of a correlational analysis. The second of these (Cassidy et al., 2016) had a sample size of 15 participants in the treatment group, and also failed to find a correlation between baseline IQ and raw IQ gain. The third of the studies (Amd & Roche, 2018) employed a respectable sample of 52 participants, but had poor adherence to the training protocol. However, that study did find that baseline IQ predicted progress with the training in terms of training stages completed using a regression analysis. Finally, a fourth study (McLoughlin, 2022) employed a larger sample of 43 experimental and 27 control participants and found a strong negative correlation between baseline non-verbal IQ (NVIQ) measured using the Kaufman Brief Intelligence test (KBit; Kaufman & Kaufman, 2004) and T1-T2 score change. The researchers found that baseline NVIQ explained 30.5% of the variance in the NVIQ from pre to post-intervention and also predicted number of training stages completed. Thus, we as of yet have no definitively clear understanding of the relationship between baseline IQ and the impact of SMART, although more studies than not (limited as they are) have found that IQ gain is not a function of baseline IQ. Notably, none of the foregoing studies employed *a priori* power analyses to adequately power these analyses. This question is worth further pursuit in-and-of-itself in a well-powered future study. Regardless of whether or not IQ gain is limited by baseline IQ and attentional skill variables, it is worth reiterating that the small sample size employed here, combined with poor participant sampling methods, have compromised the study to the point that the *prima facia* significant effects established for SMART found here have become ambiguous and difficult to interpret upon closer inspection.

It may also be considered potentially problematic that different researchers were used to administer assessments at baseline and follow-up. Depending on how this procedure is interpreted, it could be viewed as a potential source of variation within the current findings. While minor variation in scores is possible across different researchers based on subjective judgements in some of the verbal items on the WASI-II, in fact a better explanation for variation across administrations has to do with the less than ideal test-retest coefficient of the WASI-II (0.7 in the current study for the control group). That said, by employing researchers that were blind to both condition assignment *and time point*, the current procedure may even have protected more strongly against experimenter bias then might a procedure in which the researcher was blind only to condition assignment but aware of timepoint. In studies of this kind, however, random test score variations and experimenter error in the direction of the hypothesis can never be fully ruled out.

**Sample size, non-random assignment and our change indices**

While the current study would appear to have suffered from risk of Type 1 error as a result of poor participant assignment and small sample size, we may potentially risk a Type II error if we were not to consider the apparent clinical significance of individual participant performances at the descriptive level. With regard to the PAES analysis, and as per the PAES specification curve, IQ gains for the intervention group were reliable descriptively at the individual-level, with a reduction in IQ scores being recorded for only one participant in the intervention group. In contrast, eleven participants in the control group exhibited a descriptive decrease in IQ scores across the same period. In effect, the robustness of this effect at the individual (descriptive) level provides some sense that the changes observed at the group-level, although we cannot rule out these effects being due to regression to the mean. The PAES measure provides clarity on the pronouncedness of these effects for individuals across varying levels of meaningful effect sizes, allowing readers to examine the relative impact of SMART at levels which they may individually consider “meaningful”. Indeed, “meaningful” effect sizes will vary across contexts and the specific goals of the researcher. This is particularly critical given the recent emphasis on the need for statistical approaches at the individual-level within CBS (see Hayes et al., 2021; Hayes et al., in press).

 The PAES analysis found that across a range of IQ gain thresholds, participants in the intervention group were consistently more likely to have reached that gain threshold than those in the waiting control group. This is an important observation because, despite the fact that many of the IQ gains were not clinically significant (see below), the intervention group participants experienced IQ increases that were more likely larger at every calculated threshold from +1 to +15 IQ points. In effect, there was a clear shift in IQ in an upward direction more reliably for the intervention group compared to the controls, regardless of what gain threshold definition we employed. Whether this was due to SMART, regression to the mean effects, or a combination remains to be determined.

On the other hand, when we apply a more robust criterion of clinically significant change (i.e., RCI), the difference between groups is much less apparent. The RCI is calculated as a function of the observed standard deviation of scores, and test-retest coefficient, in the control group. In the current data set, the observed SD was within normal range (around 9.5) but the test-retest coefficient for the WASI-II was relatively low at r = .70. This less-than-optimal test-retest reliability may constrain the sensitivity of the RCI metric. Future studies should aim to employ assessment measures with superior test-retest reliability, such as the WAIS-V, although doing so in the absence of improved methodological control is unlikely to yield considerably different results to those obtained here. Nevertheless, it is important to consider that in principle a consistent pattern of greater IQ changes at various thresholds of meaningful change were observed for the intervention group here, although such effects require replication in a more balanced sample in order to eliminate regression to the mean as an explanation.

It should be pointed out, at this stage, that we should never mistake increases in sample size as in and of themselves constituting increases in experimental rigor. Indeed, greater rigor can be achieved by having even smaller sample sizes and improving experimental control and the incisiveness of data analysis methods (e.g., multiple baselines designs). Indeed, the very interesting individual change patterns observed within and across treatment conditions in the current study strongly suggest effects that could be studied more deeply using single case experimental designs. In effect, by pursuing the current research question regarding variability in IQ gain as a result of multiple exemplar relational skills training and as a function of baseline IQ score and attentional ability, we would be fulfilling one of the recommendations provided by Hayes et al. (2021) and moving towards a more idiographic approach within our field. Specifically, Hayes et al. (2021) recommend that progress cannot be made in the contextual behavioral science field with an overreliance on the randomized control design. Instead, more creative research designs are necessary to improve the quality of interventions and tailor these to individuals based on careful measurement of multiple covariates that contextualize the effective treatment in a given case. This indeed was the aim of the current study insofar as it considered each participant’s attentional skill and IQ level as an important contextualizing factor influencing the impact of SMART. However, the current study fell between two stools in neither providing a sufficient sample size and power to function as an adequate group design and the omission of single case experimental design techniques that would qualify it as a powerful small *n* design.

Although we must interpret these null effects honestly and openly, it is also important to point out that a single null effect is not indicative of a systemic failure of the SMART approach and theoretical model (i.e., RFT). That is, based on the rationale underlying the null hypothesis statistical testing (NHST) framework, a statistically true effect will produce false negatives in one in every twenty studies in the long run when adopting an alpha threshold of .05. Of course, at the core of the NHST framework is the necessity for sufficient power in analyses. The small samples in previous SMART research are likely a function of the generally large effects that SMART produces and constraints on time and resources. However, as the agenda of SMART research becomes more elaborate, larger samples are simply a necessity. In addition, such studies would benefit from medium or long-term follow-up measures as both a method with which two establish a more reliable picture of the effects of SMART, but also to assess the maintenance of such effects overtime for its own sake. In the meantime, however, null results such as observed in the current study should be embraced as contributing to a broader and growing pool of evidence regarding the efficaciousness (and limits) of SMART.

**Conclusion**

Our results provide several key insights into research on SMART. On the one hand, we replicated the frequently observed effect that those in the SMART condition showed greater increases in IQ than those in the control condition. However, when digging deeper, our results painted a much murkier picture: the apparent impact of SMART appeared to be driven primarily by a regression to the mean effect as a function of baseline differences in IQ and attention. When controlling for these differences, the effect of SMART was eliminated. This null effect between treatment and control groups when controlling for baseline IQ and attentional skills may suggest either that these two variables represent real boundary conditions for the observed effect, *or* that the study was underpowered sufficiently to determine the efficaciousness of SMART when controlling for any extra parameters at all. Although it would be easy to equivocate about this, our uncertainty around this only highlights the need for better-controlled experimental designs, be they in the form of larger sample sizes within NHST or single-case experimental designs. In any case, our results here highlight that to generate deeper insights into the efficacy and limitations of SMART, we must face this challenge head on, methodologically and statistically.

**References**

Amd, M., & Roche, B. (2018). Assessing the effects of a relational training intervention on fluid intelligence among a sample of socially disadvantaged children in Bangladesh. *The Psychological Record, 68*(2), 141–149.

Barnes, D., McCullagh, P. D., & Keenan, M. (1990). Equivalence class formation in non-hearing impaired children and hearing impaired children. *Analysis of Verbal Behavior, 8,* 19–30.

Bennett, G. K., Seashore, H. G., & Westman, A. G. (1990). *The Differential Aptitude Test.* San Antonio, Texas (5th ed.). Psychological Corporation.

Bernardi et al., 2017, Bernardi, M., Leonard, H. C., Hill, E. L., Botting, N., & Henry, L. A. (2018). Executive functions in children with developmental coordination disorder: a 2‐year follow‐up study. *Developmental Medicine & Child Neurology*, *60*(3), 306-313.

Broadbent, D. E., (1958). *Communication and perception*. New York: Pergammon.

Burgaleta M., Johnson W., Waber D. P., Colom R., Karama S. (2014). Cognitive ability changes and dynamics of cortical thickness development in healthy children and adolescents. *Neuroimage,* *84*, 810–819.

Cassidy, S., Roche, B., Colbert, D., Stewart, I., & Grey, I. (2016). A relational frame skills training intervention to increase general intelligence and scholastic aptitude. *Learning and Individual Differences, 47*, 222–235.

Cassidy, S., Roche, B., & Hayes, S. C. (2011). A relational frame training intervention to raise intelligence quotients: A pilot study. *The Psychological Record, 61*(2), 173–198.

Cassidy, S., Roche, B., & O’Hora, D. (2010). Relational Frame Theory and human intelligence. *European Journal of Behavior Analysis, 11*(1), 37-51.

Carol, J.B. (1976). Psychometric tests as cognitive tasks: A new ”structure of the intellect?” In L.B. Resnick (Ed.), *The nature of intelligence*. Hillsdale, N.J. :Erlbaum.

Colbert, D., Dobutowitsch, M., Roche, B., & Brophy, C. (2017). The proxy-measurement of intelligence quotients using a relational skills abilities index. *Learning and Individual Differences, 57,* 114–122

Colbert, D., Malone, A., Barrett, S., & Roche, B. (2020). The Relational Abilities Index+: Initial validation of a functionally understood proxy measure for intelligence. *Perspectives on Behavior Science, 43*(1), 189–213.

Cowan, N., Fristoe, N.M., Elliott, E.M., Bruner, R. P., & Saults, J. S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition* 34, 1754–1768.

Daley, D. & Birchwood, J. (2010). ADHD and academic performance: why does ADHD impact on academic performance and what can be done to support ADHD children in the classroom? *Child: Care, Health and Development, 35,* 455-464.

de Rose, J. C., de Souza, D. G., Rossito, A. L., & de Rose, T. M. S. (1992). Stimulus equivalence and generalization in reading after matching to sample by exclusion. In S. C. Hayes & L. J. Hayes (Eds.), *Understanding verbal relations* (pp. 69–82). Context Press: Reno.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93*, 880–896.

Dixon, M. R., Belisle, J., & Stanley, C. R. (2018). Derived Relational Responding and Intelligence: Assessing the Relationship Between the PEAK-E Pre-assessment and IQ with Individuals with Autism and Related Disabilities. *The Psychological Record.**68,*419–430.

Dixon MR, Belisle J, Whiting SW, Rowsey KE. (2014). Normative sample of the PEAK relational training system: direct training module and subsequent comparisons to individuals with autism. *Research in Autism Spectrum Disorders, 8,* 1597–1606

Dixon, M. R., Paliliunas, D., Barron, B. F., Schmick, A. M., & Stanley, C. R. (2019). Randomized Controlled Trial Evaluation of ABA Content on IQ Gains in Children with Autism *Journal of Behavioral Education***,***30*, 455–477.

Dugdale, N., & Lowe, C. F. (2000). Testing for symmetry in the conditional discriminations of language-trained chimpanzees. *Journal of the Experimental Analysis of Behavior, 73*, 5–22.

Dunkel-Jackson, S. M., & Dixon, M. R. (2018). Promoting Generalized Advanced Language Skills of Children in Intensive Behavioral Intervention with Promoting the Emergence of Advanced Knowledge Generalization Module (PEAK-G). *Behavior Analysis in Practice*, *11*, 289–306.

Edwards, L., Figueras, B., Mellanby, J., & Langdon, D. (2011). Verbal and spatial analogical reasoning in deaf and hearing children: The role of grammar and vocabulary. *Journal of Deaf Studies and Deaf Education, 16*(2), 189–197.

Farrington-Flint, L., Canobi, K. H., Wood, C., & Faulkner, D. (2007). The role of relational reasoning in children's addition concepts. *British Journal of Developmental Psychology, 25*(2), 227–246.

Fergusson, D. M., Lynskey, M. T., & Horwood, L. J. (1997). Attentional difficulties in middle childhood and psychosocial outcomes in young adulthood. *Journal of Child Psychology and Psychiatry*, *38,* 633–644 .

Fergusson , D. M. , & Horwood , J. L. (1993). Early disruptive behavior, IQ, and later school achievement and delinquent behavior. *Journal of Abnormal Child Psychology*, *23*, 183-199.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Galera, C., Melchior, M., Chastang, J. F., Bouvard, M. P., & Fombonne, E. ( 2009). Childhood and adolescent hyperactivity-inattention symptoms and academic achievement 8 years later: The GAZEL Youth study. *Psychological Medicine*, *39*, 1895–1906.

Gazzaniga, M. S., & Halpern, D. F. (2015). *Psychological science (fifth edition)*. W. W. Norton & Company.

Grice, J. W., Medellin,E.,  Jones, I., Horvath, S., McDaniel, H., O’lansen, C., Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science, 3,* 443-455.

# Hayes, S. C. (1993). Analytic goals and the varieties of scientific contextualism. In S. C. Hayes, L. J. Hayes, H. W. Reese, & T. R. Sarbin (Eds.), *Varieties of scientific contextualism* (pp. 11–27). Context Press.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition.* Kluwer Academic/Plenum Publishers.

Hayes, S. C., Ciarrochi, J., Hofmann, S. G., Chin, F., & Sahdra, B. (in press). Evolving an idionomic approach to processes of change: Towards a unified personalized science of human improvement. *Behaviour Research and Therapy*.

Hayes, S.C., Merwin, R.M., McHugh, L, Sandoz. E.K., A-Tjak, J.G.L, Ruiz, F.J., Barnes-Holmes, D., Bricker, J.B., Ciarrochi, J., Dixon, M.R., Fung, K.P.L., Gloster, A.T., Gobin, R.L., Gould, E.R., Hofmann, S.G., Kasujja, R,, Karekla, M., Luciano, C., McCracken, L.M. (2021). Report of the ACBS Task Force on the strategies and tactics of contextual behavioral science research. *Journal of Contextual Behavioral Science, 20,* 172-183.

Hock, H. H. (2003). Analogical change. In B. Josephs & R. Janda (Eds.), *The handbook of historical linguistics* (pp. 441–480). Oxford, UK: Blackwell.

Hommel, B., Chapman, C.S., Cisek, P., Neyedli, H.F., Song, J.H., & Welsh, T.N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics, 81,* 2288-2303.

Jacobson, N.S., & Truax., P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–9.

Kaufman, A. S., Flanagan, D. P., Alfonso, V. C., & Mascolo, J. T. (2006). Test review: Wechsler intelligence scale for children, fourth edition (WISC-IV).  *Journal of Psychoeducational Assessment, 24,* 278–295.

Kaufman, A. S., & Kaufman, N. L. (2004*). Kaufman Assessment. Battery for Children, Second edition*. Circle Pines, MN: AGS.

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.

Ko, L. W., Komarov, O., Hairston, W. D., Jung, T. P., & Lin, C. T. (2017). Sustained Attention in Real Classroom Settings: An EEG Study. *Frontiers in human neuroscience*, *11*, 388. https://doi.org/10.3389/fnhum.2017.00388

Langberg, J. M., Molina, B.S.G., Arnold, L. E., Epstein, J. N., Altaye, M., Hinshaw, S. P., Swanson, J.M., Wigal, T., & Hechtman, L. (2011). Patterns and predictors of adolescent academic achievement and performance in a sample of children with attention-deficit/hyperactivity disorder. *Journal of Clinical Child & Adolescent Psychology, 40,* 519-531.

Lenth R (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.2, https://CRAN.R-project.org/package=emmeans

Lonergan, A., Doyle, C., Cassidy, C., MacSweeney Mahon, S., Roche, R. A., Boran, L., & Bramham, J. (2019). A meta-analysis of executive functioning in dyslexia with consideration of the impact of comorbid ADHD. *Journal of Cognitive Psychology*, *31*(7), 725-749.

Mackay, H. A. (1985). Stimulus equivalence in rudimentary reading and spelling. *Analysis & Intervention in Developmental Disabilities, 5*(4), 373–387.

Manly, T., Anderson, V., Crawford, J., George, M., & Robertson, I. H. (2016). *Tests of Everyday Attention for Children, Second Edition (TEA-Ch2)*. London, UK: Pearson Education Ltd.

May, R. J., Tyndall, I, McTiernan, A., Rodriguez, G., & McLoughlin, S. (2022). The Impact of the SMART Program on cognitive and academic skills: A systematic review and meta-analysis. *British Journal of Educational Technology*.

McCrimmon, A. W., & Smith, A. D. (2013). Review of *the* *Wechsler abbreviated scale of intelligence, second edition (WASI-II),*by D. Wechsler].  *Journal of Psychoeducational Assessment, 31*, 337–341.

McLoughlin S, Tyndall I, Pereira A. (2020). Convergence of multiple fields on a relational reasoning approach to cognition. *Intelligence, 83*, 1-10.

McLoughlin, S., Tyndall, I., Pereira, A., & Mulhern, T. (2020). Non-verbal IQ gains from relational operant training explains variance in educational attainment: An active-controlled feasibility study. Journal of Cognitive Enhancement.

McLoughlin, S., Tyndall, I., & Pereira, A. (2022). Relational operant skills training increases standardised matrices in adolescents: A stratified active-controlled trial. *Journal of Behavioural Education*, *31*, 298-325. <https://doi.org/10.1007/s10864-020-09399-x>

Moran, L., Walsh, L., Stewart, I., McElwee, J., & Ming, S. (2015). Correlating derived relational responding with linguistic and cognitive ability in children with Autism Spectrum Disorders. *Research in Autism Spectrum Disorders, 19,* 32–43.

Mullet & Rinn, 2015.  Mullet, D. R., & Rinn, A. N. (2015). Giftedness and ADHD: Identification, misdiagnosis, and dual diagnosis. *Roeper Review: A Journal on Gifted Education, 37*, 195–207.

Nippold, M. A., & Sullivan, M. P. (1987). Verbal and perceptual analogical reasoning and proportional metaphor comprehension in young children. *Journal of Speech & Hearing Research, 30*, 367–376.

Raven, J., Raven, J. C., & Court, J. H. (1998b). *Manual for Raven’s Progressive Matrices and Vocabulary Scales. Section 5: The Mill Hill Vocabulary Scale*. San Antonio, TX: Harcourt Assessment.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech & Hearing Research, 14*(1), 5–13.

Schlinger, H. D. (2003). The myth of intelligence. *The Psychological Record, 53*, 15–32.

Schweizer, K & Moosbrugger, H. (2004). Attention and working memory as predictors of intelligence. *Intelligence, 32,* 329-347.

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., Evans, A., Rapoport, J., and Giegg, J. (2006).  Intellectual ability and cortical development in children and adolescents.  *Nature.  440*, 676-9.

Spearman, C. (1927). *The abilities of man*. New York: MacMillan.

Stankov, L. (1989). Attentional resources and intelligence: A disappearing link.

*Personality and Individual Differences, 10*, 957-968.

Thirus, J., Starbrink, M., & Jansson, B. (2016). Relational frame theory, mathematical and logical skills: A multiple exemplar training intervention to enhance intellectual performance. *International Journal of Psychology and Psychological Therapy, 16*, 141–155.

Thurstone, L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Torchiano, M. (2020). *effsize: Efficient Effect Size Computation*.  doi:10.5281/zenodo.1480624, R package version 0.8.1, https://CRAN.R-project.org/package=effsize.

Twisk, J., Bosman, L., Hoekstra, T., Rijnhart, J., Welten, M., & Heymans, M. (2018). Different ways to estimate treatment effects in randomised controlled trials. *Contemporary Clinical Trials Communications, 10*, 80-85.

Vilardaga, R., Hayes, S. C., Levin, M. E., & Muto, T. (2009). Creating a strategy for progress: a contextual behavioral science approach. *The Behavior Analyst 32*(1), 105-

33.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L, Miller, E., Bache, S.M., Müller, K., Ooms J, Robinson D, Seidel DP, Spinu V., … Yutani, H (2019). Welcome to the tidyverse.  *Journal of Open Source Software*, 43, 1686. [doi:10.21105/joss.01686](https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.21105%2Fjoss.01686&data=05%7C01%7CBryan.T.Roche%40mu.ie%7C2980a6bd186e477cb72308db141d9594%7C1454f5ccbb354685bbd98621fd8055c9%7C0%7C0%7C638125889535564701%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=g3t9zQggM07UlRvqdiiNbySQvbDhKOf9zkix68YDwrY%3D&reserved=0).

Zhang, S., Paul, J., Nantha-Aree, M., Buckley, N., Shahzad, U., Cheng, J., DeBeer, J., Winemaker, M., Wismer, D., Punthakee, D., Avram, V., & Thabane, L. (2014). Empirical comparison of four baseline covariate adjustment methods in analysis of continuous outcomes in randomized controlled trials. *Clinical epidemiology*, *6*, 227–235. https://doi.org/10.2147/CLEP.S56554

 Ziegler M (2016). *ClinicalSig.* R package version 0.1. [https://cran.r-project.org/package=clinsig](https://cran.r-project.org/package%3Dclinsig)