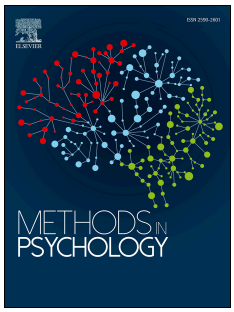

Characterising developmental disorders: Towards better group comparisons in developmental research

Gnanathusharan Rajendran, Peter Edward McKenna, Martin Corley

PII: S2590-2601(22)00012-1

DOI: https://doi.org/10.1016/j.metip.2022.100101

Reference: METIP 100101

To appear in: *Methods in Psychology*

Received Date: 12 November 2020

Revised Date: 30 May 2022

Accepted Date: 26 September 2022

Please cite this article as: Rajendran, G., McKenna, P.E., Corley, M., Characterising developmental disorders: Towards better group comparisons in developmental research, *Methods in Psychology* (2022), doi: https://doi.org/10.1016/j.metip.2022.100101.

Characterising Developmental Disorders:
Towards Better Group Comparisons in Developmental Research

Gnanathusharan Rajendran*

Heriot-Watt University

Peter Edward McKenna

Heriot-Watt University

Martin Corley

The University of Edinburgh

*Corresponding Author

Characterising Developmental Disorders:

Towards Better Group Comparisons in Developmental Research

Abstract

Investigating a developmental disorder (DD) must consider three questions: 1) Who are you measuring? 2) How are you measuring? and 3) What are you measuring? The decisions made – from selecting an appropriate comparison group and deciding what to match on, to the very measurements and statistics we use – directly inform the characterisations of a DD. We discuss the implications of these decisions through a critique of approaches: Including coping with heterogeneity, group matching and the importance of studying processes as well as outcomes. Our aim is to help researchers navigate this field of research and make informed decisions.

Keywords: Developmental Disorders; Group matching; Appropriate measurement

1. Background

The definitions of developmental disorders (DDs) have come under intense scrutiny in recent years. The Fifth Edition of the Diagnostic and Statistical Manual of Mental Disorders of the America

Psychiatric Association (DSM-V, APA, 2013) includes definitions of core deficits which have been described as "at best, a dictionary, creating a set of labels" by funding agencies such as the National Institute of Mental Health (NIMH, Insel, 2013). In the UK, the British Psychological Society's Division of Clinical Psychology has disavowed the medical model of disorders such as ADHD (British Psychological Society, 2013). So, even after decades of empirical research into DDs we are still trying to reach a consensus as to their nature.

The emphasis on empirical investigation of the differences between those considered to have a developmental disorder (DD) and their typically developing peers dates from the 1960s (e.g. Hermelin & O'Connor, 1967). Among the successes of this approach has been a better understanding of the behavioural phenotypes associated with different DDs, for instance in terms of memory, language and perception. While acknowledging new insights, it is also important to recognise that the empirical approach is confounded by the *theory* of the DD that you are investigating – and, so, may only offer a narrow interpretation of DD individuals' abilities. For example, in research into autism spectrum conditions (henceforth, autism), supporters of perceptual theories argue that 'social-first' models (e.g. the theory of mind deficit model), "…appear to us too narrow to encompass the entire range of positive symptoms or the enhanced performance of savant and non-savant autistics" (Mottron et al., 2006, p. 29). Even theoretically sound, well-crafted, and empirically robust experiments can fail to capture the essence of atypicality.

Here, we address this issue and explain why difficulties in characterising a developmental condition are inherent when making an empirical comparison between groups. We consider both experimental and statistical control for inter-group differences. We point out that even if pitfalls can be avoided, the decision of *what* to control for and *why* to control for it is theoretically grounded – and that without this acknowledgement researchers may fall foul of circularity, by eventually finding exactly what they were looking for.

2. Paper Rationale

Every child develops differently. Children typically begin to say their first word around the time of their first birthday. However, some children utter their first word at 10 months and others at 15 months (Saxton, 2010). The ages at which children first walk, read, or perform any one of a myriad of other everyday physical and cognitive feats, and the speeds at which their new skills develop, vary equally widely. Despite this heterogeneity, some children's development is clearly outside the norm. Whereas it is relatively easy to establish that a child with autism or with Williams Syndrome is 'different', the very heterogeneity of development makes it difficult to determine in exactly what way they are different, or to infer the underlying cognitive skills that are affected.

This paper is not about how to determine whether an individual child has a DD; rather, it asks how – given a diagnosis – we can effectively use research methods to generate a better understanding of the ways that groups of children differ from the norm. We start with the generally accepted causal pathway to developing a developmental disorder, i.e., that it entails biological and/or environmental disruption during defined periods in development, with an outcome that has long-term functional consequences (Leonardo & Hen, 2008). However, experiments investigating the ways in which atypical pathways are different from normative pathways require the researcher to avoid several potential pitfalls; even in the best case, an experiment requires a theory about the ways in which a developmental pathway *can* differ. Moreover, it important to appreciate that DD experiments are not true experiments because participants cannot be randomly assigned to treatment groups – and this has implications in, for example, trying to covary out any differences (see section 4.2, Factoring out differences).

So, while a clinician may instinctively recognise that a child presents atypical behaviour or cognition, it is up to the DD researcher to capture this atypicality with carefully designed experiments. Once these experiments have been conducted, the insights and conclusions that follow should be representative of the data and – importantly – be interpreted with caution. To clarify, our message is not that most extant DD research has been poorly executed. It is that research in this domain is a minefield that requires careful navigation, and that there is no 'perfect' solution.

In the following section, 'Who are you measuring?' we discuss the challenges of experimental control; chiefly, those related to group matching.

3. Who are you measuring? An issue of experimental control

3.1 Issues of recruitment and heterogeneity

Baron-Cohen et al. (1985) proposed that autism could be explained as a deficit in theory of mind (ToM: "the active process of inferring a particular mental state in a particular context", Perez-Osorio et al., 2021, p.317). Therefore, a defining aspect of using such a developmental deficit model for a DD is that it is comparative: i.e., it is only understandable with reference to a non-deficit 'norm'. So, identifying the appropriate groups for comparison[1] and determining the aspects of development that can be compared is a problem with far-reaching consequences but with no unique solutions. Identifying the appropriate groups for comparison and determining the aspects of development that can be compared is a problem with far-reaching consequences but with no unique solutions. Key questions that are often addressed using comparator groups include those of *uniqueness* (i.e. is a deficit associated with a specific disorder and that disorder alone), *universality* (does a deficit apply to all individuals who have been identified as having a specific DD), and *domain-specificity* (whether deficits in particular tasks are tied to specific modalities.

In reaching conclusions, researchers must find solutions to two problems intrinsic to DD research. The first is that individuals recruited for a DD study tend to be heterogeneous and this problem is exacerbated because people with a DD are relatively rare. This means that recruitment is resource-intensive and often opportunistic. The second is that the characteristics of a DD may not be fixed, but may vary over time (e.g. Järvinen-Pasley et al. 2008). Both issues have consequences for the selection and matching of control groups, affecting what is meant by 'typical' development. For example, typical development might be characterised as an average across diverse developmental trajectories (Pennington et al., 2007). We discuss several solutions to control group selection below.

[1] E.g. Down's syndrome and Typically Developing (TD) group in Baron-Cohen et al. (1985).

3.2 Group matching challenges

Group matching is important for the issue of *uniqueness*, also known as the 'discriminant validity problem'. Here, researchers need to be confident that the phenomenon under investigation relates to the particular disorder under inquiry and not to other populations. Also, several practical factors drive the creation of experimental designs when investigating DDs. One of the most important considerations is the difficulty of recruiting special populations.

Developmental Disorders vary greatly in their prevalence: for example, Williams syndrome (a genetic condition resulting from deletion of approximately 26 genes on the long arm of chromosome 7) is thought to occur relatively rarely (1 in 7,500 to 20,000 people: see Martens et al., 2008, for a review) compared to autism (1 in 100: Baird et al., 2006; see Elsabbagh et al., 2012 for world-wide variations). This means that clinical groups recruited to DD research are often opportunity samples and are, therefore, composed of a wide age-range of participants even when the focus is not on age-related differences per se. This mixture of age-ranges (and other individual characteristics like IQ) within the focal group is not ideal but is often the reality of both the disorder and the finite resources a researcher has in terms of recruitment. In this context, it is important to ensure that the data obtained from DD individuals is maximally able to provide insight into the disorder under investigation.

Even when high numbers of participants can be recruited, the profile of DD can be very 'uneven'. For example, Dawson et al. (2007) demonstrated that participants with autism performed well at a spatial intelligence (Block Design) task relative to other measures of intelligence. Here, matching[2] TD controls to DD participants is difficult, because it is unclear which aspects of the DD performance profiles represent an 'underlying competence'. So, a key question might be whether matching is best done using a Wechsler Test of Intelligence (e.g. WASI-II, Wechsler, 2011), or Raven's Progressive Matrices (Raven et al., 1998). This issue is neither unique to autism nor to IQ measures: For example,

[2] Historically, profile matching (such as through an IQ measure or other proxies for IQ, e.g. verbal mental age) has been conducted because an IQ less than 70 implies an intellectual disability and, so, researchers to have tried not to confound performance on a task with IQ (e.g. Hermelin & O'Connor, 1967).

individuals with Williams syndrome are often less proficient at tasks involving visual–spatial cognition, but tend to be relatively good at face processing (e.g. Järvinen-Pasley et al. 2008).

There are also substantial inter-individual differences in DD; for example, autism is a spectrum condition (Wing & Gould, 1979; Robinson et al., 2016), where individuals can be with or without an additional structural language impairment (Norbury, 2005; see below for broader discussion). Even in single gene developmental disorders, there can be substantial variation; for example, the intellectual disability in individuals with Fragile X syndrome (caused by the silencing of a single gene on the X chromosome: Cornish, Turk & Hagerman, 2008) also lies on a spectrum. In research, heterogeneity is an ever-present challenge for comparison group matching (Georgiades et al., 2013) – and the heterogeneity of the focal clinical group often means that the researcher has to go to great lengths to find a 'twin' match for each focal group member.

At first glance, looking for individual matches to comprise the comparison group (i.e. finding a series of suitable 'twins') might seem uncontroversial. However, there is no sensible definition of what a 'twin' is. This is because the notion of a 'twin' requires an *a priori* theory of the nature of the DD under investigation; that specific cognitive variables are definitional features of the DD in question, for example through the diagnostic criteria in DSM-V (APA, 2013) or The International Statistical Classification of Diseases and Related Health Problems (ICD-11, 2019). Moreover, not all matches are equal: In choosing a comparison group one is unlikely to be matching the 'twin' based on height. More likely they will be matched on characteristics such as IQ, verbal ability, gender, or calendar age.

Even if we accept the possibility of matching on 'non-definitional' characteristics, care should be taken about the concept of 'matching'. The purpose of any statistical analyses here is to confirm, rather than disconfirm, the null hypothesis. For this reason, the criterion for claiming a match should not be lack of significance at $\alpha = .05$, because the probability of a Type II error (rejecting the null hypothesis when it is false) can be unacceptably high at marginally non-significant values of *p*. Instead, it has been suggested that a *p*-value of >.50 indicates that the group distributions are sufficiently

overlapping for the groups to be considered properly matched on a given variable (Frick, 1995, Mervis & Klein-Tasman, 2004).

3.3 Methodological approaches to group matching

The various sub-divisions of DDs in DSM exemplify the complexity of the current diagnostic systems. Currently, these systems may be of greater benefit to clinicians than researchers. This is because the diagnostic systems are purely descriptive, and geared to individual cases, whereas it is up to the researcher to determine what kind of DD group they are working with (language impaired, low intellectual functioning etc.) – and to operationalise them accordingly. For example, a researcher may define a language impaired group as evidenced by current scores of at least -1.25 SD on the Recalling Sentences sub-test of the Clinical Evaluation of Language Fundamentals (CELF-4UK: Semel et al., 2004), or a high intellectual functioning group as indexed by a full-scale IQ of 70 or greater. From these criteria, it is up to the researcher to match comparison group(s) accordingly. It is also worth noting here that clinical diagnostic tools themselves may not be objective, given that they were created with someone's conceptualisation of what the disorder is. Ironically, the four cases that Asperger described in his seminal paper may not have even met the DSM-IV criteria for Asperger syndrome (Miller & Ozonoff, 1997).

As an illustration of the complexities of group matching, we cite the work of Norbury (e.g. Norbury, 2005). Norbury argues that individuals with autism may differ depending on whether they have a coexisting language disorder or not. Her solution is a methodological one (rather than statistical) of recruiting a typical group, and three clinical groups: those with autism plus language impairment, those with autism without language impairment, and those *without* autism but with a language impairment. We argue that Norbury's approach is a better solution despite being a more resource-intensive one, 'partialling out language impairment' at the group matching stage (i.e. language impairment is accounted for *without* using ANCOVA – see below for a more in-depth argument about the potential problems with ANCOVA when used in this way).

However, the consequence of multiple-matching is that there may be a limited set of tasks available to all sub-populations (in the example of autism, tasks may not be equally accessible to language impaired and non-language impaired groups). This means that a researcher may miss out on much of the characterisation of these groups (language impaired, non-language impaired, low intellectual functioning, high intellectual functioning etc.).

As an example of careful task choice, a longitudinal study tracked how the development of language interacted with the development of theory of mind (ToM) for three unique groups of children (autism, deaf children from hearing families [DoH], and TD) between pre-school and adolescence (Peterson & Wellman, 2019). The researchers acknowledged that a fair group comparison required careful selection of assessment tools that would not put any group at a disadvantage. For the autism and TD groups language was assessed using the Peabody Picture Vocabulary Test Third Edition (PPVT-III; Dunn & Dunn, 1997). However, DoH children did not complete the PPVT because some test items had two meanings which would require them to spell out in sign language their answer to disambiguate double meanings. Therefore, DoH children's responses would require greater effort in relative to the other groups, potentially underestimating their language ability. Instead, the DoH group completed the 22-item syntax subscale of the Clinical Evaluation of Language Fundamental – Preschool (CELF-P; Wiig et al., 1992), in which all items had a single meaning. Using these two standardised tests, Peterson and Wellman (2019) subsequently found no group differences in baseline language ability and, so, could investigate ToM with relative confidence that language ability would not explain a significant portion of ToM variance. Hierarchical regression revealed that the best predictor of ToM at adolescence was pre-school ToM score, showing that ToM developed with age in all three groups, despite their different developmental profiles.

Where suitable tasks can be identified, it has been recommended that the DD group and controls are also matched on a 'non-definitional' feature of the task, such as response time to neutral practice trials (Jarrold & Brock, 2004). Doing this ensures that the conclusions drawn from the data are

unlikely to contain artefacts from participants' misunderstandings of the task's instructions, or from execution requirements (e.g. older participants may be slower to react during computerised tests because they might be less familiar with some newer technologies).

Final Word on Group Matching

In summary, we are in favour of group-matching, but believe care must be taken on 1) the measures used for matching, because assumptions about what might be "invariant" between groups are driven by theory; 2) the choice of control groups, acknowledging the heterogeneity of many DD diagnoses; 3) the choice of tasks, ensuring that tasks are accessible to all participants and that differences in performance are not driven by uncontrolled differences between participants.

4. How are you measuring? An issue of analysis

4.1 Using standardised scores

One way of taking a more explicitly age-focused developmental approach is to compare the performance of individuals with a DD to TD population norms, using standardised tests. However, this approach has its own drawbacks. Consider IQ, which in DD research is calculated in the usual way, by using raw scores to create standard scores. However, these standard scores are derived from TD age-related population norms. So, if the IQ for someone with Down's syndrome is calculated using WAIS-II (Wechsler, 2011), their score will be based upon normative data from a TD rather than a Down's syndrome population.

Despite these limitations, this method of comparing against standardised scores can be a useful one – especially when trying to delineate a profile associated with a particular DD. Taking a profile approach might be one way of looking for differences between disorders that show similar characteristics. An example of the profiling approach, using a standardised assessment tool, is work by Alloway and colleagues (2009). Alloway et al. delineated the Working Memory[3] profiles of several DD

[3] Rather than being represented by a single number, a working memory profile provides four measures: verbal working memory, visuo-spatial working memory, verbal short-term memory and visuo-spatial short-term memory.

using a standardised test, namely the Automated Working Memory Assessment (Alloway, 2007). Their samples included children with Specific Language Impairment, Developmental Coordination Disorder (DCD), Attention-Deficit/Hyperactivity Disorder, and Asperger syndrome. These different DD profiles for verbal/visuospatial and working/short-term memory were compared and contrasted, showing that each DD was characterised by a distinctive Working Memory profile; a profile that could explain aspects of their unique behavioural phenotype.

In contrast, Dawson et al. (2007) showed that children with autism showed relatively elevated performance on the Block Design sub-test of the Wechsler Intelligence Scale for Children (WISC-III; Wechsler, 1991), in comparison to other subtests. However, the real strength of Dawson et al.'s study was that the children were also tested with Raven's Progressive Matrices (Raven et al., 1998). Dawson et al. found that children with autism had Raven's Matrix scores that were on average 30 (and in some cases 70) percentile points higher than their scores on the WISC. The typically developing comparison children showed no such discrepancy. Dawson et al., therefore, argued that individuals with autism might not be disproportionately impaired on tests of fluid intelligence (like Raven's Matrices) and that 'autistic intelligence' was typically underestimated if one used just the WISC.

Using standardised tests as a profiling tool is, therefore, one avenue for taking a developmental perspective – through using age-derived norms. However, this method is not without its limitations. Generally, the scores come from a static representation of the *endpoint* of a process, or an *ongoing* representation of processing. So, in the case of intelligence tests, a participant's score tells us neither about a child's underlying understanding, nor their processing style. So, to find a someone's underlying competence, you would need to look the *process* by which they complete the task. For example, the relative success of individuals with autism on the Block Design[4] might be due to their cognitive style (Happé,

[4] This test forms part of the Wechsler Intelligence Tests in which participants must copy – under time pressure – a design presented to them using cubes with differently patterned faces.

1999) rather than their intelligence. When the 2-D pictures of the blocks were pre-segmented, the comparison group's performance improved to a point where the group with autism were no longer superior. Shah and Frith (1993) suggested that participants with autism perceptually segmented the designs, such that presenting the designs pre-segmented provided no further benefit. A similar example comes from individuals with William syndrome and their relative elevated performance on standardised tests of face recognition (Järvinen-Pasley et al., 2008). Their performance on this endpoint does not reveal if they process faces in similar or different ways to typically developing individuals (see Section 5 for a more in-depth discussion about Williams syndrome and face processing).

Another criticism of using standardised tests is that a child's development is not tracked across time. So, this method cannot take into account that development might progress in a different way, for example via a differently-shaped trajectory than TD norms. As well as these conceptual points, there is also the issue of resources; i.e. creating standard scores is resource-intensive and usually done for psychometric test creation rather than experiments. One method that does gives age important credence, and one that can be used for experimental tasks, is that of using Developmental Trajectories (see Section 5). What follows is a discussion of a tool commonly used for matching, ANCOVA.

4.2 Factoring out differences

A frequently used statistical technique for adjusting for *a priori* group differences is through an ANCOVA analysis (see, e.g., Rutherford, 2001). Often, the language used to report ANCOVA hints at 'controlling for' variables such as IQ which may not have been successfully matched between groups. The problem with using ANCOVA in this way is that DD research is not truly experimental, because participants cannot be randomly allocated to treatment groups. Instead, research in this area is quasi-experimental and this has consequences for trying to 'equalise variables' which cannot inherently be equalised (Miller & Chapman, 2001). Miller and Chapman cite Cohen and Cohen (1983), who provide an extreme example that highlights the illogic of trying to 'control for' intrinsic properties: "Consider the fact that the difference in mean height between the mountains of the Himalayan and Catskill

ranges, adjusting for differences in atmospheric pressure, is zero!" (p. 425); the argument is that the mountain ranges cannot in any true sense be 'equated' by using atmospheric pressure as a covariate (see Vickers and Altman (2001) for a more typical and arguably less contentious use of ANCOVA, in treatment/intervention studies in which baseline scores are co-varied when comparing post-intervention scores).

In their autism study, Rogers et al. (2008) do acknowledge the near impossibility of matching multiple groups on multiple variables and argue that co-varying verbal mental age was appropriate because it was the most conservative approach. More specifically, their findings were so robust that using ANCOVA did not remove the presence of group differences. That is, 'controlling for' language ability also risked removing some of the variance associated with dependent variable, but – in their case – the reduction was not enough produce a null result. However, for other studies an ANCOVA approach might risk producing a null result for reasons of a Type II error (rejecting the null hypothesis when it is false). Also, although Rogers et al. (2008) found a significant main effect, the use of ANCOVA could have removed enough shared variance to result in a Type II error and, so, explain the null finding for their interaction analysis.

Final Word on using standardised measures and ANCOVA

In Summary, we advocate that 1) standardised measures like IQ can be used in DD research to help characterise particular samples of a given population and 2) *a priori* matching is a far better strategy for controlling group differences in 'baseline' variables than is using ANCOVA methods *post hoc.*

5. What are you measuring? Finding the right metrics.

After formulating the design of a study, and deciding how to match groups, the next step is to ensure that the dependent variables are sensitive and informative to the research objectives. Hypotheses focussed on potential *deficits* (absences), or *delays* (milestones reached later in time) predict quantifiable differences in outcome measures between DD and TD groups. Alternatively, searching for *deviance* places the emphasis on the cognitive process that the DD group uses, which may be qualitatively different from that used by the TD group. So, the conceptualisation of what 'difference' entails will shape the design and outcome measures of an experiment.

A cautionary tale about measurement comes from Williams Syndrome. Face processing in this condition seems relatively spared compared with visuo-spatial ability (Järvinen-Pasley et al., 2008; Martens et al., 2008; Tager-Flusberg et al., 2003). However, this may be misleading, because face processing is often measured using standardised tests, such as the Benton Face recognition task (Benton, et al., 1983) – and thereby seen through the lens of typically developing age-rated norms. However, when you look at the actual *manner* in which faces are processed (holistically or feature by feature), individuals with William syndrome may be recognising faces in a qualitatively different way from typically developing individuals (Annaz et al., 2009, Karmiloff-Smith, 2012; Leonard et al., 2011). For example, Riby et al. (2008) found that individuals with Williams syndrome showed greater accuracy than a comparison group in matching unfamiliar faces from internal features (cut outs of the faces including the eyes, nose and mouth) than external features (what was left after the internal features had been cut out: i.e. hair, eyes, chin). Note also that each participant with Williams syndrome was individually matched to three typically developing participants on one of three measures: verbal ability, nonverbal ability and calendar age.

As mentioned earlier, standardised scores offer limited insight to an individual's subjective processing, because scores come from *very end* of the cognitive process: e.g. comparing memory for faces using the Benton Face recognition task. But because there is an underlying process, there are

intermediary points at which a measurement could be taken (e.g. eye tracking would indicate what percentage of the time looking at faces is actually spent looking specifically at the eyes). Norbury (2014) argues for the use of eye-tracking in understanding the process of language production. For example, using the 'visual world' paradigm (Tanenhaus et al., 1995), Norbury and colleagues investigated the link between visual attention and language processing (e.g. Brock et al., 2008; Norbury et al., 2009). Brock et al. (2008) used the fact that eye-movements are sensitive to both semantic and phonological cues (e.g. on hearing the phrase "eat the cake", participants tend to look towards a picture of a cake even before the onset of the word "cake"; on hearing "beetle", they look more at a beaker than at a phonologically unrelated object like "carriage"). Utilising these phenomena, Brock et al (2008) found that task performance was predicted by children's language ability, but not by an autism diagnosis.

5.1 Developmental Trajectories

In terms of measurement, arguably the most important measure in understanding DD is change over time. After all, DDs may be revealed in millisecond-by-millisecond behaviour during experiments, but it may be equally important to take observations over months and years.

The Developmental Trajectory (DT) approach (Cornish et al., 2007; Karmiloff-Smith, 2009; Thomas et al., 2009) makes development the central issue, compared with the 'standardising' approach (see above) by making time *the* central consideration. In charting DD performance using standardised tests, performance of the person with DD can be derived from comparison with an age equivalent (e.g. Mental Age [MA]) control. However, the DT approach is more sophisticated because by testing participants from a range of ages one can derive a trajectory for one group of participants for a *particular* experimental task. Because the researcher simply needs an age range of participants to derive a trajectory, this method is suitable for either cross-sectional designs, longitudinal data, or a combination of cross-sectional and longitudinal. However, there are caveats with this method because the DT approach tends to use linear methods, when the relations may in fact be non-linear. Nevertheless, the DT approach does allow for useful comparisons.

Once a DT of the comparison group has been derived, the researcher can then choose to make one of three types of comparisons between the DD group and the comparison group's trajectory: 1) a theory neutral comparison, 2) to construct and compare trajectories for the DD group and the comparison group, 3) to construct trajectories using different 'time' axes (e.g. measures based on Mental Age, rather than Calendar Age).

For the theory neutral comparison a question can be asked about where a single individual in the DD group can be placed on the comparison group trajectory. This first approach is, thus, in essence, standardising the task against chronological age and asking, "where does this person with DD fit in relation to TD age norms?"

For the second type of comparison, the entire DD group is compared with the control group trajectory, and so clear links can be made between performance on the experimental task and age. In this method, the trajectories of experimental and control groups can be compared to determine if there is a delay in onset, or a slowed rate of development (see Thomas et al., 2009, for the various slopes associated with different types of trajectories). The issue here is that someone still must make a decision about where to 'anchor' the trajectories relative to each other. So, an *a priori* decision – based upon a prior theory – has already had to be made about the 'expected' outcome.

For the third type of comparison – concerning developmental aspects – a separate trajectory can be created against different age-related measures. So, in the case of autism with comorbid intellectual disability, the participant's mental age (MA) is likely to be lower than their calendar age (CA). In this case, a comparison of this type would be useful because then the trajectory for a given task against MA could be compared against the equivalent TD trajectory. This is beneficial because the researcher can see whether task performance is in consonance with the standardised measure used.

Task selection is critical because you need to have a theory about the tasks that should be affected by MA but not by the DD under investigation (see section 4.1 for further discussion). Plotting task performance against each participant's MA should 'normalise' the DD trajectory so that it 'lies on

top' of the TD trajectory. If this third kind of comparison is not made, the researcher is making an assumption that the standardised measure is a good measure for matching the participants, which it may not be. This style of DT comparison is a way of confirming, rather than assuming, that this is the case.

A useful aspect of the DT approach is data visualisation. That is, plotting performance against CA and/or MA allows the researcher to visualise the relative performances of the DD and TD groups over time. Plots like this also allow the researcher to visually examine the within-group and between-group variability, as well as to question whether the relationship between task performance and age has a linear relationship, or another kind of relationship. Work on phonological processing and reading in dyslexia by Kuppen and Goswami (2016) neatly demonstrates the power of plotting DT. By visualising task performance against CA between experimental groups these authors were able to identify, for example, which of the assessed skills were delayed (as shown by a similar gradients and poorer performance) or atypical (as shown by a group's flatter gradient and poorer performance). This allowed Kuppen and Goswami (2016) to discriminate unique abilities between children with dyslexia and age-matched low IQ poor readers (LIQPR), showing that phonological awareness – as measured by an onset oddity task – was delayed in LIQPR, but atypical in dyslexia (see Figure 1).

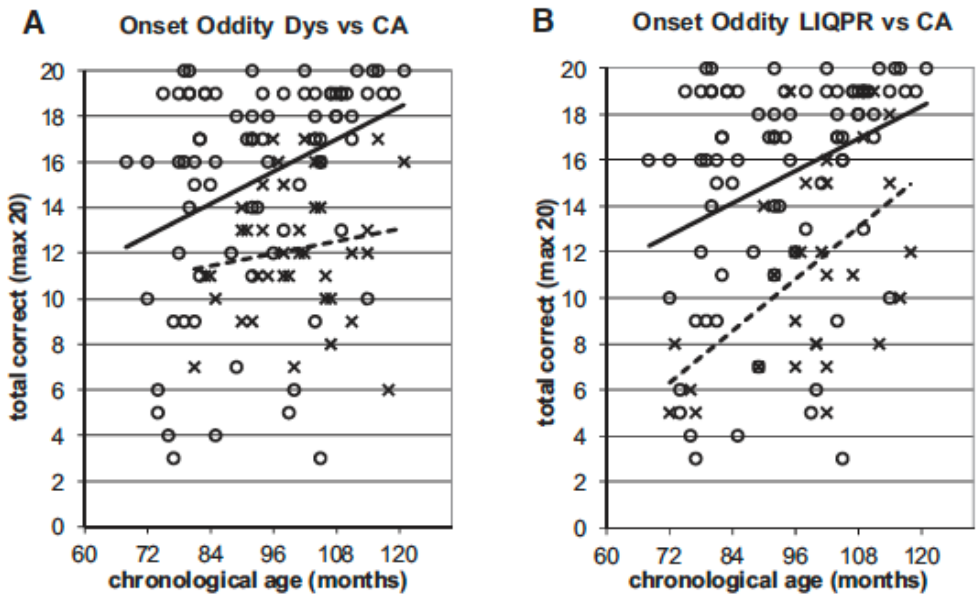

Figure 1.[5] Performance on onset oddity task against between: 1A) Children with Dyslexia (x and dotted line) vs CA matched TD (o and continuous line); 1B) Low IQ Poor Readers (x and dotted line) vs CA matched TD (o and continuous line). Plotted trajectories incorporate all data points while linear equations reflect relationship starting from youngest disorder age (Kuppen & Goswami, 2016).

An important limitation of the DT approach is that the experimental task needs to be suitable for a range of ages (and potentially for studies including both children and adults), with neither floor nor ceiling effects for the age range under investigation. This means that some tasks will not be suitable for this method, especially those tasks that have a very narrow age range window of performance (e.g. the DT approach would be inappropriate for first-order tests of false belief, because such tests are usually passed around the fourth birthday – Wellman et al., 2001).

Finally, it is important to note that the DT approach – and for that matter other methods/analyses – need not be done on their own. So, the DT method can be done *in addition* to a matching method. The DT method arguably is most useful when investigating potential developmental delays. That is, because the DT method involves tracking change over time, it is a very useful not only for detecting delays (in reaching developmental milestones), but also in saying what the exact nature of those delays might be.

Final Word on finding the right metrics

In summary, we suggest that 1) researchers investigate the *process* (the how) that someone with a DD might use to complete a task and not draw conclusions solely on the task *outcome* (the how good) and that 2) DT methods can be used to help researchers decide what they should be measuring from the outset – thereby helping them make meaningful comparisons to non-DD groups; DT plots can present these comparisons in a visually intuitive way.

[5]Permission to use kindly granted by Sarah Kuppen.

6. Conclusions

The question of what it takes to characterise a DD interleaves three questions: 1) Who you are measuring? 2) How are you measuring? and 3) What you are measuring?

6.1 Who are you measuring?

For DDs that are neurogenetic (e.g. Down's syndrome, Williams syndrome, Fragile-X syndrome etc.) it seems easier to make the case that you have a population that has been objectively defined. For DDs that are behaviourally diagnosed, however, the criteria for diagnosis are in themselves a theory of the disorder in question. So, there can be no theory-neutral behavioural diagnosis, either in terms of a diagnostic system or a specific diagnostic tool. Further, DDs that are behaviourally defined and that are strongly associated with a treatment (e.g. ADHD) might even be a risk of being defined by their treatment (Methylphenidate in the case of ADHD) than by a theory. So, the very choice of how to define your DD population has implications for what conclusion you can draw.

6.2 How are you measuring?

There is no easy answer about whom to choose as comparison participants – and this choice is often a balance between an ideal and the reality of having finite resources. Nevertheless, from the very outset, the choices you make about the composition of the comparison group stem from your theoretical or *a priori* conceptualisation of the DD in question. For example, if autism is considered a disorder in which a particular aspect of language is impaired (e.g. specific language impairment), then a [specific] language impaired comparison group is usually chosen and the various groups are matched on appropriate language-based measures. If individuals with Williams syndrome are considered to be disordered in their ability to process faces – and that this is not as a consequence of their lower than average intellectual ability (see Martens et al., 2008, for a review) – then a mental age-matched comparison group *and* a calendar age-matched comparison group might be chosen (e.g. Deruelle et al., 1999; Gagliardi et al., 2003) to tease out the potential effect of intellectual disability on face recognition. As such, we ad-

vocate the use of experimental design – through strategic recruitment of ability groups (e.g. autism, autism with language impairment, language impairment, and TD controls) – rather than statistics (like ANCOVA) to investigate characteristics of a DD.

6.3 What you are measuring?

Time is arguably the most important measurement in understanding DDs. However, trying to understand development over time in a DD has its pitfalls. For example, in using a developmental trajectory approach a comparison group still has to be matched on *a priori* theoretical assumptions about the nature of the disorder.

A potentially better way to understand DDs is to think of measurement over the duration of a cognitive process. So, standardised measures such as IQ (WASI-II, Wechsler, 2011) and memory for faces (Benton et al., 1983) measure the end points of the process – i.e. these tests assign a numerical value to a particular ability or skill. Conceivably, measurement can take place at any point in the process. For example, capturing the process by using, for example, eye tracking might indicate any deviancy in how a task is performed. So, some individuals might process faces in a feature-based way, while others do it holistically – yet both score the same in a test of face recognition. So, here a clear distinction needs to be made between process and outcome.

6.4 Summary

In summary, we argue that there is no theory-free characterisation of a DD and that this definition becomes even harder when the DD does not have a clear genetic basis. So, for behaviourally-defined DDs, this has implications for diagnosis – in that diagnosis is itself just a theory. So, it is important to acknowledge that the recruitment of both focal and comparison participants pre-supposes a theory/assumption from the outset about how you characterise that DD.

We also argue that it is important to distinguish process from outcome. So, there is more than one way of measuring performance for a given task, and that these may provide different answers. For

example, using the outcome (e.g. IQ score, face recognition score) in isolation might be misleading because the way in which an outcome is achieved might be deviant from the norm for a particular DD. Hence, the measurement tool must not only be fit for purpose, but it must also be made explicit whether the tool is measuring a process or an outcome.

Further, we argue that comparison group matching needs to be carefully thought through. A balance needs to be sought between the resources required to find appropriately matched participants *a priori* and the pitfalls of trying to statistically control for differences *post-hoc*, and on matching criteria 'p>.05' is not the same as 'these are the same'. In particular we suggest that ANCOVA should be avoided if at all possible and especially so in equating behavioural outcomes.

In sum, DD research is challenging, but the issues discussed here are a way of making informed decisions about comparison group composition, matching criteria, choice of dependent measures, choice of analyses etc. So, that researchers can make pragmatic designs decisions as well as drawing measured conclusions.

References

Alloway, T. P., Rajendran, G., & Achibald, L. M. D. (2009). Working memory in children with developmental disorders. *Journal of Learning Disabilities, 42*(4), 372-382.

Annaz, D., Karmiloff-Smith, A., Johnson, M. H., & Thomas, M. S. C. (2009). A cross-syndrome study of the development of holistic face recognition in children with autism, Down syndrome, and Williams syndrome. *Journal of Experimental Child Psychology, 102*(4), 456-486. DOI:10.1016/j.jecp.2008.11.005

American Psychiatric Association (APA). (2013). Diagnostic and statistical manual of mental disorders (5th ed.).

Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., & Charman, T. (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *Lancet, 368*(9531), 210-215.

Baron-Cohen, S. (2017). Editorial Perspective: Neurodiversity - a revolutionary concept for autism and psychiatry. *Journal of Child Psychology and Psychiatry, 58*(6), 744-747. DOI:10.1111/jcpp.12703

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the Autistic-Child Have a Theory of Mind. *Cognition, 21*(1), 37-46.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241-251.

Benton, A., Hamsher, K., Varney, N. R., & Spreen, O. (1983). *Benton Test of Facial Recognition*. New York: Oxford University Press.

British Psychological Society. (2013). British Psychological Society statement on the open letter to the DSM-5 Taskforce. Retrieved from http://www.bps.org.uk/sites/default/files/documents/pr1923_attachment_-_final_bps_statement_on_dsm-5_12-12-2011.pdf

Brock, J., Norbury, C., Einav, S., & Nation, K. (2008). Do individuals with autism process words in context? Evidence from language-mediated eye-movements. *Cognition, 108*(3), 896-904. DOI:10.1016/j.cognition.2008.06.007

Bull, M. J. (2020). Down syndrome. New England Journal of Medicine, 382(24), 2344-2352.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd Ed.). Hillside, NJ: Earlbaum.

Cornish, K., Scerif, G., & Karmiloff-Smith, A. (2007). Tracing syndrome-specific trajectories of attention across the lifespan. *Cortex, 43*(6), 672-685.

Cornish, K., Turk, J., & Hagerman, R. (2008). The fragile X continuum: new advances and perspectives. *Journal of Intellectual Disability Research, 52*, 469-482. DOI:10.1111/j.1365-2788.2008.01056.x.

Dawson, M., Soulieres, I., Gernsbacher, M. A., & Mottron, L. (2007). The level and nature of autistic intelligence. *Psychological Science, 18*(8), 657-662.

Deruelle, C., Mancini, J., Livet, M. O., Casse-Perrot, C., & de Schonen, S. (1999). Configural and local processing of faces in children with Williams syndrome. *Brain and Cognition, 41*(3), 276-298. DOI:10.1006/brcg.1999.1127.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabularly Test* (3rd Ed.). Circle Pines, MN: American Guidance Service.

Elsabbagh, M., Divan, G., Koh, Y.-J., Kim, Y. S., Kauchali, S., Marcin, C., . . . Fombonne, E. (2012). Global Prevalence of Autism and Other Pervasive Developmental Disorders. *Autism Research, 5*(3), 160-179. DOI:10.1002/aur.239

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23*(1), 132-138. DOI:10.3758/bf03210562.

Gagliardi, C., Frigerio, E., Burt, D. M., Cazzaniga, I., Perrett, D. I., & Borgatti, R. (2003). Facial expression recognition in Williams syndrome. *Neuropsychologia, 41*(6), 733-738. DOI:10.1016/s0028-3932(02)00178-1

Geogiades, S., Szatmari, P., & Boyle, M. (2013). Importance of studying heterogeneity in autism. *Neuropsychiatry, 3*(2), 123-125.

Happé, F. (1999). Autism: cognitive deficit or cognitive style? *Trends in Cognitive Sciences, 3*(6), 216-222.

Happé, F. G. E. (1994). An Advanced Test of Theory of Mind - Understanding of Story Characters Thoughts and Feelings by Able Autistic, Mentally- Handicapped, and Normal-Children and Adults. *Journal of Autism and Developmental Disorders, 24*(2), 129-154.

Hermelin, B., & O'Connor, N. (1967). Remembering of words by psychotic and subnormal children. *British Journal of Psychology, 58*, 213-218.

Insel, T. (2013). Transforming Diagnosis. Retrieved from http://www.nimh.nih.gov/about/director/2013/transforming-diagnosis.shtml

Jarrold, C., & Brock, J. (2004). To match or not to match? Methodological issues in autism-related research. *Journal of Autism and Developmental Disorders, 34*(1), 81-86.

Järvinen-Pasley, A., Bellugi, U., Reilly, J., Mills, D. L., Galaburda, A., Reiss, A. L., & Korenberg, J. R. (2008). Defining the social phenotype in Williams syndrome: A model for linking gene, the brain, and behavior. *Development and Psychopathology, 20*, 1-35. DOI:10.1017/s0954579408000011|issn 0954-5794.

Kaland, N., Callesen, K., Moller-Nielsen, A., Mortensen, E. L., & Smith, L. (2008). Performance of children and adolescents with Asperger syndrome or high-functioning autism on advanced theory of mind tasks. *Journal of Autism and Developmental Disorders, 38*(6), 1112-1123. DOI:10.1007/s10803-007-0496-8

Karmiloff-Smith, A. (2009). Nativism Versus Neuroconstructivism: Rethinking the Study of Developmental Disorders. *Developmental Psychology, 45*(1), 56-63. DOI:10.1037/a0014506

Karmiloff-Smith, A. (2012). Perspectives on the dynamic development of cognitive capacities: insights from Williams syndrome. *Current Opinion in Neurology, 25*(2), 106-111. DOI:10.1097/WCO.0b013e3283518130

Kuppen, S. E., & Goswami, U. (2016). Developmental trajectories for children with dyslexia and low IQ poor readers. *Developmental Psychology, 52*(5), 717.

Leonard, H. C., Annaz, D., Karmiloff-Smith, A., & Johnson, M. H. (2011). Brief Report: Developing Spatial Frequency Biases for Face Recognition in Autism and Williams Syndrome. *Journal of Autism and Developmental Disorders, 41*(7), 968-973. DOI:10.1007/s10803-010-1115-7

Leonardo, E. D., & Hen, R. (2008). Anxiety as a developmental disorder. *Neuropsychopharmacology, 33*(1), 134-140. DOI:10.1038/sj.npp.1301569

Martens, M. A., Wilson, S. J., & Reutens, D. C. (2008). Research Review: Williams syndrome: a critical review of the cognitive, behavioral, and neuroanatomical phenotype. *Journal of Child Psychology and Psychiatry, 49*(6), 576-608. DOI:10.1111/j.1469-7610.2008.01887.x

Matson, J. L., Fodstad, J. C., & Boisjoli, J. A. (2008). Nosology and diagnosis of Rett syndrome. Research in Autism Spectrum Disorders, 2(4), 601-611.

Mervis, C. B., & Klein-Tasman, B. P. (2004). Methodological issues in group-matching designs: alpha levels for control variable comparisons and measurement characteristics of control and target variables. *Journal of Autism and Developmental Disorders, 34*(1), 7-17. DOI:10.1023/B:JADD.0000018069.69562.b8

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110*(1), 40-48.

Miller, J. N., & Ozonoff, S. (1997). Did Asperger's cases have Asperger Disorder? A research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 38*(2), 247-251.

Milton, D. E. M. (2012) On the ontological status of autism: the 'double empathy problem', *Disability & Society, 27*(6), 883-887, DOI: 10.1080/09687599.2012.710008

Norbury, C. F. (2005). Barking up the wrong tree? Lexical ambiguity resolution in children with language impairments and autistic spectrum disorders. *Journal of Experimental Child Psychology, 90*(2), 142-171.

Norbury, C. F. (2014). Sources of variation in developmental language disorders: evidence from eye-tracking studies of sentence production. *Philosophical Transactions of the Royal Society B-Biological Sciences, 369*(1634). DOI:10.1098/rstb.2012.0393

Norbury, C. F., Brock, J., Cragg, L., Einav, S., Griffiths, H., & Nation, K. (2009). Eye-movement patterns are associated with communicative competence in autistic spectrum disorders. *Journal of Child Psychology and Psychiatry, 50*(7), 834-842. DOI:10.1111/j.1469-7610.2009.02073.x

Pennington, B. F., Snyder, K. A., & Roberts Jr, R. J. (2007). Developmental cognitive neuroscience: Origins, issues, and prospects. *Developmental Review, 27*, 428-441.

Perez-Osorio, J., Wiese, E., & Wykowska, A. (2021). *Theory of Mind and Joint Attention. In The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual*

Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition (pp. 311-348).

Peterson, C. C., & Wellman, H. M. (2019). Longitudinal theory of mind (ToM) development from preschool to adolescence with and without ToM Delay. *Child development, 90(6)*, DOI: 10.1111/cdev.13064.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Section 3. Standard progressive matrices.* Oxford, England: Oxford Psychologists Press.

Riby, D. M., Doherty-Sneddon, G., & Bruce, V. (2008). Atypical unfamiliar face processing in Williams syndrome: what can it tell us about typical familiarity effects? *Cognitive neuropsychiatry, 13*(1), 47-58. DOI:10.1080/13546800701779206

Robinson, E. B., St Pourcain, B., Anttila, V., Kosmicki, J. A., Bulik-Sullivan, B., Grove, J., . . . Daly, M. J. (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat Genet, 48*(5), 552-555. DOI:10.1038/ng.3529, URL: http://www.nature.com/ng/journal/v48/n5/abs/ng.3529.html#supplementary-information

Rogers, S. J., Young, G. S., Cook, I., Giolzetti, A., & Ozonoff, S. (2008). Deferred and immediate imitation in regressive and early onset autism. *Journal of Child Psychology and Psychiatry, 49*(4), 449-457. DOI:10.1111/j.1469-7610.2007.01866.x.

Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: a GLM approach.* London: Sage.

Saxton, M. (2010). *Child Language: Aquistion and Development*. London: Sage.

Semel, E., Wiig, E. H., & Secord, W. A. (2004). *Clinical evaluations of language fundamentals* (4th ed.). San Antonio, TX: Harcourt Assessment.

Shah, A., & Frith, U. (1993). Why Do Autistic Individuals Show Superior Performance on the Block Design Task. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 34*(8), 1351-1364.

Szatmari, P., Bryson, S., Duku, E., Vaccarella, L., Zwaigenbaum, L., Bennett, T., & Boyle, M. H. (2009). Similar developmental trajectories in autism and Asperger syndrome: from early childhood to adolescence. *Journal of Child Psychology and Psychiatry, 50*(12), 1459-1467. DOI:10.1111/j.1469-

7610.2009.02123.x

Tager-Flusberg, H., Plesa-Skwerer, D., Faja, S., & Joseph, R. M. (2003). People with Williams syndrome process faces holistically. *Cognition, 89*(1), 11-24. DOI:10.1016/s0010-0277(03)00049-0

Tanenhaus, M. K., Spiveyknowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science, 268*(5217), 1632-1634. DOI:10.1126/science.7777863

Thomas, M. S. C., Annaz, D., Ansari, D., Scerif, G., Jarrold, C., & Karmiloff-Smith, A. (2009). Using Developmental Trajectories to Understand Developmental Disorders. *Journal of Speech Language and Hearing Research, 52*(2), 336-358. DOI:10.1044/1092-4388(2009/07-0144).

Vickers, A. J., & Altman, D. G. (2001). Statistics notes - Analysing controlled trials with baseline and follow up measurements. *British Medical Journal, 323*(7321), 1123-1124.

Waterhouse, L., & Gillberg, C. (2014). Why Autism Must be Taken Apart. *Journal of Autism and Developmental Disorders, 44*(7), 1788-1792. DOI:10.1007/s10803-013-2030-5

Waterhouse, L., London, E., & Gillberg, C. (2016). ASD Validity. *Review Journal of Autism and Developmental Disorders, 3*(4), 302-329. DOI:10.1007/s40489-016-0085-x

Waterhouse, L., London, E., & Gillberg, C. (2017). The ASD Diagnosis has Blocked the Discovery of Valid Biological Variation in Neurodevelopmental Social Impairment. *Autism Research, 10*(7), 1182-1182. DOI:10.1002/aur.1832

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children–third edition: Canadian (WISC-III)*. Toronto, Ontario, Canada: Psychological Corp.

Wechsler, D. (2011). *Wechsler Abreviated Scale of Intelligence - Second Edition (WASI-II)*: Pearson.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655-684.

World Health Organization. (2019). International statistical classification of diseases and related health problems (11th ed.)

Wiig, E. H., Secord, W., & Semel, E. (1992). Clinical Evaluation of Language Fundamentals–Preschool San Antonio, TX: The Psychological Corporation/Harcourt Brace.

Wing, L., & Gould, J. (1979). Severe impairments of social interaction and associated abnormalities in children: epidemiology and classification. *Journal of Autism and Developmental Disorders, 9*, 11-29.

Zelazo, P. D., Burack, J. A., Benedetto, E., & Frye, D. (1996). Theory of Mind and rule use in individuals with Down's Syndrome: A test of the uniqueness and specificity claims. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 37*(4), 479-484.

Zwicker, J. G., Missiuna, C., Harris, S. R., & Boyd, L. A. (2011). Brain activation associated with motor skill practice in children with developmental coordination disorder: an fMRI study. *International Journal of Developmental Neuroscience, 29*(2), 145-152.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: