Purser, H. & **Van Herwegen, J**. (2016). Standardised and experimental psychological tasks: issues and solutions for research with children. In: J. Prior. & J. Van Herwegen (Eds.). *Practical Research with Children*. London: Psychology Press.

Chapter 4: Standardised and experimental psychological tasks: issues and solutions for research with children

Harry Purser & Jo Van Herwegen

4.1 Introduction

What children do and say provides us with important information about their development and wellbeing. However, there is only an indirect relationship between the former and the latter so that observing children alone cannot help us understand the underlying mechanisms. Experimental designs are there to explore specific hypotheses and examine the cause and effect relationship of different phenomena. This chapter will focus on how behavioural, experimental, and standardised psychological tasks are used within experimental designs. In the first section, we will discuss the use of standardised tasks, focusing on selecting the appropriate task, issues that relate to the administration of the task, and how to correctly interpret the scores obtained from standardised tests. In the second part of the chapter, we will focus on experimental tasks. More specifically, we will discuss the advantages and disadvantages of designing your own experimental tasks and what confounding factors should be taken into account during the design stage. Next, the issue of validity as well as general issues when using tasks with children will be explored. Throughout the chapter, we will use examples from research with children who have neurodevelopmental disorders, not only because it includes our own research, which we know best, but more importantly, because working with children with additional needs reinforces certain issues when using experimental and standardised tasks.

The term ‘test’ is often used within psychology to refer to any type of assessment that participants complete, whether on paper or in real life. However, one of the difficulties for psychology as a field is that, in contrast to other sciences, the objects of study, psychological constructs such as ‘theory of mind’, are not directly observable, compared to, for example, temperature or weight. Therefore, it is better to refer to psychological assessments as ‘tasks’ or ‘measurements’ rather than ‘tests’, because often within psychology we cannot directly test the variables of interest. In addition, the term ‘test’ carries the connotation that there is a correct or incorrect answer and this may increase anxiety in children resulting in unreliable assessments. Therefore, for the remainder of this chapter we will use the word ‘task’ instead.

4.2 Using Standardised tasks

4.2.1 Different types of standardised tasks

Standardised tasks are tasks that have been administered in a specific manner to a large group of people so that normed data has been obtained and the participant’s score can be compared to that normed population when the standardised administration is employed. There are standardised tasks for all areas of psychology, including areas such as general intelligence, attention, memory, emotions, behaviours (actions) as well as all areas of achievement (mathematics, reading, language, etc.). Although both achievement and aptitude tasks are standardised, aptitude tasks are intended to predict the participant’s future performance (such as intelligence tasks), whereas achievement tasks assess what a person has mastered or learned (such as the number of words one knows).

Standardised tasks are used with children for a variety of reasons. For example, standardised tasks are often used by educational and clinical psychologists often use standardised tasks to evaluate whether a specific child has special needs or qualifies for specialist education as well as to assess the child's mental or physical health. Within research contexts, standardised tasks are employed to evaluate whether training programmes and interventions have made an impact on a child's development or to match different groups. For example, participants with Williams syndrome (WS), a rare disorder that is caused by a genetic deletion on the long arm of chromosome seven (see Martens, Wilson, & Reuters, 2008 for a discussion), have overall learning difficulties with general IQ scores within the mild to moderate impairment range (50-70). This means that they rarely perform in line with their chronological age. Thus, within a research context their performance on experimental tasks (see below) is compared to children who are matched for mental age, i.e., children who have similar scores on a particular standardised task. For example, in our own research we have matched infants with WS to those with Down syndrome based on performance on the Bayley Scales of Infant and Toddler Development (Bayley, 1993; Karmiloff-Smith et al., 2012; Van Herwegen, Ansari, Xu, & Karmiloff-Smith, 2008). Within forensic psychology, standardised tasks are often used within child custody evaluations to assess mental health or parent-child relationships (for a discussion see Gould, 2005).

Examples of standardised aptitude tasks include intelligence tasks such as Wechsler's Intelligence Scales, which include scales for infants (WPPSI-III: Wechsler, 2002), children (WISC-IV: Wechsler, 2003), adults (WAIS: Wechsler, 1997), and an abbreviated scale that can be used with children as well as adults (WASI: Wechsler, 1999). These scales are based on the view that intelligence is made up of a number of different abilities and thus they include several sub-tasks that measure different aspects of intelligence, including verbal and performance IQ. These tasks are administered on an individual basis in a quiet room, which means that they are time-consuming and therefore costly. On the other hand, the examiner has the opportunity to carefully observe what the child is doing or where the child is looking. This is very important as children might fail to answer correctly for a number of reasons. For example, children with language difficulties might find it difficult to understand the instructions. Children with developmental disorders also often show attention difficulties and need to be brought back to task by the examiner. Finally, some of these tasks require the child to choose the correct answer from a number of different options. However, children who do not carefully look at the different options might fail the task, not because they do not have the aptitude or ability to do so, but because they just randomly chose an option. Although one has to follow the standardised procedure for scoring the task, individual administration of the tasks allows the examiner to take these possibilities into account when interpreting the child's performance.

Other aptitude tasks only include one type of assessment and thus, are easier to administer and score. In the Raven's Progressive Matrices (Raven, Raven, & Court, 2004), for example, children are shown a visual pattern that is missing a piece and they are asked to identify the missing piece from 6 options. There are different versions of the task available for use with different ages and groups, including Raven's Standard Progressive Matrices (RSPM; for use with the general population), Raven's Coloured Progressive Matrices (RCPM; for children and adults with special needs as well as elderly people and typically developing children aged 5 to 11 years old), and Raven's Advanced Progressive Matrices (RAPM; for adults and adolescents of above-average intelligence or gifted children). These tasks can be administered individually or as a group. In addition, it has been argued that this task is less culturally loaded, compared to some other aptitude tasks, as it contains minimal instructions and does not require any linguistic responses from the participant. Furthermore, the fact that this task can be administered to groups and scored automatically means that information from large numbers of children can be obtained quickly and relatively cheaply.

4.2.2 Choosing the appropriate standardised task

In order to obtain a score that reflects the participant's true abilities, it is important to choose the appropriate standardised task. The choice of which task to use might be based on the advantages and disadvantages of administering individual or group tasks. In addition, one has to take into account cultural and linguistic aspects. For example, the Wechsler's tasks are American: even the most able children in other countries might not know certain words or concepts. This is particularly problematic for the verbal sub-scales in which the child might be asked to provide a definition of a certain word or person that might be unknown to them. For this reason, in the UK the British Ability Scales III (BAS; Elliot & Smith, 2011) are often preferred instead because this task includes similar verbal and non-verbal subtasks to the Wechsler's scales, but it only includes concepts and words that British children are familiar with. Yet, in other countries there might not be an appropriate aptitude task available and one may have to use tasks that have been normed in the US or UK. In addition, when choosing a task one also has to consider the type of response that is required from the child. Some tasks require a timed response because it can be argued that intelligence is not only about finding the correct answer but also how fast one can find a solution. Such timed tasks might put children with motor difficulties and impairments at a disadvantage (for example, children with low attention span).

Another important factor when choosing what standardised task to use is the age of the participants, as normed data will only be available for a limited age-range. This is rather straightforward when working with TD groups. However, when working with atypical groups this becomes a more tricky issue as participant groups will often show uneven cognitive profiles with better performance on certain abilities compared to others. For example, children with Autism Spectrum Disorders (ASD) often perform better on non-verbal tasks that rely on visuo-spatial abilities compared to tasks that tap into verbal abilities. Therefore, when assessing children with ASD on BAS subscales, they may perform at floor (i.e., the lowest age-equivalent norm available) on certain verbal tasks while being at ceiling for nonverbal tasks (i.e., highest age-equivalent norm available). This means that no appropriate IQ or age equivalent score can be calculated and that the discrepancy between children's chronological age and mental age could have been larger should age equivalent norms from younger and older children have been available. Therefore, floor and ceiling effects should be avoided and it is important to use tasks that have been standardised with children from a wide range of ages. For example, the WASI scales can be administered with participants aged 6 to 89 years old.

4.2.3 Administration of standardised tasks

One important aspect of standardised tasks is that, in order to compare the score of a particular participant to the normed data, specific instructions have to be followed and thus only a trained psychologist should administer the task. The fact that instructions are standardised means that different people can assess a child over the child's development and that these performance scores can still be compared. However, in our experience a rigid set of instructions does not automatically imply that testing outcomes are objective. For example, a six-and-a-half-year-old boy with severe language difficulties who took part in our research was assessed on his non-verbal abilities in order to confirm whether or not he met the inclusion criteria of specific language impairment for the study. He obtained a raw score of 16 on RCPM, which put him in the 40th percentile for his age range. However, it transpired afterwards that when his educational psychologist had assessed him a few weeks before, he had only obtained a score within the 1st percentile for children of his age. Although one should not repeat standardised tasks within such a short time frame (usually not within 6 months) to avoid learning effects, it is unlikely that this difference in scores could be attributed to a learning effect or development between time 1 and time 2 alone. Through discussion with the parents and the educational psychologist, it was revealed that while the educational psychologist used a "purist" view where she did not provide feedback, the experimenter in our research study had provided general motivational feedback ("keep going", "good job") during the testing session. This suggests that motivation plays an important role when assessing children.

In our own experience in working with children with neurodevelopmental disorders, motivation is an important aspect of the assessment session and score. However, there is a debate as to whether praise should be used when assessing children. On the one hand, it has been shown that random reinforcement can have the opposite effect and decrease the motivation to respond (see Eisenberger & Cameron, 1998). In addition, one cannot ensure that different people will respond with the same motivation from reinforcement, which is why most manuals will verify what kind of feedback can be provided. On the other hand, children are used to receiving a lot of motivational feedback when performing tasks and might find it stressful when an examiner sits silently next to them. Furthermore, a review study that evaluated the impact of incentives, including praise, candy, money social reinforcement and tokens, did not find any consistent differences between studies that did and did not offer incentives (Sattler, 1988). While the debate continues, we recommend for now that researchers use motivational feedback because, other things being equal, this will be less stressful for children. It is also important that researchers remain consistent in their own testing and document the particular decisions that they made. In addition to what kind of motivational feedback is provided, one needs to take into account the mood and fatigue of the child, the length of the task to be administered, the time and day of the assessment, and the relationship between the examiner and examinee. Thus, the interpretation of the scores requires a thorough understanding of the task, the task-taker, and the conditions of the assessment session. (See also Chapters 2 and 3 for further consideration of how the conditions of the assessment session can be influential in eye tracking and brain imaging studies with children).

Due to the fact that standardised tasks require test administrators to adhere to the instructions, as well as the type of feedback that can or cannot be provided, some test makers have started to develop computer-assisted tasks. For example, the New Group Reading Test Digital (2010), a task that assesses reading ability (phonology, word reading, and passage comprehension) in children aged 7 to 16 years old, is an adaptive online achievement task that starts with the child's chronological age but is based upon the child's performance. The digital test is able to adapt the reading material during the session in order to reflect the child's actual reading ability. The fact that the task is administered on a computer means that the procedure is completely standardised, no task administrator is required, scoring happens online and automatically rather than off-line, and there is no experimenter bias involved. In addition, computerised testing also reduces the number of experimenter errors such as skipping items, not demonstrating certain items, or not accurately assessing the basal and ceiling item for a particular child.

Standardised tasks are often used in research to compare or match groups, especially in research that includes atypical groups. In order to be able to match two groups, it is important to establish that both groups approach the task in a similar way because it is possible that two groups will obtain the same score through different strategies, so that similar scores might not reflect similar levels of underlying ability. For example, research has shown that children with Down syndrome (DS) make different types of errors on RCPM compared to typically developing (TD) children (Gunn & Jarrold, 2004). This means that even when you select younger TD children who are matched for the total number of errors the DS children make, you cannot assume that the groups are actually matched for non-verbal performance as the groups might rely on atypical strategies. In contrast, children with WS have shown that they do make the same type or errors as younger TD children and you can meaningfully match the two groups for RCPM scores and ability (Van Herwegen, Farran, & Annaz, 2011). Thus, a thorough understanding of the task itself is required before a child's performance can be scored accurately.

One serious drawback related to standardised tasks is the fact that one should not usually test the same child twice within a short time span (often within 6 months) in order to avoid learning effects. This provides some constraints when evaluating an intervention programme where one might want to use standardised tasks to compare performance before and after a particular intervention. There are currently a number of standardised tasks that provide different lists of stimuli for the same subtasks. As both lists are standardised on the same population, one set of stimuli can be used pre-intervention and the other one at the end of the intervention programme. For example, the Early Numeracy Test (ENT; Van Luit, Van de Rijt, & Pennings, 1994) consists of 40 items that evaluate different aspects of young children's numerical competence. It has two analogous versions – version A and version B – so that a different set can be used at two different time points (see for example Aunio & Niemivirta, 2010).

4.2.4 Scoring standardised tasks

Once the administration of the task is completed, there are different scores that can be calculated from standardised tasks. In the first instance, one can calculate the raw score of a certain task, which is the amount of correct answers a child has given. Raw scores have the disadvantage that one cannot always directly compare scores, as children of different ages often have different starting points and different tasks might have a different total amount of items that have been administered. Therefore, most tasks will provide age equivalent norms or performance norms based on the standardised norms. Performance norms can include statistical metrics such as cumulative percentages, percentiles, *z*-scores, *t*-scores, and IQ scores. The norms of a test are based upon the distribution of the scores of the people in the normed sample. It is beyond the scope of this chapter to discuss the differences between all these different scores but there are a few pointers one has to keep in mind when interpreting them.

First of all, one has to question how each norm has been obtained and who was included within the normed sample. Often, you will find that ethnic minorities are not represented within normed samples and thus, if one has tested a child from an ethnic minority and this child obtained a lower score, it is unclear whether this lower score is caused by some cultural differences or a true representation of that child's abilities. This is called 'mismatched norming'. In addition, there are very few standardised tasks that assess all relevant cognitive abilities within one type of task, so that often a battery of standardised tasks has to be employed in order to obtain a well-rounded psychological assessment of a child. However, different standardised tasks will be normed on different groups across different countries and cultures, which might make comparison between standardised tasks difficult (see Brock, Jarrold, Farran, Laws, & Riby 2007, for a discussion).

Secondly, one needs to consider the appropriateness of each score. Standardised tasks that contain multiple sub-tasks often provide an overall score as well as scales that combine scores across different sub-tasks. For example, the Wechsler's scales provide a full scale IQ score (FSIQ) as well as a verbal (VIQ) and performance intelligence quotient (PIQ). However, one has to be careful interpreting these scale scores in atypical groups that show uneven cognitive profiles. One such example is WS: although participants with WS have FSIQs between 50-70 on average, their VIQs are often better and develop at a faster rate compared to their PIQs. Therefore, FSIQs do not provide an accurate representation about their verbal or non-verbal abilities (Jarrold, Baddeley, & Hewes, 1998). In addition, even TD children have been found to show a discrepancy between their verbal and non-verbal abilities (Brock et al., 2007).

Thirdly, although as described above raw scores do not allow comparisons between tasks or different age groups, when working with particular groups normed scores might simply not be available. For example, in the BAS pattern construction sub-task, children are asked to copy a pattern from a book using either foam squares or 3-dimensional plastic cubes for the more advanced forms. Participants obtain a score based on the time it takes them to complete an item as well as the total number of items completed. Participants with WS are particularly weak on this task and, often, normed scores are not available for their low performance. Thus, any ability score lower than 65 renders an age-equivalent score of 3-years-old. In such cases no accurate comparisons can be made between groups and thus it might be better to use ability or raw scores (see Van Herwegen, Rundblad, Davelaar, & Annaz, 2011).

In sum, standardised tasks have the advantage that they allow performance scores of different children, or scores on different tasks, to be directly compared to one another so that one can assess how a child's performance score compares to the norm. However, the fact that normed scores are obtained from a specific sample from the population has its drawbacks when the tasks are used with ethnic minority or specific atypical groups. In addition, the fact that the content of standardised tasks is set in stone only allows the researcher access to a limited amount of information.

4.3 Experimental tasks

4.3.1 Using experimental tasks

As standardised tasks only allow you to answer questions that are tested within the scope of the task, researchers often develop their own experimental tasks in which they can carefully manipulate a particular variable of interest (i.e., the experimental variable) and control for confounding variables. For example, although the Early Number Concept task from the British Abilities Scales (Elliot & Smith, 2011) allows one to assess children's early mathematical abilities across a standardised population, it includes a wide variety of mathematical abilities, including children's understanding of cardinality (i.e., the understanding of counting), digit recognition knowledge, comparison of sets, and understanding of mathematical concepts such as 'more'. Therefore, this task is too general to allow one to examine whether specific populations have issues with a particular mathematical concept only, say for example the understanding of counting.

Yet, although experimental tasks are often designed specifically to answer a specific question, some experimental tasks have been used so often with children that, even though they are not standardised, they are considered good measures to answer that specific question. For example, the Sally-Ann task (Baron-Cohen, Leslie, & Frith, 1985) is a commonly used task to assess young children's theory of mind abilities. Therefore, it is always good to check the literature first to see what tasks are commonly used within the field before starting to design your own tasks as you may well find an established task that can help you answer your question.

The aim of matching on standardised tests or mental age measures leads the researcher to indirectly equate the participant groups on their ability to perform the non-central aspects of the task. However, 'task-matching' is an alternative that achieves this equating of ability level directly. Task matching is the use of a control condition that is the same as the experimental task differing only in the information required at test. For example, if the hypothesis is that Group X are more susceptible to background noise in memory tasks than Group Y, the groups could be matched for performance on the noiseless condition of the memory task. Performance on this control condition (noiseless) can then be used to match performance of the different groups before comparing performance on the experimental condition (with background noise). It is important, however, to ensure that the control task is sensitive for the groups being matched, i.e., that the lack of any group difference is not due to floor or ceiling effects (see Jarrold & Brock, 2004). Examples abound of control conditions that fulfil the criteria above in studies that match groups on standardised tests or mental-age measures, e.g., in studies that contrast social and non-social tasks in ASD research (e.g., Klin, 2000). However, it is hard to find examples of group matching at the outset of a study for participants' performance on the task-matched control condition, which would obviate any need for matching on these additional tasks (see Phillips, Jarrold, Baddeley, Grant, & Karmiloff-Smith, 2004).

4.3.2 Designing experimental tasks

Designing your own experimental task requires careful consideration of confounds at the design stage and one has to ensure that all confounding variables have been controlled for, so that any differences between groups or conditions are caused by differences of the variable in question. Variables that are particularly important when working with children include effects caused by limited motivation, administration of the task, working memory demands as well as the verbal and motor demands of a task. We will now turn to each of these issues in more detail.

Administration of the task

Before one starts the design of an experimental task, one should consider how the task will be administered and how the responses will be recorded. For example, children with ASD perform better on traditional executive control tasks that are administered on a computer, in contrast to face-to-face tasks with an experimenter, as computerised tasks require fewer social interactions (see Kenworthy, Yerys, Anthony, & Wallace, 2008 for a discussion). However, a recent study that has directly compared computer - as well as experimenter-administered versions of executive functioning (EF) tasks in the same children with ASD has shown that there was no difference, and that problems with EF tasks observed in ASD cannot be attributed to a limited ability to engage with an experimenter or the extra social demands of such tasks (Williams & Jarrold, 2013).

Computerised tasks have the advantage that the administration can be standardised across different participants and that performance can be scored automatically. However, administering a task on computer can make it difficult for the experimenter to monitor the child's motivation carefully as the task progresses. Therefore, it is easier for the experimenter to encourage the child when enthusiasm wanes when the task is not administered on a computer. One solution could be to include regular breaks to keep the child motivated as well as giving the child a token or a reward picture after a certain number of trials have been completed. In addition, one should statistically compare performance at the end of the task to that of the start of the task or counter-balance stimulus lists across participants to ensure that motivation alone cannot explain the observed results.

Most computerised experimental tasks require children to press a certain key or button on a response pad or keyboard when viewing particular stimuli. This requires a number of abilities. First of all, the child needs to have appropriate motor skills to press that particular button or response key as well as good eye-hand coordination. More importantly, the child needs to keep in mind which button to press for a particular answer. For example, a child might be required to press the yellow button for when the stimuli on the left-hand side of the screen is correct and a red button when the answer on the right-hand side is correct (see Halberda & Feigenson, 2008 for an example). This means that the child needs to keep in mind the correct answer as well as which button to press. This can be very taxing for children's limited working memory abilities. Therefore, it might be better to use a touch screen so that children can respond by touching the correct answer on the screen itself, instead of having to link stimuli on the screen to a particular button on a response pad or keyboard. A related point is that unless memory is being assessed, it is always better to display any task-relevant stimuli until the response has been made, rather than altering the display to a 'response screen' that features only the possible responses. For example, a receptive vocabulary task might present an image for naming. If the image disappears before participants have responded, an incorrect answer might owe, at least in part, to forgetting aspects of the image, rather than to a lack of lexico-semantic knowledge.

Another alternative to response buttons might be to ask children to act out a response or to verbally provide an answer. Yet again, these methodologies have their own disadvantages in that young children have a limited vocabulary and grammar and might not be able to explain what they are thinking. Furthermore, verbal responses might disadvantage certain children with neurodevelopmental disorders. For example, children with language difficulties, but also very young children, might not elaborate as much as TD controls on their answers due to their limited language abilities, yet they may provide the correct answer when a non-verbal response is required.

Instructions of the task

Not only does one need to be careful with the type of response that is required from the children, but one also needs to carefully consider the instructions that the child is given. Ideally, instructions should be fairly short, and simple grammatical sentences should be used when working with children. In addition, one may want to consider adding a short task before the experimental tasks that assesses the child's understanding of the instructions or train the child on the type of response that is required from them. For example, it has been argued that a child's approximate number abilities are related to their number word knowledge (Mussolin, Nys, Leybaert, & Content, 2012; Wagner & Johnson, 2011). Children's ANS abilities are often assessed by a task in which children are presented with two amounts of dots on a screen and they have to indicate whether there are more dots on the left or right-hand side of the screen. Recent research has shown that when only children who passed a training task that shows they understand the meaning of the word 'more' are included in the analyses, the correlation between ANS and exact number word knowledge disappears (Negen & Sarnecka, 2015).

Another example that shows that task instructions are important comes from our own studies in which we examined figurative language comprehension (including metaphors such as 'my teacher is a dragon') in children with WS. In these studies children listen to short stories that end with a figurative expression and children are asked to indicate what the speaker means by selecting the correct picture out of three options. This question is then followed by a memory question that asks about a fact that was mentioned in the story. Any child who fails the memory question is then removed from the final data to ensure that who fail the question about the figurative expression do so because they do not understand the expression and not because they were not paying attention to the story (Rundblad, Dimitirou, & Van Herwegen, under review; Van Herwegen, Dimitriou, & Rundblad, 2013).

Motivation

As we described before, motivation is an important aspect of children's performance. Therefore, it is important to make the task relevant or interesting to children. For example, providing stimuli as part of a narrative might make the assessment more game-like and pleasant for children, which might impact positively on their performance. Instead of simply administering a digit span task, where the experimenter reads out a list of numbers to be repeated back by the participant in correct serial order, there could be a narrative wherein a 'secret agent' is imprisoned by a master criminal: the participant must listen to the secret codes and relay them to mission control, in order to break the security system and release the agent (as used in Purser et al., 2012). It is possible that even simply framing tasks as 'games' when describing them to participants, rather than only describing what the task requires, would spark interest and engagement.

Reaction times

Finally, tasks are often administered on a computer in order to obtain reaction time (RT) measures from children as RTs can provide richer data than just the number of correct answers. For example, it has been argued that difficulties with making inferences about mental states in ASD would not only be reflected in number of incorrect responses but also in the time it takes people with ASD to make these inferences (Bowler, 1997). Indeed, adolescents with Asperger syndrome do take longer to process both mental states and physical state inferences in a story context. Yet, there is a large difference in the time it takes them to process mental states versus physical states which shows that, although they have general difficulties with making inferences about stories they have been told, they find inferences about mental states especially challenging (Kaland, Smith, & Mortensen, 2007).

However, one needs to be careful when interpreting RT data from children in that RTs in children are very variable. As noted by Lange-Kuttner (2012), variability in children's reaction time data may be so large that it is not possible to detect differences between group means: within-group differences will be greater than between-group differences. This variability may be especially marked when children enter school (Kail & Ferrer, 2007), with large individual differences eventually diminishing because of the leveling effects of schooling. Unusually large variability in reaction time data seems to be associated with particular developmental disorders, such as attention deficit and hyperactivity disorder (ADHD). For example, Epstein et al. (2011) found elevated variability of reaction times in children with ADHD relative to TD controls, across five neuropsychological tasks that tapped attention and executive function abilities. These effects were neither modulated by task, nor by ADHD subtype: irrespective of the task, children with ADHD demonstrated irregular patterns of reaction times, featuring occasional long times. The authors pointed out that although these infrequent long reaction times might reflect attentional lapses (e.g., Hervey et al., 2006), the reason for these events is currently unknown.

4.3.3 Piloting and validation of new experimental tasks

Once the experimental task is set up it is important to take some time and pilot the task with some participants. This is important for all experimental tasks but especially when working with children as all adults, including experimenters who are experienced in working with children, have certain expectations of children's performance and behaviour that might be incorrect.

In addition, one needs to evaluate the validity of the task, which can be broadly defined as the ability of the task to measure the characteristics that it is designed to assess. Although the notion of validity is often subdivided into different types, they are unified in the sense that they form part of the overall justification for a particular test's application and interpretation. Messick (1980) highlighted the risk of thinking about these aspects of validity as subtypes, such as population validity, because this view might lead researchers to treat any single one of these, or even a small group, as the whole of validity. Messick suggested that if, instead, one were to use the more descriptive term ‘population generalizability’, it would guard against this confusion and emphasise that validity consists of various conditions that must be satisfied, each of which will be of differing importance, depending on the test in question.

In developing a new task, researchers might be queried about so-called concurrent validity, which might also be known simply as ‘relation to existing measures’. In practice, concurrent validity tends to consist of a moderate degree of shared variance between performance on the new task and performance on any highly cited task that purports to assess a similar ability. There are at least two reasons why researchers developing a new task might not desire concurrent validity. First, many established tasks do not stand up to a cursory analysis of their own (multifaceted) validity (see Purser, 2015 for an example), so that concurrent validity in relation to these tasks might be undesirable. Second, the new task might not be intended to assess exactly the same type of ability that the reference task was designed to assess, rendering the comparison flawed. However, concurrent validity might rightly be considered as important under some circumstances. One example is developing a culturally neutral version of an existing task: in this case, at least when testing the population for which the original task was devised, high concurrent validity would be necessary to convince us that the new test is measuring the same construct as the original.

Another aspect of validity, which is often used to scrutinise new tasks, is face validity. This is simply whether the task appears to assess what it is intended to assess. Although this aspect of validity might appear, on the face of it (no pun intended), to be somewhat vague, it can be operationalised in terms of inter-rater reliability on metrics of relevance and relatedness to the ability intended to be probed by the task (see, for example Nevo, 1985). Practically, this would involve a study that evaluated the face validity of the task. In assessing the face validity of a new vocabulary task, say, participants (the raters) would rate the task on scales or questions assessing the degree to which the task appeared to measure vocabulary, or to address the specific aspect of vocabulary that the task was designed to measure (see also Chapter 5 for a detailed discussion of inter-rater reliability).

A further, and very important aspect of validity is construct validity, which concerns the degree to which a task actually assesses (rather than appears to assess) what it is intended to assess. It is beyond the scope of this chapter to discuss this in detail, but it is worth noting that good construct validity relies heavily on the quality of the underlying theoretical concepts on which the task is based (see Messick, 1995). For the researcher, ensuring construct validity essentially means ensuring that the scientific basis of the task is sound: the theoretical framework is empirically supported and internally consistent, and the characteristics of the measure are what the theoretical framework would predict. For example, if our theory of vocabulary acquisition entails a gradual accretion of known words over chronological age, then we would expect to see a steady increase of our vocabulary measure when plotted against age.

4.4 Conclusions

In this chapter we have discussed some practical aspects that are important when using either standardised or experimental tasks. Standardised tasks are often used when making a clinical diagnosis, to decide whether a child belongs to a certain experimental group or to match experimental groups to one another. As the procedure and instructions are standardised, the normed outcomes of these tasks can be compared between groups and testing sessions as well as over time.

Limitations of standardised tasks include the problem that they can only be administered by a trained professional and that often the standardised procedure and instructions do not take into account or allow permutations for cultural differences or language issues. More importantly, as with any task, they only provide information about the task at hand and thus often one is obliged to design a new experimental task in order to manipulate the dependent variable in question.

Designing your own experimental task can be time consuming, especially when one has to control for many variables and one has to spend enough time piloting the task to ensure validity. However, if designed properly they can be used to match participants on non-central tasks demands, with the inclusion of an appropriate control condition that differs from the main experimental task only in the information required. The logic behind conventional matching procedures, based on standardised tasks or mental-age assessments, is to attempt to equate groups for these non-central demands, but these procedures approach this only indirectly.

Regardless of whether one uses a standardised task, or develops a new experimental task, there are a few issues one has to take into account when working with children. First and most importantly, when working with children one has to take development seriously. This means that one needs to use tasks that are age-appropriate and when one works with atypical groups this often requires tasks that span a large age range. This can be quite challenging both when it comes to finding a standardised task as well as when creating a new experimental task. Yet, this is very important in order to avoid floor and ceiling effects as we have discussed. Secondly, one needs to ensure that children pay attention to the task at hand. This can be done by monitoring where children are looking, allowing children to take breaks or by including motivational stimuli and stories when designing your own experimental task.

Finally, those administering the tasks should be wary of their own eye movements and avoid looking at the correct answer, in order not to give the correct answer away as children in general look where adults are looking in order to learn new things.

Still, tasks, whether they are standardised or experimental will only render information about the variables that are being tested. In addition, the tasks do not tell us anything about how children perform in their daily life. For example, performance on language tasks, whether standardised or experimental, does not inform us how children communicate within different social settings such as in the home or classroom. So some research questions will require a more qualitative or mixed approach which are discussed elsewhere within this volume (see also Chapters 10, 11, 13, and 14 for further discussion of mixed methods research).

Practical tips

1. Keep instructions simple and include a pre-test to check that children understand the wording used.
2. Motivation is an important aspect of children's performance. Therefore, when you design your task ensure that it includes a narrative or stimuli that children will enjoy.
3. Although reaction times (RTs) provide you with rich data sources, RTs are often very variable in children for a number of reasons (e.g., concentration during the task, motor abilities, distractability, etc.) and thus, unless one has a control task that matches the experimental task exactly except for the variable at interest, one has to question whether they should be used in research with children.

References

Aunio, P., & Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy. *Learning and Individual Differences, 20*, 427-435.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition*, *21*, 37-46.

Bayley, N. (1993). *Bayley Scales of Infant Development* (2nd ed.). San Antonio, TX: The Psychological Corporation.

Bowler, D. M. (1997). Reaction times for mental-state and non mental-state questions in 'theory of mind tasks': Evidence against logico-affective states in Asperger's syndrome. *European Child and Adolescent Psychiatry, 6*, 160-165

Brock, J., Jarrold, C., Farran, E. K., Laws, G., & Riby, D. M. (2007). Do children with Williams syndrome really have good vocabulary knowledge? Methods for comparing cognitive and linguistic abilities in developmental disorders. *Clinical Linguistics & Phonetics, 21*, 673-688.

Eisenberger, R., & Cameron, J. (1998). Reward, intrinsic interest, and creativity: New findings. *American Psychologist, 53*(6), 676-679.

Eliot, C. D., & Smith, P. (2011). *British Ability Scales* (3rd ed.). BAS*3*. London: GL-Assessment.

Epstein J. N., Langberg, J. M., Rosen, P. J., Graham, A., Narad, M. E., Antonini, T. N., … Altaye, M. (2011). Evidence for higher reaction time variability for children with ADHD on a range of cognitive tasks including reward and event rate manipulations. *Neuropsychology, 25*, 427-441.

Gould, J. (2005). Use of psychological tests in child custody assessment*. Journal of Child Custody, 2*, 49-69.

Gunn, D. M., & Jarrold, C. (2004). Raven's matrices performance in Down syndrome: Evidence of unusual errors. *Research in Developmental Disabilities*, *25*(5), 443-457.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The approximate number system in 3-, 4-, 5-, 6-year-olds and adults. *Developmental Psychology, 44*(5), 1457-1465.

Hervey, A., Epstein, J. N., Curry, J. F., Tonev, S., Arnold, L. E., Conners, C. K., … Hetchman, L. (2006). Reaction time distribution analysis of neuropsychological performance in an ADHD sample. *Child Neuropsychology, 12*, 125–140.

Jarrold, C., Baddeley, A. D., & Hewes, A. K. (1988). Verbal and non-verbal abilities in the Williams Syndrome phenotype: Evidence for diverging developmental trajectories. *Journal of Child Psychology and Psychiatry*, 39(4), 511-523.

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

Kail, R. V., & Ferrer, E. (2007). Processing speed in childhood and adolescence: Longitudinal models for examining developmental change. *Child Development, 78*, 1760-1770.

Kaland, N., Smith, L., & Mortensen, E. L. (2007). Response times of children and adolescents with Asperger syndrome on an 'advanced' test of theory of mind. J*ournal of Autism and Developmental Disorders, 37,* 197–209.

Karmiloff-Smith, A., D'Souza, D., Dekker, T., Van Herwegen, J., Xu, F., Rodic, M., & Ansari, D. (2012). Genetic and environmental vulnerabilities: The importance of cross-syndrome comparisons. *Proceedings of the National Academy of Sciences of the United States of America*, *190*(2), 17261-17265.

Kenworthy, L., Yerys, B. E., Anthony, L. G., & Wallace, G. L. (2008). Understanding executive control in autism spectrum disorders in the lab and in the real world. *Neuropsychology Review, 18*(4), 320–338.

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The Social Attribution Task. *Journal of Child Psychology & Psychiatry, 41*(7), 831-46.

Lange-Küttner, C. (2012). The importance of reaction times for Developmental Science: What a difference milliseconds make. Inaugural Issue*, International Journal of Developmental Science, 6*, 51-55.

Martens, M. A., Wilson, S. J., & Reutens, D. C. (2008). Research review: Williams syndrome: a critical review of the cognitive, behavioural, and neuroanatomical phenotype. *Journal of Child Psychology & Psychiatry*, *49(6),* 567-608.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Mussolin, C., Nys, J., Leybaert, J., & Content, A. (2012). Relationships between approximate number system acuity and early symbolic number abilities. *Trends in Neuroscience and Education, 1*, 21-31.

Negen, J., & Sarnecka, B. W. (2015). Is there really a link between exact-number knowledge and approximate number system acuity in young children? *British Journal of Developmental Psychology, 33*(1), 92-105.

New Group Reading Test Digital. (2010). UK: GL-Assessment.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*(4), 287-293.

Phillips, C. E., Jarrold, C., Baddeley, A. D., Grant, J., & Karmiloff-Smith, A. (2004). Comprehension of spatial language terms in Williams syndrome: Evidence for an interaction between domains of strength and weakness. *Cortex, 40*(1), 85–101.

Purser, H. R. M. (2015). Experimental difficulties in neurodevelopmental disorders: evidence from Down syndrome. In: J. V. Herwegen & D. Riby, (Eds.), *Neurodevelopmental Disorders: Research challenges and solutions* (pp.199-218). Hove: Psychology Press.

Purser, H. R. M., Farran, E. K., Courbois, Y., Lemahieu, A., Sockeel, P., & Blades, M. (2012). Short-term memory, executive control and children's route learning. *Journal of Experimental Child Psychology*, *113*, 273–285.

Raven, J., Raven, J. C., & Court, J. H. (2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales.* San Antonio, TX: Harcourt Assessment.

Rundblad, G., Dimitriou, D., & Van Herwegen, J. (under review). Comprehension of lexicalised metaphors and metonyms: a developmental study of typically developing children and children with Williams syndrome.

Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Jerome M. Sattler Publications.

Van Herwegen, J., Ansari, D., Xu, F., & Karmiloff-Smith, A. (2008). Small and large number processing in infants and toddlers with Williams syndrome. *Developmental Science, 11*(5), 637-643.

Van Herwegen, J., Dimitriou, D., & Rundblad, G. (2013). Development of novel metaphor and metonymy comprehension in typically developing children and Williams syndrome. *Research in Developmental Disabilities, 34*, 1300-1311.

Van Herwegen, J., Farran, E., & Annaz, D. (2011). Item and error analysis on Raven's Coloured Progressive Matrices in Williams Syndrome. *Research in Developmental Disabilities, 32*(1), 93-99.

Van Herwegen, J., Rundblad, G., Davelaar, E. J., & Annaz, D. (2011). Variability and standardised test profiles in typically developing children and children with Williams syndrome. *British Journal of Developmental Psychology,* 29, 883-894.

Van Luit, E. E. H., van de Rijt, B. A. M., & Pennings, A. H. (1994). *Utrechtse Getalbegrip Toets (Early Numeracy Test)*. Doetinchem: Graviant.

Wagner, J. B., & Johnson, S. C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cognition, 119*, 10-22.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (1999). *Manual for the Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: Psychological Corporation.

Williams, D., & Jarrold, C. (2013). Assessing planning and set-shifting abilities in autism: are experimenter administered tasks and computerised versions of tasks equivalent? *Autism Research, 6*(6), 461-467.