Dynamic assessment as an early screening tool for identifying New Zealand children at risk of reading difficulty upon school entry

A thesis presented in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

in Education

at Massey University, Manawatū, New Zealand

Susan Bisschoff

2019

Abstract

The purpose of this study was to investigate the use of a dynamic assessment as a screening tool for identifying children at risk of reading difficulty. Unlike traditional static assessment, dynamic assessment includes a teaching stage within the assessment and aims to determine what the child can do independently as well as what they have the potential to do when given quality input.

At the start of their formal schooling, 165 New Zealand children were administered a dynamic assessment of phonological decoding, along with several static measures of emergent literacy skills. At the end of their first year at school, these same children's reading abilities were assessed using multiple early reading measures. The results were analysed to determine whether measures administered at the beginning of formal schooling significantly predict future reading ability, and whether there is a significant difference in the ability of the static and dynamic measures to predict future reading difficulty and in their respective predictive classification accuracy.

Results indicated that the dynamic assessment of decoding was able to predict future reading difficulty with a high level of accuracy and that it provided superior predictive ability and classification accuracy to that of the static measures of emergent literacy. Furthermore, combining the dynamic and static measures did not improve the overall ability of the dynamic measure alone to predict future reading difficulty. The ease and efficiency of administration of the dynamic assessment, as well as its ability to provide information pertinent to supporting remedial intervention, provided evidence of this measure's acceptability as an effective universal screening tool.

Taken together, the findings indicate that a dynamic assessment of decoding can accurately predict future reading difficulty and that it has the potential to meet the other important characteristics of an effective universal screening tool. This provides support for the use of a dynamic assessment of phonological coding as a universal screening tool for the prediction of reading difficulty at the start of children's formal schooling.

Acknowledgements

The saying goes that “it takes a village to raise a child”, and it would seem the same is true for a doctorate. I am incredibly grateful to all those in my doctoral village who gave me guidance, support, and encouragement during the course of my doctoral journey.

First, I would like to thank my supervisors, Dr Tara McLaughlin, Dr Sally Clendon, and Assoc Prof Alison Arrow. I was truly blessed to have such knowledgeable, supportive, and insightful supervisors. Each of you offered feedback from a unique perspective on my thesis as it unfolded. I am grateful for the challenging questions you asked, detailed comments and constructive criticism you gave, and your ongoing support over the past years. I count it a privilege to have had the opportunity to benefit from your wisdom, and I look forward to continuing to do so in future. I would also like to thank Prof Tom Nicholson for his meaningful input during the initial stages of my doctoral research.

My sincere gratitude goes to all the schools who agreed to participate in this research, and to the parents who consented to their children’s participation. Special thanks go to the teachers who allowed me to work with the children in their classes, and those who took the time to provide me with the details of the children’s reading skills. In particular, I want to thank all the children who took part in the study. I am grateful to have had the opportunity to meet these wonderful children, to spend time finding out more about their literacy journey, and to learn from them.

I want to give special thanks to my family. This thesis would not have been possible without them. To my husband, Dustin, thank you for your steadfast support through the many weekends of having to keep the boys entertained while I worked and for encouraging me every step of the way. To my children, Declan and Finley, thank you for being so patient and understanding when

it seemed I was always working, and for giving me plenty of hugs when I needed it most. Declan, your challenges with reading when you first started school, motivated me to begin this study, and your amazing literacy ability today has shown me first-hand that with early and targeted input, these challenges do not need to persist. Finley, thank you for reminding me, in your own special way, when the work-life balance was becoming tipped too heavily in the work direction.

Finally, I want to thank my mother and father. Mom, thank you for your unwavering support and love. Dad, thank you for listening to me talk endlessly about the exciting, challenging, frustrating, and inspiring aspects of this research, and for the hours you spent pouring over references and the like. Your keen interest in my research and loving support has meant so much to me.

Contents

Appendices

List of tables

List of tables in appendices

List of figures

List of boxes

Chapter 1:
Introduction

Reading with accuracy, fluency, and comprehension is a key ability on which all school-based learning and achievement depends. A child who falls behind in reading also tends to fall behind in other curriculum areas. There is an abundance of research that shows that children who have difficulty learning to read at the very beginning of their schooling are likely to continue to have difficulties reading throughout (Boets et al., 2011; Corriveau, Goswami, & Thomson, 2010; McCardle, Scarborough, & Catts, 2001; Ortiz et al., 2012). This puts them at a disadvantage in terms of academic achievement, future (post-school) education, and as a result, restricts employment prospects. Struggling readers often become adults who continue to struggle with reading, and as such do not get the opportunity to fully engage in all the opportunities available to their more literate contemporaries. As Torgesen puts it "Clearly, children who become adults with low levels of literacy are at an increasing disadvantage in a society that is creating ever-higher demands on effective reading skills within the workplace" (2002, p. 8). In addition to these practical concerns, reading difficulties and the concomitant challenges have a negative effect on an individual's psychological and emotional well-being. Self-esteem and motivation, including motivation to read, are negatively affected in those who struggle to read, and this impact is felt from an early age (Burden, 2008; Lyon, 2003).

Reading difficulty (RD) affects a significant number of school-aged children in New Zealand, with estimates ranging from 5% to 10% (Dyslexia Foundation of New Zealand, 2015; New Zealand Ministry of Education, 2007). Although there is some disagreement as to the exact nature and aetiology of RD, there is a general consensus that early diagnosis and intervention are crucial (Kantor, Wagner, Torgesen, & Rashotte, 2011; New Zealand Ministry of Education, 2008; Petersen & Gillam, 2015; Wilson & Lonigan, 2010). It is well-documented that literacy

achievement in early schooling declines more rapidly for children who are below-average readers (Wilson & Lonigan, 2010) and that a reading difficulty, once established, is difficult to remedy (Petersen, Allen, & Spencer, 2016; Torgesen, 2002). For this reason, preventing a potential reading difficulty is seen as preferable to trying to "cure" an established difficulty. Early intervention is likely to be more effective, efficient and cost-effective than trying to remedy reading difficulties once that have already taken hold (Kantor et al., 2011; Tunmer, Chapman, Greaney, Prochnow, & Arrow, 2013; Vaughn & Fuchs, 2003). For this reason, internationally there has been a move away from a "wait-to-fail at reading" approach to one of early detection and prevention. Converging evidence points to the need for assessment to predict reading difficulty risk and early intervention to take place even before the child begins formal reading instruction (Bishop & League, 2006; Bridges & Catts, 2011; Huang, Moon, & Boren, 2014; Lonigan, Purpura, Wilson, Walker, & Clancy-Menchetti, 2013). In New Zealand, this would be before the child starts formal schooling at age five. Numerous studies have shown that early identification of children at risk of RD and early, intensive, evidence-based intervention can play a major role in preventing later reading problems (Bishop & League, 2006; Bridges & Catts, 2011; Huang et al., 2014; Lonigan et al., 2013; O'Connor & Jenkins, 1999).

Early intervention is particularly important for those children who need it the most, such as those with dyslexia. For example, research has shown that children with dyslexia take significantly longer to master specific reading skills than other children (Nijakowska, 2010). Thus, it is important that these children get help as soon as possible, particularly in the light of research findings that indicate that with age, reading difficulties become more severe and harder to remediate (Nijakowska, 2010; Morlini, Stella, & Scorza, 2014; Tunmer & Greaney, 2010). Early intervention can have a marked positive impact on the severity of the reading difficulty and the concomitant effects on the child's other studies, self-esteem, anxiety, and motivation to read (Norton & Wolf, 2012).

Universal screening for reading difficulty risk

Early intervention is contingent on early assessment that accurately predicts the child's reading risk status. Assessment used to predict possible future reading difficulties aims to screen children for risk of reading difficulty *before* they experience difficulties, rather than to diagnose an existing reading difficulty. This identification of risk of reading difficulty allows for early intervention to take place, even before children learn to read, with the aim of ameliorating or even preventing a reading difficulty from developing. The challenge is to identify universal screening measures that can accurately identify those at risk at this early time.

A range of traditional static tests of emergent literacy skills such as letter naming, phonological awareness, and rapid naming have been shown to be good predictors of future reading ability (Bishop & League, 2006; Catts, Nielsen, Bridges, & Liu, 2014; Schatschneider, Francis, Carlson, Fletcher, & Foorman, 2004). However, these tests have generally been less effective in predicting which children are likely to go on to develop a reading difficulty. Floor effects are common with these traditional static tests and this negatively impacts on their classification accuracy as they either under- or over-identify reading difficulty risk (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009; Petersen et al., 2016). Possible reasons for these floor effects include that children at this young age may struggle to understand the assessment instructions, or that they are not yet able to independently perform the tasks required (Bridges & Catts, 2011; Catts et al., 2009). In response to these shortcomings, new research, particularly in the United States of America and Europe, has focused on the use of dynamic assessment as a screening tool for reading difficulty risk (Gellert & Elbro, 2018; Petersen et al., 2016). Dynamic assessments include a teaching stage within the assessment and aim to determine what the child can do independently as well as what they have the potential to do when given quality input. There is emerging evidence that dynamic assessments of reading do not exhibit floor effects and can more accurately predict future reading difficulty at this early stage than traditional static assessments (Petersen et al, 2016).

Rationale for current study

In New Zealand, many children are given their first formal literacy assessment at the end of Year 1, using Clay's (2013) *Observation Survey*. The results of this assessment are used to determine which children need to be referred for remedial action, most commonly in the form of *Reading Recovery* during the child's second year at school (Education Review Office, 2018). However, there is wide agreement amongst researchers that prevention of early reading difficulties should begin much earlier than this and that intervention before or when the child starts formal reading instruction is more effective than remedial action later in primary school (Moyle, Heilmann, & Berman, 2013; Scanlon, Vellutino, Small, Fanuele, & Sweeney, 2005; Tunmer et al., 2013). For this reason, it could be argued that this current ''wait-to-fail at the end of Year 1" approach in New Zealand schools is doing our children a disservice.

Given these concerns around current practices relating to assessment for reading difficulty in New Zealand schools and the issues identified with traditional static assessments for younger children, this study seeks to examine whether a dynamic assessment of phonological decoding can be used to accurately identify risk of reading difficulty in New Zealand children at the start of their formal schooling. The use of dynamic assessment as a screening tool for risk of reading difficulties is relatively new and at the time of writing there were only a few longitudinal studies that had examined the use of dynamic assessment of decoding in children at the start of their formal schooling to predict future reading difficulty. One of these was conducted with Danish kindergarten children, and involved children being taught the letter sounds corresponding to non-alphabetic symbols, and how to blend these into words (Gellert & Elbro, 2018). Three other studies were conducted by Petersen and colleagues (Petersen & Gillam, 2015; Petersen et al., 2016; Petersen, Gragg, & Spencer, 2018), in the United States of America, which has a very different education system to that of New Zealand and where universal testing and screening is routine. Furthermore, a relatively large proportion of the participants used English as an

additional language, with Spanish as their first language. A sample representative of the New Zealand population would have a much smaller proportion of participants with English as an additional language, and there would be a variety of different first languages (e.g., Māori, Fijian, Mandarin, Afrikaans, Hindi, and so on). All three of these United States based studies provided support for the predictive validity and classification accuracy of a dynamic assessment of decoding to predict risk of reading difficulties. However, as Petersen and Gillam (2015) point out, "Because an assessment procedure appears to have evidence of predictive validity for one group of children does not mean that it has universal applicability across all cultures and languages" (p. 16). To date, there has been no research in New Zealand into the use of dynamic assessment as an early universal screening tool for risk of reading difficulty. The New Zealand Ministry of Education has emphasised the need for evidence-based research to guide decision making in order to provide educational services that address the needs and goals of all learners, and allow for better targeting of investment, resources, and support (2016). This study aims to contribute evidence-based research to add to the growing body of research worldwide into the use of dynamic assessment as a tool for predicting reading difficulty risk to support the implementation of early, targeted interventions for those children who need it most. It also seeks to address the current research gap into the use of universal screening tools to predict the risk of reading difficulty within the New Zealand context.

In support of this aim, this study utilises a range of data analysis procedures to examine and compare commonly used static assessments of emergent literacy, and a dynamic assessment of decoding, in terms of their ability to accurately predict risk of future reading difficulty. This includes logistic regression analyses to determine whether the static and dynamic reading measures administered at the start of formal schooling were able to predict reading difficulty status after a year at school, and receiver operating characteristic analyses to examine how accurate these measures were in terms of correctly identifying a child's reading difficulty risk.

Structure

This thesis consists of six chapters, including this one. In chapter 2, a review of literature germane to this study is presented with the focus on when screening for reading difficulty should take place, which skills should be tested to screen for reading difficulty, and how these should be assessed. This includes a review of relevant theories of reading acquisition and reading difficulties. The chapter describes static assessment tools commonly used to screen for reading difficulty and compares and contrasts these with the use of dynamic assessment. This is followed by a brief review of the key considerations used to evaluate the measurement quality and acceptability of assessments used to predict reading outcomes, and the literature comparing static and dynamic measures in terms of their measurement quality. Finally, the current study is framed in terms of a review of the most pertinent literature and gaps are highlighted in the existing research that this study aims to address.

Chapter 3 outlines the aim of the research, and the specific areas of investigation. The research design for the study is detailed, including the participants and how they were recruited, as well as the assessment measures employed and the procedures for their administration, scoring, and analysis.

Chapter 4 presents the findings of the research, focussing on each of the areas of investigation. In Chapter 5, these results are interpreted and discussed in the context of previous research in this area. Finally, Chapter 6 concludes the thesis by reflecting on the contribution made by this research, the implications for the New Zealand context, as well as the limitations of the research and suggestions for further research.

Terminology

To follow are definitions of key terminology used throughout this study. These definitions reflect the current usage of these terms at the time this study was undertaken and, where relevant, how this terminology applies within the context of the current study.

Reading difficulty

In this study, the term *reading difficulty* refers to a condition where a child has difficulty reading compared to other children of the same age. As reading difficulty status for this study is determined at the end of a child's first year at school, the determination of a reading difficulty is based on levels of ability in early reading skills that children at this stage are expected to have acquired. This includes measures of the child's ability to correctly identify and name the letters of the alphabet, segment three- and four-phoneme words into their individual phonemes, phonologically decode pseudowords, read real words, and read early levelled reading books.

Reading difficulty risk

Reading difficulty risk refers to a child's risk of developing a reading difficulty. For the current study, assessment was undertaken at the start of children's formal schooling, before formal reading instruction had begun, to determine which children were at risk of developing a reading difficulty. As such it was not used to determine whether a reading difficulty was already present, but rather to indicate whether there was an increased likelihood that a reading difficulty might develop.

Reading status

In this context of this study, *reading status* refers to whether the child is classified as having, or not having, a reading difficulty after a year at school. For a child classified as having a reading difficulty after a year at school, their reading status would be *reading difficulty*; for a child classified as not having a reading difficulty, their status would be *no reading difficulty*.

Universal screening for reading difficulty

Universal screening refers to the brief assessment of all children to identify those at risk of developing a reading difficulty. Universal screening allows for those identified as at risk to be given targeted early intervention and/or to be tested further to acquire a more detailed understanding of the child's risk profile.

Screening tool

A *screening tool* is used to identify the presence of a condition or risk of a condition developing in individuals that currently show no signs or symptoms of the condition. In the case of reading difficulty, a screening tool is used to identify children at risk of developing a reading difficulty, even before any indication that such a reading difficulty may be present (i.e., before formal reading instruction has begun). The purpose of a screening tool is to identify those at risk so that interventions can be put in place to help ameliorate or even prevent a reading difficulty developing.

Predictive validity

Predictive validity is the ability of a measure to predict a future outcome (Field, 2015). In the context of this study, predictive validity refers to the ability of a measure administered to children at the start of their formal schooling (predictor measure), to predict future reading performance as determined by outcome measures administered a year later. Predictive validity is established by determining a statistically significant correlation between a predictor measure and an outcome measure (Mislevy & Rupp, 2010).

Predictive modelling

Predictive modelling involves the use of models to predict a future outcome. These models aim to predict the value of an outcome variable based on a predictor variable or variables. To determine the quality of a predictive model, the predictive validity of such fitted models is

reported. Where the outcome variable is dichotomous (i.e., a variable that has only two possible values such as *reading difficulty*/*no reading difficulty*), logistic regression, area under the receiver operating characteristic curve (AUC), and measures of classification accuracy can be used to determine the quality of a predictive model. Two key ways in which to evaluate a predictive model is to look at predictive power and goodness-of-fit statistics.

Model goodness-of-fit

The *model goodness-of-fit* statistic is used to compare models to establish whether it is possible to improve the ability to predict an outcome by adding predictors to the model. The -2LL statistic (also referred to as the log likelihood ratio) is used to compare the goodness of fit of two statistical models, with the model chi-square X^2 statistic indicating the difference between the -2LL for a model containing the predictor variable, and that of a null model (Pett, 2016; Tabachnick & Fidel, 2012).

Predictive power

In the context of a model with a dichotomous outcome, *predictive power* refers to how well the outcome variable can be predicted based on the predictor variables (Allison, 2014). Two commonly reported predictive power statistics are pseudo-R^2 and AUC. These values range from 0 (meaning no predictive power) to 1 (meaning perfectly predictive).

Classification accuracy

Classification accuracy is an additional way of estimating the criterion-related validity of a screening tool and refers to how effectively a tool can correctly identify a particular condition. In the context of this study, it refers to how well a tool can (a) correctly identify a child as *at risk* who will go on to have a reading difficulty; and (b) accurately identify a child as *not at risk* who will not have a reading difficulty in the future. There a several descriptors of classification accuracy including sensitivity, specificity, and receiver operating characteristic curve (ROC curve).

Sensitivity

Sensitivity refers to the ability of a measure to correctly predict the existence of the target condition (e.g., a reading difficulty). It is the proportion of true positives in a total group of individuals with the target condition. In the current study, sensitivity refers to the percentage of children correctly identified at the start of school as being at risk of a reading difficulty, out of the entire group of children classified as having a reading difficulty after a year at school.

Specificity

Specificity refers to the ability of a measure to correctly predict the absence of the target condition. It is the proportion of true negatives in a total group of individuals with the target condition. In the current study, specificity refers to the percentage of children correctly identified at the start of school as not being at risk of a reading difficulty, out of the entire group of children classified as not having a reading difficulty after a year at school.

ROC curve

An ROC curve gives a visual presentation of sensitivity and specificity. The overall classification accuracy of a particular predictor is measured by the area under the ROC curve (AUC), with an area of 1 representing a perfectly predictive measure, and an area of .5 indicating predictive accuracy that is no-better-than-chance.

Acceptability

Acceptability refers to the perception key stakeholders have of the agreeability or palatability of a given assessment tool (Proctor et al., 2011). The acceptability of a tool within a school setting is frequently linked to factors such as cost, ease of administration and scoring, and accuracy.

Summary

In this chapter, the importance of early and accurate assessment to screen for the risk of reading difficulty was introduced, and the rationale and aim of the current study were discussed. The structure of the thesis was outlined, and key terminology were also presented. In the chapter that follows, the literature pertinent to assessment to screen for the risk of reading difficulty is reviewed in greater detail.

Chapter 2:
Literature review

Early identification of reading difficulty risk allows for a "prevention rather than cure" approach to reading difficulties as it makes early intervention possible, which in turn may ameliorate or even prevent reading difficulties. As such, early identification and subsequent early targeted intervention, are more efficient in terms of resources such as time, money, and effort than later identification and attempting to remediate already established reading difficulties (Compton et al., 2010; Kantor et al., 2011; Petersen et al., 2016). Although there are several important reasons to test for reading difficulty risk before formal reading instruction has begun, one may wonder whether it is possible to assess reading skills before the child has received any formal reading instruction, and if so, which skills are best able to predict future reading ability.

This chapter reviews literature that addresses the question of which skills should be assessed to accurately screen for reading difficulty risk (Section 1). It also reviews theory and previous research germane to the question of what form this assessment should take (Section 2). In this regard, it includes an overview of the problems commonly associated with traditional static measures used to screen for reading difficulty risk. Section 3 reviews literature relevant to dynamic assessment, an assessment approach that focusses on the construct of modifiability (or responsiveness to input), and how dynamic assessment may be able to address issues with traditional static measures. Section 4 outlines the characteristics of effective measurement to predict future reading status, and how these are evaluated, and includes a comparison of static and dynamic measures in this regard. In Section 5, research most closely related to the current study is reviewed to highlight the gaps in existing research which this study aims to address. Finally, Section 6 gives a summary of the key points revealed by the review of the literature related to the prediction of risk of reading difficulty in young children.

Section 1: What should be assessed?

Decisions on which components of reading ability to include in a universal screening tool are chiefly based on determining which are likely to predict future reading difficulty most accurately and which will be able to provide useful information for preventative or remedial action to be taken. This section discusses relevant theories of reading development and reading difficulties, as well as evidence from pertinent longitudinal and intervention research studies to help identify key components of reading that are predictive of future ability.

A theory of reading provides a framework for reading research, instruction, and remediation. It is important to work from a theory of reading that aligns with the particular type and aspect of research one is undertaking. The focus of the current study relates to the prediction of future reading difficulties in beginning readers. Therefore, this section reviews theory and previous research that focusses on beginning reading and addresses the following questions: (1) What does it mean to read?; (2) How does reading develop?; (3) Which components/skills in beginning readers provide effective predictors of their later reading ability?

What does it mean to read?

The question "What does it mean to read?" is more complex than one might expect: Is a child reading when they can say the printed words on a page?; or does the child need to be able to both decode and understand what they mean to really be reading? Alternatively, is reading being able to understand how the meaning of individual words relate to each other in phrases, sentences, paragraphs, and the text as a whole?; or does it go even further than this: Does the child need to be able to use what they have read and understood to map to their existing knowledge and create new knowledge? In addressing the question of what it means to read, it is important to consider the age or stage of learning because what may be considered reading for a 6-year-old would be

different than that for an adult. Understanding what it means to read at a particular age or stage is important in determining the skills children need to become successful readers, and therefore which types of skills may predict future reading difficulties.

There are several theories that attempt to explain what it means to read. One of the most influential theories of reading, which has been widely supported by several research studies (e.g., Catts, Adlof, & Weismer, 2006; Kendeou, Savage, & van den Broek, 2009; Roberts & Scott, 2006; Stuart, Stainthorp, & Snowling, 2008), is *The Simple View of Reading* (SVR) (Gough & Tunmer, 1986; Hoover & Gough, 1990). According to the SVR, reading can be defined as the product of decoding and language comprehension. For an individual to be said to be reading, they must be able to both decode words in the written text and use this to understand the text as a whole. Decoding and comprehension are both essential for skilled reading. Hence the formula is: Reading = Comprehension X Decoding (R = C X D) rather than: Reading = Comprehension + Decoding (R = C + D). In a model where R = C + D, this would imply that one could be said to be "reading", even if one of the two components (decoding or comprehension) was absent.

Decoding, in the context of the SVR, refers to "efficient word recognition: the ability to rapidly derive a representation from printed input that allows access to the appropriate entry in the mental lexicon, and thus, the retrieval of semantic information at word level" (Hoover & Gough, 1990, p. 130). However, to be truly reading, it is not enough to recognise the words in a written text, one also needs to understand what these words, put together in sentences, paragraphs, and so on, mean in the context of the text as a whole (i.e., comprehension).

Language comprehension, in the context of the SVR, refers to the ability to take "lexical information (i.e., semantic information at the word level) and derive sentence and discourse interpretations" (Hoover & Gough, 1990, p. 131). In the same way that decoding without comprehension is not reading in the true sense, someone who is unable to decode written text, but has good language comprehension (i.e., oral language comprehension), is not able to read.

As both decoding and language comprehension skills are essential components of reading according to the SVR, reading difficulties could arise in three different ways. Firstly, a child can have adequate decoding skills, accompanied by inadequate comprehension skills; in other words, they have no difficulty accurately decoding (reading out) printed words, but have difficulty constructing meaning from the text they have decoded. These children are identified as having *hyperlexia* (Hoover & Gough, 1990), or a *specific reading comprehension deficit* (Catts et al., 2006). Secondly, a child can have adequate language comprehension skills, accompanied by inadequate decoding skills (*dyslexia*). These children do not have any difficulty understanding oral text, but they struggle with decoding (and therefore also accessing the meaning) of written text. Finally, a child can have inadequate language comprehension and decoding skills (*mixed* or *general reading difficulty*).

Despite these different possibilities for how reading difficulties can arise, there is an abundance of research studies that have shown that most children who experience reading difficulties have early and ongoing difficulties with accurate and fluent word identification skills. Research into the relationship between word identification, language comprehension, and reading comprehension shows that children who have poor reading comprehension typically have poor word identification skills (in terms of both fluency and accuracy), and that those who have poor word identification skills show impaired reading comprehension, even if they have adequate language comprehension skills (Gough & Tunmer, 1986; Vellutino et al., 1996).

Furthermore, while word decoding and language comprehension are both required for reading comprehension, the contribution made by each of these, and the relationship between them, differs across the development of reading skills from emergent to expert reader (Catts et al., 2006). Therefore, to determine whether skills related to decoding or language comprehension are more likely to predict individual differences in readers, it is important to consider when the testing is taking place, as screening tools that have strong predictive validity at one phase in reading development

may be less valid for a different phase. The next section reviews the literature pertaining to reading development, as this sheds light on the specific component skills most likely to predict the reading ability of different readers at different stages of development, with the focus being on what the existing literature reveals regarding beginning readers.

How does reading develop?

According to Hoover & Gough (1990), for children in the early stages of learning to read (in the early school years), decoding and language comprehension are unrelated, but both correlate with reading comprehension. Importantly, during the first few years of learning to read, decoding correlates substantially more strongly with reading comprehension than language comprehension, with decoding having a coefficient of about 0.55 (large effect) and language comprehension, 0.35 (medium effect) (Hoover & Gough, 1990). In the later school years, decoding and language comprehension are more closely related, and both still relate to reading comprehension. However, from this point onwards, the relationship between language comprehension and reading comprehension becomes stronger than that between decoding and reading comprehension (Adolf, Catts, & Lee, 2010; Roberts, Mohammed, & Vaugh, 2010; Stuart et al., 2008; Yeong, Fletcher, & Bayliss, 2014).

In the current study, the participants are beginning readers for whom learning to decode individual words is the primary skill that needs to be acquired. This is not to say that comprehension has no role, even at this early stage, or that once a child has acquired word recognition skills, they will automatically be able to develop reading comprehension. However, accurate, efficient, and rapid word recognition allows a child's cognitive resources to be freed up for higher level tasks related to comprehending the text, whereas the child who has deficient word reading skills, gets "stuck" focussing on the decoding of individual words and is therefore unable to attend to the meaning of the text (Ehri, 2005a; Ouellette & Fraser, 2009; Share & Shalev, 2004).

In their definition of decoding within the SVR, Hoover and Gough (1990) point out that for beginning readers, the particular element of decoding that plays the most important role, is the ability to derive phonologically-based representations of novel printed words. In other words, the child needs to be able to pronounce (either silently or aloud) the written word. This in turn allows them to access their mental lexicon of known words to recognise the printed word. Therefore, phonological processes are key to the process of *learning* to read, and for beginning readers, it would be important to assess their ability to phonologically decode novel words.

For beginning readers, the primary role played by word recognition skill, and in particular the ability to phonologically decode novel words, is supported by several theories and models, including Ehri's *Phase Theory* (2005) and Share's (1995) *Self-Teaching Hypothesis*. Furthermore, both Ehri's *Phase Theory* and Share's *Self-Teaching Hypothesis* provide elucidations of why phonological skills, as well as alphabet knowledge, underpin the ability to phonological decode text, which in turn is key to the process of learning to read.

According to Ehri (2005b) there are four ways in which words can be decoded.

1. *Phonological recoding*: The process of converting printed letters (graphemes or letter strings) into sound units (phonemes or syllables) and blending these to sound out a word.

2. *Analogy*: Using known words to read novel words, for example, using the known word *dog* to "read" the novel word *log*. However, reading by analogy is not a strategy available to beginning readers as they have not yet acquired a sufficient store of words in their mental lexicon.

3. *Prediction (contextual guessing)*: Using letter and context clues to guess novel words.

4. *Sight word (orthographic) reading*: Reading words from memory by recognising words that have been read before. This is what forms the foundation of efficient and accurate reading comprehension. The competent reader no longer needs to phonologically decode every word they encounter, with the vast majority of words being recognised and understood from memory. This allows the reader to devote their attention to understanding and engaging with the text, rather than focussing their effort and attention on the laborious process of matching the letters on the page to sounds.

Ehri (2014; 2005b) distinguishes four phases in the process by which children learn to read words by sight. The *pre-alphabetic* phase occurs in the earliest period. During this phase, children do not know much about the alphabetic principle and as such do not use grapheme-phoneme connections to read words. Any words they do read are recognised visually and from contextual clues (e.g., they may recognise the 'shapes' that make up their own name, or they might recognise *KFC* from contextual clues such as the KFC logo and/or where the sign is).

During the *partial alphabetic* phase, children start to learn the sounds and/or names of the letters in the alphabet and start to use these to connect to sounds. However, because they do not yet have full knowledge of the alphabetic system (especially vowels, diphthongs, and so on), and are not yet able to segment a word into all its phonemes, they are unable to map all the letters on the page to sounds. Most often they are only able to map the letter sounds at the beginning and end of a word, because these are easier to perceive. They then use "contextual guessing" to "fill in the gaps" to read the word. Unfortunately, this tactic is often unsuccessful, and children confuse words that have the same initial and final letters (e.g., *pan* and *pen*; *look* and *lock*). There is some evidence that children with reading difficulties get "stuck" in the partial alphabetic phase (Ehri & Saltmarsh, 1995) and that, as a result, they have difficulty accurately (and efficiently) phonologically decoding novel words. For this reason, assessment of alphabet knowledge during this phase can help to identify those children who may be at risk of a reading difficulty.

In the *full alphabetic* phase, children can form complete matchings between the letters in a written word and the phonemes in the pronunciation of that word, because they are familiar with the main grapheme-phoneme links. They can also segment the spoken word into all its phonemes and match this to the graphemes they see in the written word. It is during this phase that children start to learn sight words more easily.

As children increase the number of sight words stored in memory, they enter the *consolidated* phase. During this stage they are more familiar with the morphological features of words and the grapheme-phoneme links are no longer restricted to individual letter-sound (grapheme-phoneme) connections. Instead links begin to occur between larger letter-sound chunks such as rimes, syllables, morphemes, and whole words.

David Share's *Self-Teaching Hypothesis* addresses the question of *how* individuals who are learning to read move from having to phonologically decode words to being able to recognise words visually; in other words, how orthographic learning (learning to visually recognise words from memory) takes place. Share (1995) identifies three ways in which orthographic learning can take place:

1. *Direct instruction*: rote learning of new words.
2. *Contextual guessing*: using "clues" from the surrounding text to predict unfamiliar words.
3. *Phonological recoding and the self-teaching mechanism*: each time the learner successfully decodes (phonologically) an unfamiliar word, they have the opportunity to develop orthographic knowledge, which in turn is the key for skilled word recognition.

Share argues that of the three, "only phonological recoding offers a viable means for printed word learning" (1995, p. 152). Direct instruction is untenable because of the huge number of new words young learners encounter each year. Learners can, and do, acquire a small number of words by rote association (e.g., via flash cards or basal readers) and there is evidence that this can be helpful in the very early stages of reading acquisition because a relatively small number (about 100-150) of

high frequency words account for a significant portion of the vocabulary in children's reading material. Furthermore, a significant proportion of these high frequency words are phonically irregular (e.g., *said*, *their*, *write*, *people*). However, the sheer number of vocabulary items a skilled reader needs to recognise, and the cognitive load required to learn these by rote, makes direct instruction an insufficient strategy for efficient reading acquisition beyond the earliest stages.

Contextual guessing (prediction) is likewise a poor means by which to build word recognition skill because natural text has low predictability, resulting in a very high probability of incorrect guessing (Share & Stanovich, 1995). To develop an orthographic lexicon, it is not enough to correctly guess the *meaning* of a particular word; the specific word itself needs to be identified. If, for example, the child uses contextual clues such as pictures to guess that the word *stool* says *chair*, they do not get the opportunity to acquire the orthographic representation of the word *stool* because, instead of attending to the orthographic features of the word itself, they are simply (incorrectly) guessing what the word might mean. However, this does not mean that there is no place for contextual information in learning to read. On the contrary, the *Self-Teaching Hypothesis* posits that contextual information can play an important role in aiding exact word pronunciations based on impartial decoding. However, this role is ancillary to reading acquisition, and because natural text is not highly predictable, contextual guessing is insufficient for reading acquisition.

Unlike direct instruction and contextual guessing, the ability to independently "recode" printed words into their spoken form presents a feasible and efficient strategy for reading acquisition. This is because whenever a child successfully decodes a novel written word, they get the opportunity to learn word-specific orthographic information. It is this word-specific orthography that lies at the heart of proficient word recognition. The central idea of the *Self-Teaching Hypothesis* is that by phonologically recoding (or "sounding out" a word), the child is able to engage with and therefore teach themselves the orthographic features of these words, which in turn is the basis of sight-word reading (Share, 1999).

The more frequently a child is exposed to, and correctly decodes a particular word, the sooner the child is able to form an orthographic representation of that word, and therefore no longer needs to rely on phonology to access the item. For this reason, from early in the process of learning to read, children are likely to be able to read high-frequency words orthographically, with little need for phonological processing. However, new or less familiar words for which children have not yet formed orthographic representations will depend more heavily on phonological processing to be decoded (Share, 1995). Later in the process of reading development, phonological recoding becomes increasingly influenced by the child's growing orthographic knowledge. Rather than relying on simple grapheme-phoneme (letter-sound) correspondences, the child increasingly uses what they have learnt about the "rules" of the target language to aid in accurate recoding of lexical items. This is line with what Ehri (2005b) refers to as the *consolidated* phase of learning to read words by sight, as during this stage children have an increased number of sight words in memory and are more familiar with the morphological features of words, allowing for individual letter-sound connections to be replaced by links between larger letter-sound units and even whole words.

Therefore, in summary, self-teaching is dependent on two processes: the phonological process (which is the primary process) and the orthographic process (which is considered secondary). Even though phonological recoding "provides the opportunity for self-teaching, … one's ability to assimilate word-specific orthographic information determines how quickly and accurately orthographic representations are acquired as a result of phonological recoding" (Loveall & Conner, 2013, p. 110).

This theory is supported by several studies that have shown that children with greater pre-existing phonological recoding skill, and/or who more accurately phonologically recoded target words, showed greater orthographic learning of those words (e.g., Conners, Loveall, Moore, Hume, & Maddox, 2011; Nation, Angell, & Castles, 2007; Ouellette & Fraser, 2009; Suárez-Coalla, Ramos, Álvarez-Cañizo, & Cuetos, 2014). Conners and colleagues (2011) investigated whether the relationship between phonological recoding and word identification (i.e., sight-word reading) was mediated by orthographic knowledge. They explored whether the *Self-Teaching Hypothesis* applies to individual differences in reading skills in that this hypothesis may point to the possibility that children who have superior phonological recoding abilities will also acquire greater orthographic knowledge, which in turn allows them to have better word recognition skills. Results of their research indicated that the relationship between phonological decoding and word identification is significantly mediated by orthographic knowledge.

Numerous studies have shown that readers who are typically developing only need to phonologically recode a novel word a few times to acquire orthographic information about that word, and correctly read the word as a sight word (e.g., Bowey & Muller, 2005; Ehri & Saltmarsh, 1995; Share & Shalev, 2004). In contrast, those with reading difficulties require significantly more exposures to a word, and children with the most severe reading difficulties (e.g., dyslexia) have difficulty acquiring orthographic learning even after many exposures to the target words. For example, Suárez-Coalla and colleagues (2014) investigated the ability of dyslexic children to develop orthographic representations of the words they read. They found that dyslexic children did not show evidence of forming orthographic representations after six exposures to target words, but children without dyslexia (who were the same age or reading level) were able to do so. In line with the theory that phonological processing plays a primary role and orthographic processing a secondary role in reading acquisition, this could indicate that in terms of decoding, a child can experience difficulties because of a phonological deficit,

an orthographic deficit (in terms of ability to acquire orthographic knowledge), or a combination of both. Therefore, for early readers, assessing both their ability to phonologically recode novel words, as well their ability to acquire orthographic knowledge about words, would be important.

There is evidence that the self-teaching process by which children learn to recognise words from memory (i.e., orthographic reading) can begin even before a child is able to sound out and blend (i.e., phonological recoding). For the beginning reader, the simpler grapheme-phoneme correspondences are "sufficient to kickstart the self-teaching mechanism which is then able to refine itself in the light of expanding orthographic knowledge" (Share, 1995, p. 156). Therefore, the two skills that need to be in place for the child to be able to "self-teach" in an alphabetic orthography are 1) phonemic awareness: the understanding that spoken language consists of units of sound (phonemes) that can be segmented, blended, or changed in different "patterns" to make all the words of the language; and 2) alphabetic knowledge: understanding and knowing the different written symbols that correspond to these sound units. In other words, seen in the context of Ehri's theory of phases of learning to read words, the self-teaching mechanism can take effect as early as the partial alphabetic phase, although it is at its most efficient once children are in the full alphabetic phase (Ehri, 2005b). As a result, for children in the earliest stages of learning to read, such as those just starting school, assessment of phonemic awareness and alphabetic knowledge would be important to assess, as they underpin the beginning readers' ability to self-teach orthographic knowledge, which in turn provides the foundation for orthographic reading.

Summary

The Simple View of Reading posits that to be a successful reader, both the ability to successfully decode words and to comprehend what has been read are needed. However, for the beginning reader, the ability to successfully decode words is the primary skill that needs to be acquired. Being able to read words accurately and efficiently frees up a child's cognitive resources for tasks related to text comprehension. According to the Ehri's *Phase Theory* and Share's *Self-Teaching Hypothesis*, to learn how to read words orthographically (i.e., sight word reading), a child first needs to acquire the ability to convert printed letters into sound units and blend these to sound out a word (i.e., phonological recoding). In an alphabetic orthography such as English, two emergent literacy skills are required for the child to be able to phonologically decode words and therefore begin the process of learning the orthographic representations required for automatic word recognition (i.e., sight word reading): phonemic awareness and alphabet knowledge. Therefore, in the light of these theories the key skills that are likely to be useful to assess to determine future reading ability in beginning readers including phonemic awareness, alphabet knowledge, phonological decoding of novel words, and orthographic learning.

Key predictors of reading ability identified in previous research

Numerous research studies support the central role of phonological decoding in learning to read as identified in the theories discussed above (e.g., Bosse, Chaves, Largy, & Valdois, 2015; Conners et al., 2011; Loveall & Conner, 2013; Ricketts, Bishop, Nation, & Pimperton, 2011; Ziegler, Perry, & Zorzi, 2014). A meta-analysis of around 500 research articles found that the best predictor of decoding skill in a child's first year of formal schooling and beyond is the child's ability to decode pseudowords in the preschool period (National Early Literacy Panel, 2008). A good reason to assess pseudoword reading as a predictor of future reading ability before the onset of formal reading instruction, is that with increased reading experience it becomes more problematic to use pseudowords to accurately measure phonological decoding ability. For example, if a pseudoword is similar to a real word (e.g., *pog* versus *dog*), the child may read the word by analogy: they may compare the novel word (*pog*) to the known word (*dog*) stored in their lexicon and use this to help them decipher the word. This use of orthographic decoding strategies would reduce demands on phonological (sub-lexical) decoding (Cotton & Crewther, 2012); therefore, testing children using pseudowords before they have started to develop their orthographic lexicon provides a cleaner measure of phonological decoding ability.

Despite the theoretical advantages of assessing pseudoword reading before the beginning of formal reading instruction, for novice readers, assessing pseudoword decoding is generally not considered to be a viable option as the vast majority of children have little or no decoding ability at this stage. This is because, as the theories discussed earlier suggest, at this age, children do not yet have the phonological recoding knowledge to be able to sound out pseudowords, and thus measures of pseudoword decoding are likely to be subject to floor effects. For this reason, researchers have instead turned to the component skills and precursors of phonological decoding to identify potential predictors of reading skill.

In this regard, the National Early Literacy Panel study found five key variables that provide significant prediction of later literacy outcomes: alphabetic knowledge, phonological awareness, rapid naming, writing own name, and phonological short-term memory. Even when other variables such as oral language, concepts about print, and visual perception skills were controlled, these variables provided significant prediction of later literacy outcomes (National Early Literacy Panel, 2008). Indeed, as is discussed in the sections that follow, there is almost universal agreement in the research literature that alphabet knowledge, phonological skills and, to a lesser extent, rapid naming, are significant predictors of later reading ability.

Alphabet knowledge and phonological skills

The importance of alphabet knowledge and phonological skills suggested by the theories discussed earlier, has been supported by a number of training and intervention studies that have shown how instruction targeted at improving letter sound mapping and phonological awareness have positively impacted on word identification, spelling, and reading ability in general (National Early Literacy Panel, 2008; Share, 2004; Share & Shalev, 2004; Vellutino, Fletcher, Snowling, & Scanlon, 2004). There is also an indication that these skills are interrelated, with difficulties around phonological abilities negatively impacting on alphabet knowledge. The *phonological coding deficit hypothesis* posits that children with reading difficulties show deficits in the representation, storage and/or retrieval of speech sounds (phonological abilities) and this in turn impedes the acquisition of alphabet knowledge (Nijakowska, 2010). As Share (1995) points out, for beginning readers a letter name or sound is a type of pseudoword in the sense that it is a novel phonological string, and therefore a phonological deficit could hinder the learning of letter names and sounds, which in turn may help to explain why alphabet knowledge is such a good predictor of future reading status.

In this sense, deficits in phonology can be seen as being at the heart of most reading difficulties. For children with the most severe reading difficulties, phonological deficit is

present prior to formal reading instruction and is the main cause of later reading problems, with the severity of the phonological deficit predicting the severity of the later reading difficulty (e.g., Bridges & Catts, 2011; Kantor et al., 2011; Lervåg & Hulme, 2014; National Early Literacy Panel, 2008; Tunmer & Greaney, 2010).

The importance of phonological skill in learning to read is also supported by research in the areas of connectionism, neuroimaging, and behaviour genetic studies. *Connectionist computational models* are computer applications that simulate detailed features of behaviour, including simulations of how children learn to read, competent reading, and reading difficulties. Connectionist research has demonstrated the importance of phonological decoding for early learners and that it continues to play a role, even for skilled readers, for example, when they encounter a novel word (Seidenberg, 2007). Since the start of the new millennium, development and improvements in magnetic resonance imaging (MRI) have made possible a number of *neuroimaging studies* with children to identify the neural underpinnings of phonological awareness and these support the importance of phonological skill in reading development (Butterworth & Kovas, 2013; Gabrieli, 2009; Nijakowska, 2010; Norton & Wolf, 2012; Shaywitz, Lyon, & Shaywitz, 2006; Shaywitz, 2005; Vellutino et al., 2004; Wolff, 2014). Recently there have also been several *genetic and twin studies* that provide insights into the biological and behavioural differences related to reading abilities, and the relative influence of genetic and environmental variables on future reading ability. These also point to phonological skill being key to individual differences in reading ability and that phonological deficit is an endophenotype (biomarker which indicates familial risk) of reading difficulties as it is associated with both literacy deficits and familial risk (Moll, Loff, & Snowling, 2013; Nijakowska, 2010; Vellutino et al., 2004; Vellutino & Fletcher, 2005).

When assessing phonological awareness, it is important to consider the course of phonological skills development in children. For example, in line with Ehri's *Phase Theory*, children increase

their sensitivity to smaller units within words as they get older, at first being able to detect and manipulate syllables, then onsets and rimes, and finally phonemes. For many preschool children, many tasks involving phoneme substitution and/or manipulation are too difficult and are therefore likely to suffer from floor effects (Aaron, Joshi, & Quatroche, 2008; Anthony & Francis, 2005; Torppa et al., 2007). However, by the start of formal schooling, rhyme identification and production, syllable identification, blending, and deletion, as well as first phoneme isolation are within the abilities of most children. By the end of their first year at school, most children are also able to produce rhymes, break words down into their individual sounds (phonemic segmentation), and blend two phonemes (Mather, Wending, & Kaufman, 2011).

In summary, for beginning readers, the assessment of alphabet knowledge and phonological awareness is important in identifying potential reading difficulties. For children who have not yet started formal reading instruction, phonological awareness assessment could take the form of tasks such as first phoneme isolation and rhyme identification, whereas after about a year at school, children can be assessed used using more complex phonological awareness tasks such as phonemic segmentation.

Rapid naming and fluency

Some argue that the ability to rapidly name objects, colours, letters, or words is merely a component of phonological skill and that it engages phonological processing abilities such as the rate of retrieval of phonological codes in long-term memory and articulatory planning (Arnell, Klein, Joanisse, Busseri, & Tannock, 2009; Share, 1995; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997). Rapid automatized naming (RAN) is included in some widely used standardised tests of phonological processing such as the *Comprehensive Test of Phonological Processing - CTOPP-2* (Wagner, Torgesen, Rashotte, & Pearson, 2013) where the rapid naming tasks are considered to be measures of the ability to efficiently retrieve phonological information from long-term memory.

However, a number of recent studies have shown that phonological awareness and RAN account for unique variance in reading achievement (Arnell et al., 2009; Cronin, 2011; de Jong, 2011; Norton & Wolf, 2012; Wolff, 2014); and that phonological processing and RAN are only moderately correlated (Wolf & Bowers, 1999). Other researchers posit that instead of being a sub-component of phonological processing, RAN taps processes important to the acquisition of orthographic representations of words (Manis, Seidenberg, & Doi, 1999; Wolf & Bowers, 1999).

However, despite there being little consensus as to *why* rapid naming predicts future reading ability and the specific nature of its relationship to other reading-related skills, there is little question that it *is* a stronger predictor, in particular of reading fluency (e.g., Cronin, 2011; Furnes & Samuelsson, 2011; Jones, Branigan, & Kelly, 2009; Lervåg & Hulme, 2014; McAlenney & Coyne, 2015; Wolff, 2014). As mentioned earlier, skilled reading relies on both accurate and efficient (rapid and automatic) recognition of words (Ehri, 2005b). When recognition of words is laborious and slow, it makes it difficult for the child to devote cognitive resources to comprehension of the text as a whole. According to Jones and colleagues (2009), over the past 25 years, more than 100 published studies have used RAN as a measure of reading skill. In some cases, it has been shown to be a more reliable longitudinal predictor of reading than phonological awareness (Norton & Wolf, 2012; Torgesen et al., 1997; Wolf & Bowers, 1999) and rapid naming deficits are exhibited by 60 to 70% of individuals with reading or learning disabilities (Norton & Wolf, 2012). As with deficiencies in phonological awareness, rapid naming deficits have been shown to be clear and prevalent, even for high-functioning adult readers with dyslexia (Jones et al., 2009).

Rapid naming is most commonly assessed using RAN measures. In these measures, children are presented with colours, pictures of familiar objects, numbers, letters, or words and asked to name them as quickly as possible. The child being tested needs to be familiar with the items to be named and for this reason the type of items used varies by age group.

For children who have not yet started to learn to read or have not yet learnt the alphabet, rapid naming can be assessed using colours or familiar objects. Although colours are commonly used with this age group, it should be kept in mind that an estimated 6% of New Zealand males have some form of colour vision deficiency (New Zealand Health Technology Assessment Clearing House, 1998). Therefore, testing with familiar objects may be a more suitable form of RAN for children who are not yet able to read or name letters.

Alphanumeric items are generally used for children who have started reading instruction, and older children are usually asked to rapidly name pseudowords or real words. For children who have already started to learn how to read, assessing not only the accuracy, but also the speed at which they read pseudowords and real words, have been shown to be strong predictors of future reading skill (Florit & Cain, 2011; Joshi & Aaron, 2012; Sprenger-Charolles, Siegel, Béchennec, & Serniclaes, 2003). Even though pseudowords or real words are most commonly used with older children, it should be noted that rapid naming of familiar objects, colours, and alphanumeric items, continue to be useful in predicting reading performance into young adulthood (Arnell et al., 2009).

There are two main formats in which RAN items can be presented: as discrete or serial RAN. In the case of discrete RAN, items are presented one at a time and the naming latency (i.e., the time from when the item has been presented to when the child names the item) is measured. The discrete RAN score is calculated as the mean naming latency over all presented items. For serial RAN, items are presented in several rows (or columns) and need to be named sequentially as quickly as possible. The serial RAN score is calculated as the total time taken to name all the presented items. In general, serial RAN has been found to have a stronger correlation with reading of both individual words and serial words than discrete RAN (Wolf & Bowers, 1999), although there is some evidence that this is only the case for younger, beginning readers. In the case of advanced readers (those who can read most words by sight), RAN format correlates most closely with the reading outcome format, with discrete RAN best predicting discrete word

reading and serial RAN being a better predictor of serial word reading (de Jong, 2011; Zoccolotti, De Luca, Marinelli, & Spinelli, 2014). However, because serial RAN (rather than discrete RAN) better predicts both serial and discrete word reading in beginning readers, it has become the most commonly used format of RAN in studies with this age group of learners.

Summary

Reading is a complex process, and as a result it is unsurprising that there are several factors that can impact the development of reading ability and that can be used as indicators of potential risk of reading difficulties. However, by narrowing the focus to those components most closely linked to individual differences between beginning readers in their ability to learn to read, the evidence strongly supports the role of phonological skills, including phonological awareness and phonological decoding ability. There is also strong evidence for the role of alphabet knowledge and RAN as indicators of potential risk of reading difficulties, although there is some indication that these are also influenced by, and have an influence on, phonological awareness.

In the next section, the focus is on how to assess these key components of reading skill. This includes a discussion of static assessment tools commonly used to screen for risk of reading difficulty, the limitations of these tools, as well as alternative approaches to static assessment. This is followed by a detailed discussion of how one such alternative approach - Dynamic Assessment - has been developed to address the shortcomings of static assessment.

Section 2: What form should assessment take?

The previous section focused on which skills should be assessed to predict reading difficulty. In this section, the focus moves to the form this testing should take. Assessment used to screen for reading difficulties has most often been in the form of traditional, static assessments of phonological awareness, alphabetic knowledge, rapid naming, and word reading. This section looks at several shortcomings of these static assessments, as well as alternative assessment approaches that attempt to address these limitations.

Static assessments give a snapshot of the child's abilities at a single point in time and are widely used because they are standardised, quick and easy to administer, norm-referenced, and produce results that classify students clearly. Static assessments measure the skills and knowledge a child has already acquired and is able to apply at a given point in time and within a test context. They are generally characterised by minimal, highly scripted, and standardised assessor input, with the assessor (at most) only giving initial instructions on what the child needs to do, and sometimes a short practice of the test task (Camilleri & Botting, 2013).

However, static assessments of early reading skills have several important limitations, which has led to an increased focus on alternative forms of assessment for the identification of future and current reading difficulties. Although research has shown that there are significant positive correlations between phonological awareness, decoding skill, and reading skill, static assessments of these abilities to identify risk of reading difficulty commonly suffer from inadequate *sensitivity*: high rates of false negatives with children not being identified as being at risk who later develop reading difficulties (Torgesen, 2002); and *specificity*: high rates of false positives, with children incorrectly being identified as at risk who do not go on to develop a reading impairment (e.g., Boscardin, Muthén, Francis, & Baker, 2008; Caffrey, Fuchs, & Fuchs, 2008; Kantor et al., 2011; McAlenney & Coyne, 2015; Nijakowska, 2010; Torgesen, 2002;

Vellutino, Scanlon, Zhang, & Schatschneider, 2008). The deleterious consequences of over- and under-prediction are immediately apparent. On the one hand, providing intensive, targeted intervention for children identified as at risk for a reading difficulty is resource-heavy. Over-prediction results in unnecessary allocation of these resources. On the other hand, under-prediction means that children who need help do not get it as soon as it's required, with lasting negative consequences for their reading and academic development.

According to Petersen and Gillam the "majority of research pertaining to the predictive evidence of validity of early reading measures has revealed that a balance between good sensitivity and good specificity is difficult to obtain" (2015, p. 3). They go on to report that several studies have shown over-prediction of risk status, with specificity ranging from 47% - 67%; and several have shown under-prediction of risk status, with sensitivity ranging from 21% to 69%. Over-prediction (i.e., insufficient specificity and false positives) is particularly a problem in the case of very young children and those who have had limited preschool literacy experience or have culturally and linguistically diverse backgrounds (Petersen et al., 2016). In the case of young children who have not yet received any formal reading instruction, this lack of specificity is often the result of floor effects. Many children who are just starting their formal schooling are not able to perform the tasks included in a standardised static test. Floor effects may mean that screening tools are unable to distinguish between children who perform poorly on these measures as a result of experiential or instructional deficits, rather than those who are truly at risk for reading difficulties (Bridges & Catts, 2011; Petersen et al., 2016). Unfortunately, those screening protocols that do show acceptable sensitivity and specificity, tend to be too time-consuming and inefficient to be used as a universal screening tool when children begin formal schooling (Compton et al., 2010).

Some previous studies have found that measures given later, when children can be assessed using measures that resemble actual reading tasks (rather than emergent literacy skills), have greater predictive power (McAlenney & Coyne, 2015). Therefore, some would argue that it may be better to wait until the end of Year 1 or the start of Year 2 to assess for risk of reading difficulties. However, as McAlenney and Coyne (2015) point out, waiting to identify children in need of reading intervention is not an appropriate solution as there is overwhelming evidence of the importance of identifying those at risk for reading difficulties as early as possible, and providing them with early targeted intervention. Intervention studies have shown that even before the start of formal reading instruction, children can be taught phonological awareness, letter knowledge, and other emergent literacy skills, and that explicit instruction in these areas can have a positive impact on reading outcomes (Dufva, Niemi, & Voeten, 2001; Hulme & Snowling, 2009; Lonigan et al., 2013; Lundberg, Frost, & Petersen, 1988).

Another possible reason for the poor specificity of static assessments is the way in which they are administered. To ensure standardisation and avoid assessor impact on the assessment results, static assessments aim to limit interactions between the assessor and person being assessed. There are, therefore, two results possible: unaided success or unaided failure (Fuchs, Compton, Fuchs, Bouton, & Caffrey, 2011). As static assessments usually provide little or no practice before testing or feedback during testing, "failure" exhibited by very young children and those from disadvantaged backgrounds, may be a result of being unable to understand the test instructions or familiarity with the test setting and conditions, rather than an inability to perform the task itself (Caffrey et al., 2008; Spector, 1992).

There is substantial evidence that norm-referenced assessment tools result in over-prediction of children who belong to cultural or linguistic minority groups. These children may perform poorly on traditional standardised tests because of factors that are not directly related to the child's actual or potential ability. Factors could include, for example, the use of test formats or content

that are less familiar to children from minority backgrounds, or norms that are based on children from majority backgrounds (Moyle et al., 2013). For similar reasons, children from socioeconomically and educationally disadvantaged backgrounds tend to be over-predicted for reading and other learning difficulties (de Beer, 2010). Furthermore, because static assessments only assess what the child is currently able to do, they do not distinguish between the child who has performed poorly because they have a (potential) reading difficulty, and the child whose prior literacy and learning experiences lead to their poor performance.

Regarding children who have not yet received formal reading instruction, another limitation of static assessments is that they are only able to assess emergent skills such as phonological awareness and letter identification. Static assessment is typically used to assess decoding skill only once the child has received significant reading instruction, by which point reading difficulties may already have emerged. In the light of the importance of intervention at the earliest possible time, this is a significant shortcoming of static measures.

Alternatives to static assessment

Since the 1990s there has been increased interest in using response to instruction or intervention as an alternative to static assessment to identify reading difficulty and reading difficulty risk. The concept of *responsiveness* or *response to instruction/intervention* refers to an approach to the early identification and support of young learners with, or at risk of, reading difficulty. Universal screening and iterative progress monitoring are used to provide data about a learner's responsiveness to instruction and level of achievement. These data are used to identify those students who are at risk of reading difficulty and therefore require intervention (Fuchs & Fuchs, 2006; Stoiber & Gettinger, 2016).

Response to Intervention (RTI) is most commonly used in North America where it is usually takes the form of a three-tiered approach. At Tier 1, learners receive high quality, research-based instruction in a general classroom setting. Periodic screening and progress monitoring are used to identify at-risk learners and these learners receive supplemental instruction in the general classroom. At the end of this stage (which could last up to 8 weeks), learners who show significant progress (i.e., good "responsiveness") are returned to the normal classroom program. Learners who demonstrate inadequate responsiveness at Tier 1 are moved to Tier 2, where they are given intervention directed specifically to their needs. These interventions are typically provided in small-group settings as ancillary to general classroom instruction, and this stage could last up to 20 weeks. Learners who despite these directed interventions continue to show inadequate progress may then be moved to Tier 3 where they receive intensive, individual reading interventions targeted to their specific difficulties (Kaminski & Powell-Smith, 2017; Stoiber & Gettinger, 2016).

Several researchers have produced evidence in support of the use of RTI for the early identification of children who may be at risk of developing reading difficulty (McAlenney & Coyne, 2015; Vellutino et al., 2008). However, RTI as a framework for identification of reading difficulty-risk is not without its critics. Chief criticisms include the length of time involved in identification of reading difficulty at different tiers. Targeted intervention for those who need it most may be delayed for several months, or even years. This means that, in a sense, RTI becomes another wait-to-fail approach (Al Otaiba et al., 2014; Compton et al., 2012; Fuchs et al., 2011; Gustafson, Svensson, & Fälth, 2014; Kantor et al., 2011; Lidz & Peña, 2009). This delay in intervention is made worse because, in most cases, RTI models are only implemented once the child begins formal schooling (Gustafson et al., 2014; Moyle et al., 2013).

Another criticism is that RTI generally employs the use of static (one-point-in-time) tests for screening students at risk of reading difficulty. However, as discussed earlier, there are several limitations of using static tests with young children and those from socioeconomically or educationally disadvantaged backgrounds, including that floor effects make the results of these assessments difficult to interpret (Bridges & Catts, 2011; Caffrey et al., 2008; Petersen et al., 2016). This leads to false positives (i.e., the misidentification of many young children as having reading problems).

Further criticism is that RTI is a resource-heavy approach that is time-consuming, requires additional expertise and input from teachers and other school professionals and other resources for implementation (Gustafson et al., 2014). Furthermore, the identification of reading difficulty within an RTI model is based on learners who lack responsiveness to *high quality*, *research-based* instruction or intervention. This requires ongoing professional development and training to ensure that teachers provide instructional programmes that can expect to lead to positive outcomes. If there are discrepancies between the fidelity and integrity of the instruction/intervention provided within an RTI model, then it becomes impossible to determine the learner's level of responsiveness to effective instruction/intervention (Gustafson et al., 2014; Tunmer & Greaney, 2010; Vaughn & Fuchs, 2003).

In response to these shortcomings of static assessment and RTI, an alternative approach to assessment, known as dynamic assessment, has recently been investigated. The concept of dynamic assessment has been around for more than 80 years but has, until recently, focussed mainly on general intelligence testing. However, recently there has been renewed interest in the use of dynamic assessment to address the shortcomings of using traditional static assessments or RTI to identify (risk of) reading difficulties. In the section that follows, an overview is given of the use of dynamic assessment to predict reading future reading difficulty.

Section 3: Dynamic assessment

There is no single definition for dynamic assessment as it encompasses several variants and approaches and has been variously termed *learning potential assessment* (e.g., Budoff, 1987; Sternberg & Grigorenko, 2002), *mediated learning* and *structural cognitive modifiability* (Feuerstein, Feuerstein, & Falik, 2010); *testing the limits* (Carlson & Wiedl, 1979), *assisted learning and transfer by graduated prompts* (Campione & Brown, 1990), *learning test* (Guthke & Stein, 1996), *dynamic testing* (Sternberg & Grigorenko, 2002; Swanson, 1992), and *assisted assessment* (Campione, 1989). However, what all these approaches have in common is the inclusion of some form of intervention (teaching) as integral to the assessment process.

In this section, the focus is fourfold: (1) a review of the literature related to the theoretical underpinnings of dynamic assessment; (2) an overview of the forms of dynamic assessment; (3) a comparison of the different formats which dynamic assessment may take; and (4) a summary of the relationship between dynamic assessment and RTI.

Theoretical underpinnings and models of dynamic assessment

Dynamic assessment has its roots in the theories of Vygotsky and Feuerstein. From Vygotsky's *Social-Cultural Theory* comes a key principle underlying dynamic assessment: the concept of the *Zone of Proximal Development* (ZPD) (Poehner, 2008; Tzuriel, 2001).

According to Vygotsky, the ZPD is, "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978, p. 86). The ZPD is unique to each child; some will require greater guidance and assistance to master new learning, while others will do so with much less input. The Vygotskian approach is not only interested in what the child can do today (i.e., the "present child") but also

what the child will be able to do in the future (i.e., the "future child") (Bodrova & Leong, 2007). It argues that if one only looks at what the child can do today without assistance (i.e., what is within their *Zone of Actual Development*), the child's full capacity is not being measured because the potential to master new learning has not been included in the assessment. This potential to acquire new learning is known as the child's *learning potential* (Cho & Compton, 2015) or *cognitive modifiability* (Dörfler, Golke, & Artelt, 2009). For Vygotsky, assessment can only accurately predict a child's development if it includes measures of both the child's independent performance and measures of their assisted performance (Lidz & Gindis, 2003).

In line with this, in a dynamic assessment, children are given tasks that are within their ZPD: tasks they have the potential to master with guidance and support, but which they are not yet able to perform independently. This differs from traditional (static) assessment where children are only tested on items that they should already be able to perform independently (i.e., the lowest level of the ZPD). For dynamic assessment, the aim is not only to ascertain what the child can do independently, but also what they have the potential to achieve with guidance and assistance. Furthermore, dynamic assessment enables the assessor to investigate the type and extent of guidance or assistance the child needs to reach this level of potential. For example, at a certain point in time, two children may both be unable to perform a specific skill and, in this sense, appear be at the same level of performance. However, one child may actually be a lot closer to being able to perform the skill than the other and given some input (e.g., explicit instruction, scaffolding, or some other form of intervention), this distinction between the two children could be revealed: one child may be able to perform the skill with only minimal support, whereas the other child may not be able to perform the skill, even with extensive assistance. The child who can perform the skill with only minimal support was probably already close to being able to produce the target skill and would be more likely to respond to core classroom instruction (i.e., typical high-quality classroom-based instruction). The other child, however, would likely need much more extensive and supportive input to attain the target skill.

Feuerstein's *Structural Cognitive Modifiability Theory* was developed independently of Vygotsky's theory, but they share much in common. Central to Feuerstein's theory is the idea that humans are not static beings, but that they are *modifiable* – in other words, their cognitive functioning and structures can be modified (Feuerstein et al., 2010). This modifiability is best achieved through *mediated learning experiences* (MLE). According to Feuerstein's MLE theory, MLE "occurs when a person (mediator) who possesses knowledge, experience, and intentions mediates the world, makes it more understandable, and imparts meaning to it …" (Feuerstein et al., 2010, p. 24). Based on these theories, Feuerstein and his colleagues developed the *Learning Propensity Assessment Device* (LPAD), which is a battery of assessment tools and methods of application that aim to assess the individual's *modifiability* (Feuerstein et al., 2010). Modifiability is related to learning potential and refers to the child's "propensity to learn from new experiences and learning opportunities and to change one's own cognitive structures" (Tzuriel, 2014, p. 17) . It has been hypothesised that modifiability is an independent construct which, if measured appropriately, can be assessed without confounding factors such as language, prior learning, or socioeconomic status biasing the results (Petersen & Gillam, 2015).

These key concepts from Vygotsky and Feuerstein have been reconceptualised and employed in many different contexts by researchers, teachers, and clinicians in a wide range of fields including intelligence testing, additional language learning, mathematics, and speech and language therapy (e.g., Jeltova et al., 2011; Safa & Beheshti, 2018; Wang & Chen, 2016). Recently, in an attempt to address some of the shortcomings of static assessment and RTI measures, there has also been increased interest in the use of dynamic assessment for predicting reading difficulty-risk (e.g., Bridges & Catts, 2011; Caffrey et al., 2008; Camilleri & Botting, 2013; Coventry, Byrne, Olson, Corley, & Samuelsson, 2011; Fuchs et al., 2011; Gustafson et al., 2014; Kantor et al., 2011).

Forms of dynamic assessment

Dynamic assessment is actualised in several different forms. Campione (1989) provides a useful taxonomy for describing the similarities and differences across three general dimensions: focus, intervention, and target.

Focus

Focus refers to the measures derived from the assessment. In general, dynamic assessments could have as their focus: (a) measures of the change which has taken place as a result of input from the assessor (e.g., measures of the difference between pre- and post-test scores or of post-test performance) or (b) measures of the processes presumed to underlie the change that occurs (e.g., measures of learning strategy used, response to teaching, or extent of assistance needed). There appears to be some evidence that dynamic assessment that focusses on measuring the processes underlying change (learning) more closely correlate with future achievement than those that focus on measuring the change itself. For example, Petersen and Gillam (2015) compared the predictive validity afforded by a range of different scores produced during a dynamic assessment of decoding of pseudowords in kindergarten children. The task yielded the following scores.

- *Pre-test post-test gain scores in pseudoword reading*: Difference between score on the pre-test and score on the post-test.
- *Residuum gain scores*: How much the child still has to learn (i.e., the difference between their pre-test performance and the ceiling on a task). This is determined by calculating the percentage of the remainder between the pre-test score and the ceiling. For example, if there is a total of six test items and child A gets three correct on the pre-test and six correct on the post-test, their residuum gain is 100% (they needed to learn three more and they learnt all three). If child B has a pre-test score of 0 and

gets all six items correct on the post-test, their residuum gain is also 100% (they needed to learn six more and they learnt all six). Petersen and Gillam included this measure to try to account for the fact that children who have the most to gain (i.e., the greatest difference between their pre-test scores and the task ceiling) have the potential to show larger gains after the input phase of dynamic assessment.

- *Modifiability score*: This reflects the child's responsiveness during the input phase of the dynamic assessment and the extent to which the child applied what s/he was taught in the post-test.

In their analysis of which of these scores was most predictive of first-grade criterion measures, Petersen and Gillam found that the pre-test post-test gain score was not significantly predictive of first-grade criterion measures, while the residuum gain score did significantly correlate with the first-grade criterion measures with correlations between .33 and .39. However, it was the modifiability score that proved to be most predictive of future reading performance, yielding excellent classification results and moderately high to high correlations (between .46 and .51) with first-grade criterion measures. This may indicate that the child's response to input as it relates to being taught how to phonologically decode pseudowords (i.e., modifiability) may be a better predictor of future reading ability than that of change in phonological decoding ability after instruction alone. In summary, in terms of *focus*, dynamic measures that focus on the processes that underlie the change that occurs because of intervention (modifiability) are better able to predict future reading performance than those that focus on the change itself (i.e., difference between the pre- and post-test scores).

Intervention

Intervention refers to the form of input provided by the assessor during the assessment. In broad terms, one can distinguish between idiographic and nomothetic approaches or, as Caffrey and colleagues (2008) refer to it within the context of dynamic assessment, *clinically-oriented* and *research-oriented* dynamic assessment. These approaches differ in terms of how standardised and replicable the assessments should be.

Clinically-oriented dynamic assessment focusses on diagnosis and remediation, with the assessments specifically designed to attempt to produce change in the learner. It is more focussed on the individual learner than research-oriented dynamic assessment and generally uses a non-standardised approach with the assessor input varying from unstructured scaffolding to mediation that is contingent on the learner's responses. This means that the examiner's prompts, questions, teaching and feedback is non-standardised, making inter-individual comparisons difficult. Clinically-oriented dynamic assessment is strongly influenced by non-epistemic values such as usability, worthiness, and applicability (Grigorenko, 2009). In contrast, research-oriented or psychometric dynamic assessment, emphasises epistemic values such as objectivity, reliability, testability, accuracy, precision, generalisation, simplicity of concepts, heuristic power, repeatability and statistical analysis instead. In this regard, the focus on standardisation means that research-oriented dynamic assessment is more comparable with standard, static assessment measures (Campione, 1989). In a research-oriented dynamic assessment approach, the assessor input is not contingent on the learner's responses and can vary from scripted graduated prompts to standardised (sometime scripted) instruction. This more standardised approach facilitates inter-individual comparisons.

Dynamic assessment that employs standardised, non-contingent feedback has been shown to correlate more strongly with future achievement than that when feedback is individualised to the learner. In a meta-analysis of 24 studies exploring the predictive validity of dynamic

assessment, Caffrey and colleagues (2008) found that dynamic assessment studies with non-contingent feedback correlated .56 with future achievement, whereas those with contingent feedback correlated only .39 with achievement. Today, the majority of dynamic assessment tools used for prediction of future achievement employ standardised feedback, generally in the form of a scripted protocol for either *graduated prompts* (e.g., Bridges & Catts, 2011; Camilleri & Botting, 2013; Cho, Compton, Fuchs, Fuchs, & Bouton, 2014; Coventry et al., 2011; Fuchs et al., 2011; Fuchs et al., 2007; Spector, 1992) or *standardised instruction with reducing support* (e.g., O'Connor & Jenkins, 1999; Petersen & Gillam, 2015; Petersen et al., 2016).

In summary, in terms of intervention, dynamic measures that use standardised, non-contingent input are better able to predict future reading performance than those employing contingent feedback. This standardisation draws more comparability with static measures, whilst still offering the important difference of a teaching phase as part of the testing, allowing for the measurement of change in ability rather than a static measurement of current ability.

Target

Target refers to whether domain-general or domain-specific skills are being assessed using a dynamic assessment. Domain-general skills such as intelligence are those that operate identically regardless of the field or area of learning, whereas domain-specific skills are those considered to be specific to a particular domain such as reading or mathematics. The historical roots of dynamic assessment are in the field of intelligence testing and as such the target of such assessments has been domain-general. Today there are still those who support domain-general approaches (e.g., Feuerstein et al., 2010), however, more recently there has been increased support for domain-specific approaches. Theoretical support for a domain-specific approach comes from those who emphasise intra-individual variability across domains. As such there is no homogeneous, single or general zone of proximal development, but instead, several different domain-specific zones of proximal development including, for example,

numerical, verbal, visual, and auditory content domains (Campione & Brown, 1990; Sternberg, 2003). Recent studies have also shown that while dynamic assessment is effective in assessing learning potential within specific domains, it does not measure learning potential which is generalisable across domains (Cho & Compton, 2015; Tissink, Hamers, & van Luit, 1993). Furthermore, there are those who point out that domain-general assessments do not provide helpful information to teachers in terms of the action they can take to remediate existing or potential difficulties in specific domains (Campione, 1989; Campione & Brown, 1990). In summary, in terms of the *target* of the dynamic assessment, there is support for a domain specific approach over that of a domain general approach.

Dynamic assessment formats

It is possible to distinguish between two basic formats of dynamic assessment: 1) a *pre-test-input-post-test* format and 2) an *input-within-test* format. The pre-test-input-post-test format is the most commonly employed in research-oriented dynamic assessment and is sometimes referred to as the "sandwich" format because it consists of only a single layer of assessor input surrounded by testing (Grigorenko, 2009; Sternberg & Grigorenko, 2001). The following is a common pre-test-input-post-test protocol.

1. *Pre-test:* The learner's performance at the beginning of the test is measured. This can be equated to a static assessment, as it measures what the child is currently able to do, and involves the assessor giving no additional support or input other than telling the child what they need to do.
2. *Input*: The learner is provided with input from the assessor in an attempt to enhance their performance and to assess how well they learn and respond. As mentioned earlier, this input could vary from unstructured scaffolding to standardised scripted instruction.

3. *Post-test*: The learner's performance is measured again after they have received input from the assessor. This is used (sometimes in conjunction with the assessor's judgement of how well the child responded to input – in step 2) to determine their responsiveness to input (i.e., their modifiability).

The *input-within-test* format is most commonly used in clinically-oriented dynamic assessment and is sometimes also referred to as the "cake" format because it includes multiple layers of input from the assessor in between testing (Grigorenko, 2009; Sternberg & Grigorenko, 2001). The first assessment item is presented and, if the child provides the correct or expected response to the assessment item, this is acknowledged by the assessor and the second assessment item is presented. However, if an incorrect/unexpected response is given, the assessor immediately provides input to help the child correctly solve the item. This input could be in the form of hints, feedback, or graduated prompts. In the case of graduated prompts, the child is given increasingly supportive prompts in an attempt to enable them to provide the correct/expected response. The level of prompts required for the child to master the task item is used as an indicator of the his/her responsiveness to input (modifiability).

Dynamic assessment and RTI

Dynamic assessment shares much in common with RTI (for a review see Grigorenko, 2009; Lidz & Peña, 2009; Robinson-Zañartu & Carlson, 2013; Sternberg & Grigorenko, 2002; Wagner & Compton, 2011), with some arguing that they are two sides of the same coin (Grigorenko, 2009) and that RTI and dynamic assessment should be blended into a single model (Lidz & Peña, 2009). Dynamic assessment and RTI share the same goal of finding approaches to instruction and intervention that will result in successful learning for all learners (Lidz & Peña, 2009), and the same focus on responsiveness to instruction/intervention (modifiability). However, as a screening tool for reading difficulty risk, many see dynamic assessment as more efficient and cost-effective than RTI. Unlike RTI, where it takes weeks or even months before learners are

identified as at risk for reading difficulty, dynamic assessment usually takes place during one session of testing that lasts only a few minutes (Caffrey et al., 2008; Fuchs et al., 2011). So rather than waiting weeks or even months to "fail" (i.e., not respond to general classroom education), dynamic assessment can be used to screen children for risk of reading difficulty even before they begin formal reading instruction, giving them the opportunity to receive intervention early on, before they start to fall behind their peers.

Another key challenge of RTI is that it does not have a standard implementation, with different schools approaching assessment within the RTI model in different ways (Fuchs et al., 2007). However, when scripted standardised input protocols guide its use, dynamic assessment may be more straightforward to implement and easier to achieve reliably than RTI (Caffrey et al., 2008). Furthermore, there is some evidence that dynamic assessment may be better able to predict individual differences in responsiveness to input than RTI measures (Cho et al., 2014).

Summary

When the aim is to use dynamic assessment as a screening tool for reading difficulty risk, a review of the literature pertaining to the different forms and formats of dynamic assessment provides support for the use of a dynamic assessment in pre-test-input-post-test format that utilises standardised rather than contingent input, focuses on measuring modifiability (responsiveness to instruction) rather than pre-test post-test gain scores, and which is domain-specific rather than domain-general. The literature reveals that although RTI and dynamic assessment may have much in common, with a shared focus on responsiveness to instruction, dynamic assessment may be more effective as part of a preventative model that seeks to identify children at risk of reading difficulty very early, before a reading difficulty develops. The section that follows examines literature regarding the relative construct validity, predictive validity, and classification accuracy of static and dynamic measures as this pertains to the prediction of reading difficulties.

Section 4: Measurement quality and comparison of assessment approaches

Effective, informative assessment approaches in education are based on key standards for measurement quality. These often include *Standards for educational and psychological testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.), 2014). These standards include criteria against which tests should be evaluated to ensure they are well-constructed and well-interpreted. This in turn supports testing that provides sound information on which decisions can be made to benefit individuals, and others affected by these decisions (e.g., schools, family, etc). In additional to measurement quality, the acceptability of a particular measure is an important consideration as it impacts on the utilization of the measure in an applied context such as a school (O'Donnell & Miller, 2011). This section focuses on the key considerations used to evaluate measurement quality of tests used to predict reading outcomes and provides a review of the literature comparing static and dynamic measures to predict future reading outcomes in terms of their measurement quality. It also includes an overview of the concept of acceptability in terms of its importance to a screening tool for reading difficulty.

Measurement quality

When comparing the measurement quality of tools used to predict reading outcomes, key considerations are the validity (including construct and predictive validity), and classification accuracy of the measures. Construct validity refers to the ability of an assessment to measure what it asserts to be measuring (Brown, 1996) and is important because measures used to predict future reading status rely on accurate measurement of reading ability at a minimum of two points of time: the initial testing (time 1, predictor measure) and testing at a later time

(time 2, outcome measure). For example, to ensure the construct of reading is being measured at time 1 and time 2, it is important to consider the differences in what it means to read at those two different time points. As mentioned earlier, for beginning readers, the construct of reading encompasses the ability to phonologically decode individual novel words. For more experienced readers, the ability to decode words is less important and the ability to read sentences, paragraphs, etc, and to comprehend what has been read is considered to encompass the concept of reading.

Predictive validity refers to the ability of predictor measure to accurately predict the results of an outcome (i.e., criterion) measure. A measure is said to have predictive validity if the scores on that measure are significantly correlated with the scores achieved on an outcome measure. Linear regression may be employed to assess the relationship between continuous predictor and outcome measures and can therefore be used to determine how well a predictor measure is able to accurately predict a continuous outcome. However, linear regression is not able to provide information as to the ability of a predictor measure to predict a *dichotomous* outcome (e.g., reading difficulty/no reading difficulty), and therefore indicate the ability of the predictor measure to correctly predict group membership (e.g., reading status) (Pett, 2016). In contrast, logistic regression provides information about the probability that a specific condition will occur (e.g., reading difficulty/no reading difficulty) given certain conditions (e.g., scores on predictor measures). It is used to determine whether a discrete outcome (e.g., reading difficulty/no reading difficulty) for individual cases can be predicted from a set of continuous, categorical, and/or dichotomous predictors (Tabachnick & Fidel, 2012).

Although logistic regression can provide information about the predictive validity of a measure, it does not provide information about how accurately the measure is able to predict the classification of an individual as having a specific condition (e.g., reading difficulty). In this regard, Jenkins and colleagues recommend that “when selecting screening measures for early

identification, use classification accuracy as a selection guide" (Jenkins, Hudson, & Johnson, 2007, p. 584). Predictive classification accuracy refers to the ability of a measure to correctly predict a condition (e.g., reading difficulty). Common measures of predictive classification include sensitivity, which is the ability to correctly predict the existence of the target condition (i.e., true positives); specificity, which is the ability to correctly predict the absence of the target condition (i.e., true negatives); and area under the receiver operating curve (AUC), which examines the overall accuracy of a particular predictor measure by plotting the true-positive rate (i.e., sensitivity), against the true-negative rate (i.e., 1-specificity) for all possible cut points. In general, within the field of educational testing and testing for risk of reading difficulty in particular, sensitivity and specificity rates of 80% are considered to be the minimum acceptable, and even higher rates of sensitivity are usually considered preferable (Bridges & Catts, 2011; Compton, Fuchs, Fuchs, & Bryant, 2006; Petersen et al., 2016). In the case of an AUC, a value greater than .80 is considered good, and greater than .90, excellent (Catts, Nielsen, Bridges, Liu, & Bontempo, 2015; Compton et al., 2006; Hosmer, Lemeshow, & Sturdivant, 2013).

The sections that follow examine the literature related to the relative construct validity, predictive validity, and classification accuracy of static and dynamic measures. The focus is specifically on research in the field of reading and in particular, prediction of reading difficulties in young children.

Construct validity

Within the domain of reading, only a handful of studies have aimed to validate dynamic assessment as measuring a distinct construct to that measured by static assessment. Fuchs and colleagues (2011) examined the construct and predictive validity of a dynamic assessment of decoding. They used exploratory factor analysis on the prediction battery which included dynamic assessment; static assessments of alphabetic knowledge, RAN, phonemic awareness,

oral vocabulary, and listening comprehension; and teacher ratings of attentive behaviour and hyperactive or impulsive behaviour. The results of the analysis showed that dynamic assessment loaded on a first factor along with language and IQ. However, it was relatively distinct from speeded alphabetic knowledge and RAN, as well as task-oriented behaviour. In their study using a dynamic assessment of phonological working memory and semantic memory, Swanson and Howard (2005) found some support for dynamic assessment as measuring learning potential distinct from verbal IQ and a static measure of working memory. Cho and Compton (2015) conducted a study to determine whether a dynamic assessment of decoding measured early reading learning potential differently from static decoding and intelligence tests. Their study provided support for the hypothesis that dynamic assessment measures early reading learning potential distinct from that measured by assessments of general intelligence and current (actual) decoding skill. Furthermore, in their study using a dynamic assessment of phonological decoding to predict risk for reading difficulty in a group of bilingual Latino kindergarten children, Petersen and Gillam (2015) found that the children's language ability (both in their first and second language) did not relate to the results of the dynamic assessment (modifiability score), providing some support for modifiability being a separate construct to that of language ability.

With regard to threats to construct validity, dynamic assessment differs from static assessment in how it views test-retest effects. In the case of static assessments, test-retest effects are unwanted and considered a threat to construct validity. However, in the case of dynamic assessment, changes in scores due to assessor input are expected and, as an indicator of learning potential, can actually be seen as increasing construct validity (Carlson & Wiedl, 1979; Petersen & Gillam, 2015; Robinson-Zañartu & Carlson, 2013; Swanson & Howard, 2005).

Predictive validity

In this section, some pertinent research studies related to the relative predictive validity of dynamic assessment are reviewed. In a meta-analysis of 24 dynamic assessment studies, Caffrey and colleagues (2008) found that the predictive validity of dynamic assessment varied depending on the specific academic domain, but that in the domain of reading, dynamic assessment measures explained significant variance in the prediction of reading achievement. Since then, there has been increasing empirical evidence of the predictive validity of dynamic assessment in terms of reading achievement and difficulties (e.g., Bridges & Catts, 2011; Cho & Compton, 2015; Cho et al., 2014; Compton et al., 2010; Fuchs et al., 2011; Petersen & Gillam, 2015; Petersen et al., 2016; Stevenson, Bergwerff, Heiser, & Resing, 2014).

In one of the first studies to investigate the ability of dynamic assessment to predict beginning reading progress, Spector (1992) assessed a group of emergent readers using static assessments of receptive vocabulary, letter and word recognition, and phonemic awareness, as well as a dynamic measure of phoneme segmentation. Almost a year later, the emergent readers were retested on reading, spelling, and phonemic awareness. The study found that when the outcome measures of phonemic awareness and word recognition were regressed on the predictive measures, dynamic phoneme segmentation accounted for a significant amount of variance. Bridges and Catts (2011) investigated the predictive validity of a dynamic assessment of phonological awareness and found that when compared to a static version of the same test and to other commonly used static measures of phonological awareness, the dynamic assessment accounted for a significant amount of unique variance. Dynamic assessment of decoding has also been shown to contribute unique variance to outcome measures of word identification and reading comprehension achievement beyond that explained by widely used static tools used to predict reading development such as alphabetic knowledge, phonemic awareness and RAN (Fuchs et al., 2011). It has also been found to

uniquely explain a small but significant additional variance in word reading growth beyond that accounted for by several static measures including those of decoding, phonological awareness, RAN, and IQ (Cho et al., 2014; Cho & Compton, 2015).

Predictive classification accuracy

Although most studies have found that dynamic assessment accounts for a significant but small amount of variance over and above that of static measures, as Cho and colleagues (2014) point out, this does not necessarily indicate the extent of the practical significance of dynamic assessment. They suggest that examining the classification accuracy of a dynamic assessment (particularly its sensitivity and specificity) could be one way of making such a determination. Until very recently, researchers have advocated that dynamic assessment should not be used on its own as a tool to predict future achievement but should instead be used to supplement static assessment (Caffrey et al., 2008). In particular, they recommend the use of dynamic assessment as a possible way to improve the predictive validity of highly sensitive static tools (O'Connor & Jenkins, 1999; Grigorenko, 2009; Gustafson et al., 2014). As mentioned earlier, static assessments tend to produce a significant number of false positives (over-identification of reading difficulty) because they do not consider the possible effects of experiential or instructional differences, or the cultural and linguistic backgrounds of the participants. As dynamic assessment includes instruction as part of the assessment, it has been shown to reduce linguistic and cultural bias, and to improve specificity. Several studies have shown that dynamic assessment measures can add to the classification accuracy of early reading achievement, particularly in terms of reducing the number of false positives (Bridges & Catts, 2011; O'Connor & Jenkins, 1999; Spector, 1992).

In contrast to an approach where dynamic assessment is viewed as supplementary to static assessment, research by Bridges and Catts (2011) raises the possibility that a dynamic screening of phonological awareness may be best used as the primary measure to predict for risk of reading difficulties. While they found that a dynamic measure of phonological awareness significantly reduced the number of false positives produced by a static screening tool, they found that nearly the same results could be achieved using the dynamic measure alone.

Three studies by Petersen and colleagues (Petersen & Gillam, 2015; Petersen et al., 2016; Petersen et al., 2018) provide support for this possibility that a dynamic measure, on its own, may yield superior classification accuracy over static measures. In the first of these studies, Petersen & Gillam (2015) investigated the predictive validity of a dynamic assessment of decoding at the beginning of formal schooling to identify risk of later reading difficulty. They assessed a relatively small sample (n = 63) of bilingual (Spanish/English) learners from low socioeconomic backgrounds who were just starting formal schooling (mean age = 5.25 years), and who were pre-screened for risk of language impairment before they started schooling. At the beginning of their first year of formal schooling, the participants completed a dynamic pseudoword decoding task as well as two widely used standardised static measures of phonological awareness and alphabetic knowledge. Close to the end of their second year at school, the participants were tested again with commonly used standardised measures of pseudoword decoding, oral reading, and sight word identification. In analysing the relative classification accuracy of the dynamic and static predictive measures, Petersen and Gillam found that the static measure classified more than 95% (60/63) of the participants as being at risk of reading difficulty. They indicated that they did not analyse these results further as when "a measure classifies nearly all children as being at risk, the inevitable result will be near-perfect sensitivity with the alternating near-imperfect specificity" (Petersen & Gillam, 2015, p. 15). In contrast, they found that using a dynamic assessment yielded high classification accuracy with both sensitivity and specificity of greater than 80% in relation to all three outcome

measures. The dynamic assessment yielded 100% sensitivity for both the oral reading fluency (ORF) and sight word identification (WID) measures, along with specificity of 88% and 80% respectively. Although the sensitivity for the pseudoword reading criteria was slightly lower (86%), with a specificity of 85%, it still comfortably meets the minimum acceptable levels for sensitivity and specificity accepted by many (e.g., Boan, Aydlett, & Multunas, 2007; Carran & Scott, 1992; Plante & Vance, 1994).

In a later study, Petersen and colleagues (2016) used an adapted version of the aforementioned dynamic assessment of pseudoword decoding to assess for risk of future reading difficulties in an unscreened large sample (n = 600) of kindergarten children (mean age = 5.5 years) with mixed cultural, linguistic and socioeconomic backgrounds. As in the earlier study, Petersen and colleagues compared the classification accuracy of two static measures (phonological awareness and alphabetic knowledge) with that of the dynamic assessment of pseudoword decoding. In this case, two different strategies were employed in the input phase of the dynamic assessment, with one group of participants being taught to decode using a sound-by-sound strategy and the other being taught an onset-rime decoding strategy. Towards the end of the participants' second year in formal schooling, they were assessed again using five different static criterion measures of: sight word reading, pseudoword reading, nonsense word correct letter sounds fluency, phonemic awareness, and alphabetic knowledge.

The results showed that the static predictive measures yielded numerous false positives with specificity levels around 50% for the entire sample and even lower specificity levels for the English L2 participants. Sensitivity levels were also below 80% for the entire sample. However, the dynamic assessment measures yielded better sensitivity levels (81% and 92% for the onset-rime and sound-by-sound group respectively) for the entire sample, and even higher for the English L2 participants. Often, improved sensitivity results in lower specificity, but in the case of the dynamic assessment measures, the specificity levels were also higher than those yielded by the static measures: 81%

(onset-rime group) and 88% (sound-by-sound group) for the entire sample. Petersen and colleagues (2016) then also looked at whether combining the static and dynamic measures would improve the classification accuracy compared to the dynamic measures alone. They found that combining the measures did not significantly increase the predictive classification accuracy of the dynamic assessments alone. This would appear to provide some support to the suggestion made by Bridges and Catts (2011) that dynamic assessment could possibly be used as the primary measure for assessing risk of reading difficulties.

In a 2018 follow-up study, Petersen and colleagues (2018) followed the same children from their 2016 study through to the fifth grade. The results indicated that the dynamic assessment administered during kindergarten was able to predict, with a good level of accuracy, which children would have difficulty decoding up to six years later.

There are several possible reasons why dynamic assessment, in the form employed by Bridges and Catts (2011) and Petersen and colleagues (2015, 2016, 2018) may be better able to accurately predict future reading performance than static measures or other implementations of dynamic assessment. Firstly, in comparison to static assessment which provides a snapshot of a learner's level of performance at a particular point in time and is only interested in measuring the *product* of student learning, dynamic assessment is interested in both the *product* and the *process* of student learning. It does not only ask "What can this learner do now?", but also "What could this learner do in future if they were given high quality input?" and "How well does this learner respond to input?" (Grigorenko, 2009; Gustafson et al., 2014; Lidz & Peña, 2009; Petersen & Gillam, 2015). Those children who show greater response to input during dynamic assessment are presumed to be more likely to benefit from classroom instruction and therefore show higher levels of reading achievement. As such, dynamic assessment may be better than static measures at predicting future performance (Caffrey et al., 2008; Spector, 1992).

Secondly, in the case of children who have not yet received formal reading instruction, static measures are only able to assess emergent literacy skills such as phonological awareness and letter identification. In contrast, dynamic assessment can be used to measure skills that the child has not yet acquired (but which are within the child's ZPD) because it includes a teaching component within the test. For example, children beginning reading instruction can be taught a skill such as decoding and the child's response to the teaching given can be measured. For this reason, dynamic assessment can include content (such as decoding pseudowords or real words) that more closely corresponds to reading itself (Petersen et al., 2016).

Thirdly, as mentioned earlier, research has shown that static assessment tends to produce a significant number of false positives (over-identification of reading difficulty) because it does not take account of the possible effects of experiential deficits or the different cultural and linguistic backgrounds of the children (Bridges & Catts, 2011; Camilleri & Botting, 2013; Petersen & Gillam, 2015; Plante & Vance, 1994; Robinson-Zañartu & Carlson, 2013; Vellutino et al., 1996; Vellutino, Scanlon, Small, & Fanuele, 2006). As Petersen and Gillam (2015) point out, for children learning to read in an additional language, static assessments will not only reflect their ability to read in that language, but also their experiential variance in their first language reading. There is abundant evidence that a child's first language reading or early literacy skills significantly affect reading in an additional language (e.g., Edele & Stanat, 2015; Hoti et al., 2011; Koda, 2007; Sparks, Patton, Ganschow, & Humbach, 2012). As children learning English as an additional language are not a homogeneous group, and have very different first language backgrounds and experiences, this will impact on static assessment results which will reflect the child's ability to transfer their first language knowledge to an additional language through cross-linguistic transfer. Dynamic assessment essentially circumvents differences in first language reading or emergent reading skills by shifting the assessment focus from "the construct of what a child currently knows to the construct of what

a child can learn – a construct that is potentially independent of test content and prior knowledge" (Petersen & Gillam, 2015, p. 5). It is measuring the construct of modifiability, which is less influenced by socioeconomic, linguistic, and prior learning experiences (de Beer, 2010). As dynamic assessment includes instruction as part of the assessment, it has been shown to reduce linguistic and cultural bias, and to improve specificity (Bridges & Catts, 2011; Elbro, Daugaard, & Gellert, 2012; Grigorenko, 2009; Lidz & Haywood, 2014; Lidz & Peña, 2009; Petersen & Gillam, 2015). These findings are of particular interest in the New Zealand context where a significant percentage of children learn to read in a language (English or Māori) that they do not speak frequently at home[1], and the population is becoming increasingly culturally diverse (Statistics New Zealand, 2013).

A likely reason that the dynamic assessments used by Bridges and Catts (2011) and Petersen and colleagues (2015, 2016, 2018) were better able to accurately predict future reading outcomes than other implementations of dynamic assessments is the specific approach taken. The dynamic measures used standardised input in the form of scripted graduated prompts (Bridges & Catts, 2011) and scripted instruction (Petersen et al., 2016; Petersen & Gillam, 2015); focussed on measuring modifiability in terms of responsiveness to input (Petersen et al., 2016; Petersen & Gillam, 2015) and extent of input needed (Bridges & Catts, 2011); and was domain-specific, assessing phonological awareness (Bridges & Catts, 2011) and phonological decoding (Petersen et al., 2016; Petersen & Gillam, 2015). As discussed earlier, domain-specific dynamic assessments which use standardised input, and focus on measuring modifiability may have better predictive ability than those which are domain-general, use contingent input, and focus on pre-test post-test gain scores (Caffrey et al., 2008).

[1] In the 2010/2011 Progress in International Reading Literacy Study (PIRL 2010/11), 26% of Year 5 students reported that they only "sometimes" or "never" spoke the test language at home (Chamberlain, July 2013)

Acceptability

While validity and classification accuracy are key considerations related to the measurement quality of a screening tool for reading difficulties, the acceptability of the tool is crucial to ensure the tool is widely adopted in an applied setting (O'Donnell & Miller, 2011; van der Vleuten, 1996). Acceptability refers to the perception that a given tool is agreeable or palatable to key stakeholders such as the teachers and schools that could potentially utilize the tool (Proctor et al., 2011). The validity and classification accuracy of a tool will impact its acceptability as teachers and schools need to have confidence that the tool is able to accurately predict which children are at risk so that intervention resources can be directed to those children who really need it. Furthermore, the usability of a tool in a particular setting will affect how likely the tool is to be accepted and therefore used in that setting. Usability includes considerations such as the cost of the tool, the training required to learn how to administer the tool, the ease of test administration and scoring, and whether any specialized equipment is needed (Brown, 1996). To date there has been no specific acceptability research conducted specific to the use of dynamic assessment as a screening tool for reading difficulties.

Section 5: The current study

This section aims to frame the current study in terms of the literature most closely related to the study and highlight the gaps which this study aims to address. The literature reviewed has highlighted the importance of early intervention to ameliorate and even prevent the development of reading difficulties in children. However, early intervention is contingent upon the early and accurate identification of reading difficulty risk in children, preferably even before they begin formal reading instruction. In reviewing theory and previous research related to the key predictors of future reading at this early stage of reading development, clear trends emerge: The key skills to be assessed are phonological decoding ability and the emergent literacy skills on which this depends, namely alphabetic knowledge and phonemic awareness, as well as rapid naming, which has been found to be a good predictor of future reading ability.

While traditionally static tests of alphabet knowledge, phonemic awareness, and pseudoword decoding have been used to predict future reading status, these tests are frequently plagued by floor effects and poor classification accuracy. Dynamic assessment has shown some promise in addressing these issues, however, the use of dynamic assessment as a screening tool for risk of reading difficulties is relatively new and to date there have been very few longitudinal studies which examine the use of dynamic assessment with children before or at the start of formal schooling. Most of these studies have focussed on assessing phonological awareness skills (e.g., Bridges & Catts, 2011; Gellert & Elbro, 2018) or assessing decoding using artificial letters (i.e., symbols) rather than real letters (e.g., Gellert & Elbro, 2018). Given the central role of phonological recoding in the process of learning to read, it is evident that a dynamic assessment using phonological decoding of pseudowords may be a better indicator of future reading ability. Not only is the ability to phonological decode pseudowords more closely correlated with future reading outcomes than phonological awareness, but it is also within most children's zone of proximal development at the time they start school.

At the time of writing, there have only been three longitudinal studies that specifically investigated the use of dynamic assessment of decoding in 5-year olds to predict future reading difficulty, all conducted in the United States of America by Petersen and colleagues (2015, 2016, 2018). These studies were similar, with the 2015 and 2016 studies both focusing on the use of a dynamic assessment of decoding at the start of kindergarten, with follow-up assessment of the same children at the end of first grade to determine the ability of the dynamic assessment to predict reading difficulty. The 2018 study was a follow-up on the 2016 research in which the same children from the 2016 study were followed through to the fifth grade.

In the 2016 study, Petersen and colleagues focussed on comparing the classification accuracy of the dynamic assessment and commonly used static assessments of emergent literacy. For the dynamic assessment, Petersen and colleagues used a continuous score for their logistic regression and receiver operating characteristic analyses. This was calculated by adding the score from the learning rating scale (the assessor's rating of the participant's responsiveness to teaching) and the total post-test strategy score (which reflects the extent of the participant's use of the strategy they were taught to decode each pseudoword). Furthermore, a dichotomous score that provided a classification of at risk or not a risk for each child was calculated using the child's learning score (as an indication of the assessor's judgement of the child's modifiability), sound gain scores, and strategy scores, with the learning score being weighted heavily. The findings from the study indicated that the dynamic assessment of decoding had high levels of predictive validity and classification accuracy, and that predictive ability and classification accuracy was significantly superior to that of static measures of emergent literacy. The results from the follow-up (Petersen et al., 2018) study showed that this dynamic assessment administered at the start of schooling was able to accurately predict reading difficulty up to the fifth grade (Year 6 in New Zealand).

A relatively large proportion of the participants in the studies by Petersen and colleagues (2015, 2016, 2018) were Spanish-speaking and used English as an additional language. A representative sample of the New Zealand population would have a much smaller proportion of participants for whom English is an additional language, and there would be a mix of first language (e.g., Māori, Fijian, Mandarin, Afrikaans, Hindi, and so on). Furthermore, these studies were conducted in the United States of America, which as a very different education system to that of New Zealand, and where, unlike New Zealand, universal testing and screening for reading difficulty is routine. The question remains whether, using a protocol like that developed by Petersen and colleagues (2015, 2016), a dynamic assessment of decoding would demonstrate similarly high levels of predictive validity and classification accuracy when used with a very different demographic sample. Also, as Petersen and colleagues (2016) point out, although the findings from their use of a sound-by-sound dynamic assessment of decoding appear promising, theirs was the first study to use such a strategy and the results need to be replicated, particularly in terms of the way in which the *continuous* and *dichotomous* scores were obtained from the dynamic assessment. In this regard it is possible that refinements to the scale used by Petersen and colleagues (2016) to guide assessor's judgement of each child's learning could improve this measure's predictive validity and classification accuracy. The scale used by Petersen and colleagues (2016) includes only four focus areas: errors, confidence, disruptions, and the rate at which the child completes the decoding task. The addition of other elements associated with learning ability may be beneficial in better guiding the assessor to a final judgement of the child's learning.

Furthermore, Petersen and colleagues (2016) pointed out that it may be possible to further improve the classification of a dynamic assessment measure of decoding in several ways. Firstly, by increasing the number of pseudowords in the test or increasing the time spent on the input (teaching) phase of the assessment. Secondly, by using pseudowords at post-test

with the same letters, but in a different order (e.g., *pog* … *gop*), to obtain a measure of generalisability. The child's ability to transfer their newly acquired skills to analogous tasks may provide useful information as it is a good predictor of how responsive they will be to instruction (Robinson-Zañartu & Carlson, 2013).

The current study aims to investigate whether a dynamic assessment of decoding, using procedures like those employed by Petersen and colleagues (2015, 2016), and administered to New Zealand children at the start of their formal schooling, can be used to predict their future reading status. Furthermore, this study will investigate whether, like the aforementioned studies, this dynamic assessment of decoding is able to produce superior predictive validity and classification accuracy to that of traditional, static assessments of the emergent literacy skills of alphabet knowledge, phonemic awareness, and rapid naming.

Section 6: Summary

A review of the literature related to the prediction of risk of reading difficulty in children has revealed the following key points. Early prediction of reading difficulty, even before a child has started formal instruction, is important as it allows for early intervention, which in turn affords the most promising prospect for reading difficulties to be ameliorated, or even prevented. There is substantial research that indicates that such early prediction of reading difficulty is viable, using tests of emergent and early literacy skills, in particular phonological awareness, alphabet knowledge, rapid naming/fluency, and phonological decoding.

To date, the dominant form of reading difficulty assessment is static; however, this method of assessment has several limitations, including that static measures frequently suffer from floor effects and poor classification accuracy (sensitivity and specificity). To address these limitations, there has been growing interest and research into the use of other forms of assessment, including RTI and dynamic assessment, both of which include an element of intervention (teacher input) within the assessment, helping to reduce the prevalence of floor effects, and improve classification accuracy. However, dynamic assessment has several additional advantages over RTI, most notably that it is quick and easy to administer before children start formal reading instruction. This means that targeted and intensive intervention can be provided to those children needing it as early as possible rather than waiting several months (as is the case with response to intervention with its three-tier, multiple month approach).

However, to date, there has been very few longitudinal studies that focus on the ability of a dynamic assessment, administered to learners at the start of formal schooling, to predict future reading difficulty, and no such studies conducted in the New Zealand school context. This study aims to address this gap and to add to the existing body of evidence pertaining to the use of dynamic assessment as a screening tool for reading difficulty.

Chapter 3:
Methodology

This study investigated the use of static and dynamic assessments to predict the risk of reading difficulty. A longitudinal design was used to explore whether static and dynamic measures administered at the start of schooling were significantly correlated with reading ability at the end of one year at school, and to investigate and compare the predictive validity and classification accuracy of these measures in terms of reading difficulty status.

As outlined in the previous chapter, there is a significant body of research that highlights the importance of early diagnosis and intervention of reading difficulties, and growing evidence of clear and strong correlations between the emergent and early literacy skills of alphabet knowledge, phonological awareness, rapid naming, and decoding, and later reading ability. However, static assessments of emergent and early literacy skills commonly suffer from inadequate classification accuracy and in particular, high rates of false positives (poor specificity), with children incorrectly identified as being at risk for later reading difficulty. In response, there have been several studies that have investigated the use of dynamic assessment to address the shortcomings of traditional static assessments. However, there has been very limited research focussing specifically on children at the start of their formal schooling and with pseudoword decoding as the target skill being tested. Furthermore, there has been no research to date on the use of dynamic assessments to screen for reading difficulty risk in New Zealand children upon school entry. This study aimed to contribute evidence-based research on the use of universal screening tools to predict the risk of reading difficulty within the New Zealand context, and to address the current research gaps in this area.

In this chapter, the aim of this study and the research design are outlined. This includes details of the participants, the assessment measures employed, and the procedures for their administration and scoring. A brief overview is also given of the statistical analyses employed in the study. Additional details about the analyses are provided in Chapter 4.

Research aim/purpose

The purpose of this study was to investigate the use of a dynamic assessment tool with children at the start of their formal schooling to predict reading difficulty after a year at school. In order to do so, the following were examined:

1. The relationship between static and dynamic reading measures, administered at the beginning of formal schooling, and reading performance at the end of a year at school.

2. The ability of static and dynamic reading measures, administered at the beginning of formal schooling, to predict future reading difficulty.

3. The difference in predictive classification accuracy between the static and dynamic measures in terms of future reading difficulty.

4. Whether there is an increase in the predictive ability and classification accuracy when the static and dynamic measures are combined, versus used on their own.

5. The predictive ability and classification accuracy of the dichotomised (*at risk*, *not at risk*) scores for the static and dynamic predictor measures to provide information that can be readily understood and easily applied in a school setting.

Research design

This study employed a prospective longitudinal study that examined and compared the predictive ability and classification accuracy of static and dynamic measures of emergent and early literacy, administered to New Zealand children upon school entry. At the start of their first year of formal schooling, children were administered a battery of static measures of emergent literacy, as well as a dynamic assessment of pseudoword decoding. At the end of their first year at school, the children's reading ability was re-assessed using a battery of static measures, as well as school evaluations of their reading level.

As discussed in the preceding chapter, there has been a growing emphasis on early identification and intervention of reading difficulties, particularly at preschool level (Bailet, Repper, Murphy, Piasta, & Zettler-Greeley, 2011; Catts et al., 2014; Kantor et al., 2011). However, for the purposes of this current study, administering the predictor measures during preschool would have caused logistical difficulties. In New Zealand, children attending a particular early childhood education centre will go on to attend any number of different primary schools. This makes tracking and follow up of the children difficult as they move from preschool to primary school and would likely result in an unacceptably high rate of attrition. Instead, it was decided to administer the predictor measures as soon as possible after children had commenced formal schooling. This made it easier to follow the same children throughout the study and at the same time minimise the possible impact of curriculum and instructional differences as a confounding factor. Furthermore, in New Zealand, children do not all start school on the same date. Instead, most children begin school on or close to their fifth birthday. Therefore, an additional advantage of testing the children as soon as possible after they enter school is that the confounding influence of age is reduced.

As ancillary to the main study, exploratory work was undertaken to determine the scoring procedure for the dynamic assessment in the form of pilot testing and statistical analysis. Details regarding this work are provided in Appendix A, Section A.1.

Ethical considerations

Guided by discussions with the supervisory panel for this doctoral study, as well as the Massey University *Code of ethical conduct for research, teaching and evaluations involving human participants* (2010), several ethical implications were considered. These centred around issues of informed consent; maintenance of confidentiality; risks of harm; ownership of data; and the principles of the Treaty of Waitangi.

As the participants in the study were children under the age of seven, written parental consent was sought for each child to participate (see Appendix B for the parent/caregiver information and consent forms). Oral assent for participation was also sought on an iterative basis from all children involved in the study. For example, "I'm Su and I want to find out about your reading. Is it okay if we do some reading activities together? You don't have to if you don't want to."; "Can we do one more reading activity together? You don't have to if you don't want to."; and so on.

Each child participating in the study was assigned an alphanumeric study code (study ID) prior to data collection. A file listing each participant's name and their unique study ID (e.g., BKO936415) was stored separately from data documents. When collecting data, only the participant's unique study ID was used on all data documents and data collected. All data was stored in electronic form and password protected. Consent forms and hard copies of data documents were stored in a locked filing cabinet, with data and consent forms stored separately.

Consideration was given to the fact that some children might feel anxious about talking to someone they don't know and/or participating in a new situation (the assessment). To address this potential concern, the researcher asked teachers to introduce her to the children in the class and where possible to allow her to spend some time interacting with the children during normal class activities before the testing session. This gave children the opportunity to become familiar with the researcher before assessment began. Another potential issue was that children might start to feel tired, frustrated, or otherwise not willing to participate in one or more of the assessments, especially if they perceived it as difficult. To address this, it was decided that while children would be encouraged to continue to participate in the assessment, if they indicated they were too tired or no longer willing to participate, the assessment would be discontinued, and an attempt made to assess the child later.

The research considered the fact that teachers may want to be given results from testing of the children in their classes. Before testing began, the researcher explained to teachers that she would not be able to give them this information for two key reasons: (1) It could impact on their teaching practices and thereby affect the results obtained from the criterion measures; and (2) it would be unethical to give teachers information based on an assessment that is still being evaluated in terms of its predictive validity and accuracy (the dynamic assessment).

Careful consideration was given to the principles of the Treaty of Waitangi and the researcher liaised with a senior Māori adviser at Massey University who agreed to provide support and advice as necessary, including arranging and facilitating hui for whānau if needed.

Before proceeding with the study, approval for the research was obtained from the Massey University Ethics Committee: Southern B, Application 16/27. See Appendix D for a copy of the Ethics Committee approval.

Recruitment and participants

The schools

Ten schools in North and West Auckland agreed to participate in this research (see Appendix C for information and consent forms for schools and teachers). These schools represent a range of school deciles: from decile 2 to decile 9. In New Zealand, school deciles indicate the extent to which schools draw their students from low socio-economic (SES) communities (New Zealand Ministry of Education, 2015). The five socio-economic indicators are household income, whether income support is being provided, crowded households, parents with no qualifications, and parents with low-skilled occupations. Decile 1 schools are the 10% of schools that draw the greatest proportion of their students from low SES neighbourhoods, whereas the 10% of schools with the lowest proportion of students from low SES communities are decile 10. It should be noted that the schools' decile does not indicate the overall SES mix of the school (New Zealand Ministry of Education, 2015) and all the decile 5, 6, and 7 schools that participated in this study draw students from a range of socio-economic groups: including those coming from homes that could be classified as socio-economically disadvantaged, as well as those that could be classified as socio-economically privileged.

The children

Start-of-school (T1) – Predictor measures

All children starting school during Term 4 of 2016 and Term 1 of 2017 were invited to participate in the study. Teachers at the participating schools were asked to give information sheets and consent forms to the parents of all children starting school during this period. Consent forms were received for 165 children, and all these children were included in the study.

No children were excluded from the study as the intent was to mirror the reality in classrooms as closely as possible, where all children aged five and over can attend school, regardless of whether they have suspected or diagnosed learning, behavioural, or other difficulties, and irrespective of their ethnicity, home language, or other demographic characteristics. Furthermore, this research aimed to establish how well static and dynamic measures predicted further reading ability at all ability levels, not only for children who may be identified as having a higher probability of having a future reading difficulty.

Age at testing

Although many children in New Zealand begin school on or very soon after the day of their fifth birthday, most of the participant schools had a 'settling in' period for children before beginning formal reading instruction. Testing was arranged to take place no more than 4 weeks after children had started formal reading instruction in schools. For the T1 predictor testing, most children were between 5 years and 5 years 2 months, with a mean of 5 years 2 months. In all cases, children who were older than 5 years 2 months at the time of testing had either started schooling later than is the norm (e.g., some months after they turned five), or they had remained in a 'transition class' and had not yet started formal reading instruction.

Gender

Slightly more boys (53.9%) than girls (46.1%) participated in the study. This somewhat reflects the gender distribution for all schools in Auckland in 2016: 51.4% male; 48.6% female, as at 1 July 2016 (New Zealand Ministry of Education, n.d. -a).

Language and ethnicity

Parents were asked to indicate the main language(s) spoken in the child's home and the ethnic group(s) to which they feel their child belongs.

English was identified as the main language for the majority of the children (74.0%), and 1.2% identified Māori as the main language spoken at home. An Asian language (Chinese, Korean, Japanese, etc) was identified as the main language for 17.6%, a European language (e.g., German, Russian) was identified as their first language by 5.0% of children, and 2.2% identified a Polynesian language (e.g., Samoan) as their main language.

The majority (50.5%) identified as NZ European, followed by 20.8% who identified as Asian, 10.4% who identified as NZ Māori, 6.6% as Other European, 6.1% as Pasifika, and 4.2% as Middle Eastern/Latin American/African (MELAA). This somewhat reflects the ethnic breakdown of the West and North Auckland regions, which together average at NZ European = 42.1%; Asian = 26.3%; NZ Māori = 14.5%; Pasifika = 10.8%; Other = 5.6% (New Zealand Ministry of Education, n.d. -a).

After one year at school (T2) – Outcome measures

A year later, 136 of the children tested at T1 were available for testing, representing an attrition rate of 17.6%. A comparison of the children who were tested at both T1 and T2 (longitudinal group) and those who were only available for testing at T1 (attrition group) is provided in Chapter 4.

Measures and procedures

Once parental consent had been obtained, children were tested individually during school time. The testing took place in a quiet room to reduce the likelihood of external distractions, and the testing sessions were video recorded to allow for later verification of test scoring. All testing was conducted by the author of this thesis, a doctoral candidate with several years of assessment experience. Training consisted of familiarisation with the administration and scoring rules for the different measures, as well as practice administration during pilot testing (refer to Appendix A). To follow is a description of the predictor measures administered at the start of formal schooling (T1), and the outcome measures administered after a year at school (T1).

Start-of-school predictor measures (T1)

Table 3.1 summarises the measures administered at the start of formal schooling. The dynamic assessment of pseudoword decoding was conducted first, followed by the static measures which were administered in the order shown in the table. For ease of reading, in this section, the static measures will be described prior to the description of the dynamic measure.

Table 3.1: Summary of start-of-school (T1) predictor measures

Static measures	
CTOPP-2 RON[a]	Comprehensive test of Phonological Processing (2nd edition) – Rapid Object Naming
DIBELS Next[b]	Dynamic Indicators of Basic Early Literacy Skills (DIBELS)
• LNF	• Letter Naming Fluency
• FSF	• First Sound Fluency
Dynamic measure	
DA	Dynamic Assessment of Pseudoword Decoding

[a] Wagner et al., 2013
[b] Good & Kaminski, 2011

Static predictor measures

Each child was administered static measures of rapid object naming, alphabet knowledge, and phonemic awareness. As discussed in the Literature Review, all three of these emergent literacy skills have been shown to correlate significantly and strongly with later reading ability. The specific measures used have all been widely used in international studies, including those investigating the use of these measures as screening tools for future reading difficulty. The static measures used are discussed below, in the order in which they were administered.

CTOPP-2 RON

The *Comprehensive Test of Phonological Processing – Rapid Object Naming* (CTOPP-2 RON) is a rapid automatized naming task in which children are asked to name familiar objects. Children are shown six different familiar objects repeated six times in a pseudorandomized array (no object appearing consecutively on the same line). The child is asked to name as many objects as possible as quickly as they can, and the assessor times how long it takes the child. The number of seconds it takes the child is the *CTOPP-2 RON* raw score. If a child made more than four errors when naming objects, testing was discontinued, and the child was not awarded a score.

The *CTOPP-2 RON* raw score was converted to a scale score from 1 to 18 (as set out in the *CTOPP-2 Examiner's Manual*). Children who were not awarded a score because they made more than four errors were awarded a RON scale score of 1 (lowest possible scale score). It was decided to use the scale score, rather than other provided normative scores such as age and grade equivalents; this is because the scale score provides the clearest indication of a child's subtest performance (Wagner et al., 2013).

Descriptive terms corresponding to the scale scores are provided in the *CTOPP-2 Examiner's Manual*, with these ranging from *very poor* to *very superior* (Wagner et al., 2013). Only those classified as *poor* or *very poor* (scaled score of 5 or less) were classified as *at risk* for the purposes of this study.

Alternate form, test-retest, and scorer reliability coefficients for the *CTOPP-2 RON* were reported as .86, .86, and .96 respectively, indicating good to excellent reliability (Wagner et al., 2013). Criterion-prediction evidence of validity measured across six other measures of rapid naming produced a mean coefficient of .70 (Wagner et al., 2013).

DIBELS Next

Dynamic Indicators of Basic Early Literacy Skills (DIBELS) *Next* is a set of brief measures used to assess developing literacy and reading skills. The measures were developed in the United States of America for children from kindergarten (start-of-school in New Zealand) through to sixth grade (Year 7 in New Zealand). For the static predictor measures, the DIBELS Next kindergarten measures of *Letter Naming Fluency* and *First Sound Fluency* were used.

DIBELS Next Letter Naming Fluency (LNF) was used to assess alphabet knowledge (letter names). In this task, children are shown a page with upper- and lowercase letters in random order. They need to name as many letters as they can in one minute and are given one point for each letter correctly named within this time.

DIBELS Next First Sound Fluency (FSF) was used to assess phonemic awareness skills. This task requires children to distinguish and produce the initial sound in as many orally presented words as possible. The assessor presents the words orally for up to one minute, or until 30 stimulus words have been presented. Children are given one point for each initial sound correctly produced.

Excellent alternate form and inter-rater reliabilities were reported for both the LNF and FSF, with coefficients ranging from .85 to .99 (Dewey, Powell-Smith, Good, & Kaminski, 2015). However, there was weaker evidence of predictive validity with moderate to moderate-strong correlations ranging from .39 to .52 reported between the FSF and the external criterion measure *Group Reading Assessment and Diagnostic Evaluation – GRADE* administered at the end of the kindergarten year (Dewey et al., 2015).

The DIBELS Next Composite score is a combination of the LNF and FSF scores and has been shown to provide the most accurate indication of a learner's reading ability (Good et al., 2013).

Reliability coefficients for the Composite score ranged from .66 (alternate form) to .97 (inter-rater). The Composite score had a predictive validity coefficient of .50 with the GRADE criterion measure administered at the end of the kindergarten year, indicating moderate-strong correlation (Good et al., 2013).

An advantage of *DIBELS Next* is that it includes *benchmark goals* and *cut points for risk* for three different times at each year level (beginning, middle, end) (Good & Kaminski, 2011). The *benchmark goals* are target scores that indicate satisfactory reading progress. Children who achieve these goals are likely to achieve future reading goals without the need for additional support or intervention. The *cut points for risk* indicate "a level of skill below which the student is unlikely to achieve subsequent reading goals without receiving additional, targeted instructional support" (Good & Kaminski, 2011, p. 23). Based on the *benchmark goals* and *cut points for risk*, three classification categories for the *LNF* and *FSF* are indicated: *core*, *strategic*, and *intensive*.

A classification of *core* indicates the child has achieved the benchmark goal and is likely to achieve later reading outcomes if s/he receives core/standard classroom instruction (i.e., there is no need for additional support or intervention). A classification of *strategic* indicates that the child has not achieved the benchmark goal but is still above the cut point for risk. Children in this range should receive some additional, strategic support in the skill areas they are having difficulty with and should be monitored regularly to check that they are making adequate progress. Scores below the cut point for risk are classified as *intensive*. Children in the *intensive* range are likely to need *intensive* intervention such as small group or individual instruction, additional instruction time, and so on.

For the predictor measures, the beginning of kindergarten benchmark goals and cut points for risk were applied. Following the model of Petersen and colleagues (2016), only children classified as *intensive* on the DIBELS measures were classified as *at risk* for the purposes of this study. Although children in the *strategic* range could be considered at some risk, according to Good and Kaminski,

in this range, "a student's future performance is harder to predict" (2011, p. 24). Therefore, using the DIBELS Next *intensive* classification as an indicator of *at risk* status allows for the most conservative comparison with the dynamic assessment measures (Petersen et al., 2016).

Dynamic predictor measure

Dynamic assessment of pseudoword decoding

Within four weeks of children starting formal reading instruction, a dynamic assessment of pseudoword decoding was administered to each child individually. The administration followed a similar procedure to that used by Petersen and colleagues (2015, 2016) and consisted of three stages: (1) The pre-test stage where the assessor asked each child to decode six pseudowords; (2) the instruction stage where each child was given instruction on how to decode each of the pseudowords, using a sound-by-sound strategy; and (3) the post-test stage where the assessor asked the child to decode the same six pseudowords in a different order, as well as two new pseudowords, which consisted of the same letters but in a different order (e.g., *pes* from *sep*).

Pseudowords: Pseudowords are nonsense words that obey the phonotactic rules of a particular language. Pseudoword reading tasks are used to assess a child's ability to apply grapheme-phoneme knowledge to decode a word. They require the child to link the appropriate phoneme with each grapheme in the pseudoword and to blend these together to form a pronounceable "word". As discussed in Chapter 2, pseudoword decoding has been shown to be the best single predictor of word reading for both normal and poor readers, and difficulty in pseudoword reading has been shown to be the most reliable indicator of reading difficulties, particularly in early readers. The additional two pseudowords used in the post-test stage were added as a measure of generalisability. This was used to provide an indication of how successful the child was in applying the principle they had been taught to more complex problems, without assistance.

As English is the medium of reading instruction for the children, all the pseudowords used were English pseudowords (i.e., phonotactically legal within the context of English). The initial six pseudowords (used during all three stages of the test) followed the same consonant-vowel-consonant (CVC) pattern. The short vowel /e/ and final consonant /p/ were the same in each word, creating rhyming words (*nep, kep, tep, mep, sep, hep*).

Careful consideration was given to choosing sounds common in all the language groups likely to be present in the sample: English, te reo Māori, Pasifika languages (e.g., Fijian, Tongan), Mandarin, etc. For example, the /w/ sound is not present in many languages (e.g., Samoan, Tongan, Cook Island Māori, Fijian, Niuean, and Standard Chinese), and is rarely used in English words. Therefore, even though it is used frequently in te reo Māori, the /w/ sound was not included in any of the pseudowords in the test. In contrast, while there is no /s/ sound in te reo Māori, the sound was still included in one of the pseudowords because the /s/ sound is very common in English words. As a result of the sound's dominance in English and other languages, it is likely that Māori-speakers would already be familiar with the sound. Furthermore, as the /s/ sound was present in only one of the target words (*sep)* it would mean that, at most, the child would encounter one sound not native to their first language, and as such this should not have a major impact on the overall result.

It is recognised that many languages do not have consonant-vowel-consonant (CVC) as the basic structure for their syllables. However, as CVC is the dominant structure for the English words encountered by beginning learners, it was considered valid to use in this assessment. Rather than using a mix of CVC forms, it was decided to use an onset-rime pattern, with the rime (/ep/) used in all six words. This was done to support learning of what would be entirely new skills for most of the children, namely phonological decoding and blending. Keeping the final two phonemes constant reduced the complexity of the learning required in the single short session

and at this early stage in literacy development. The two new pseudowords given at post-test consisted of the same letters, but in a different order, as a measure of generalisability.

Pre-test stage: Children who, during the pre-test stage, correctly decoded three or more of the first four pseudowords, or nine or more of the phonemes within these words, did not continue to the instruction stage. Following the procedures used by Petersen and Gillam (2015), these children were considered unlikely to be at risk for future reading difficulty. These children still completed both the static predictor measures (at the start of school) and static outcome measures (at the end of their first year at school).

Instruction stage: As soon as the pre-test stage had been completed, each child who did not meet the discontinuation criteria during the pre-test stage was told that s/he was going to be taught how to read. The instruction stage consisted of two rounds of teaching during which the assessor utilised standardised input and feedback to teach each child individually how to decode the pre-test pseudowords. Research has shown that dynamic assessment that uses standardised input has superior predictive validity to those which use contingent input (Caffrey et al., 2008).

During the first round of instruction, each child was taught how to use a sound-by-sound strategy to decode the same six pseudowords from the pre-test. Petersen and colleagues (2016) compared the classification accuracy of a dynamic assessment using an onset-rime decoding strategy and one using a sound-by-sound strategy and found that the dynamic assessment using a sound-by-sound decoding strategy may produce superior classification accuracy. As a result, it was decided to employ a sound-by-sound decoding strategy in this study, rather than onset-rime.

The following procedure was applied for each of the six pseudowords (using the pseudoword tep as an example): The assessor pointed to the initial phoneme /t/ and said, “This letter says /t/. Say /t/.” Then the assessor pointed to each of the remaining phonemes in turn and said,

"This letter says /e/. Say /e/. This letter says /p/. Say /p/". The assessor then demonstrated how to blend each of the phonemes, "Put the letters together and you get the word *t-e-p*, *tep*. What is this word?".

The second round of instruction followed immediately after the first round. In the second round, the level of assessor scaffolding was reduced, and the child was asked to imitate the assessor's reading. Using the pseudoword *tep* as an example, the assessor said /t/ while pointing to the letter *t.* The child was asked to imitate this model by saying /t/. If the child failed to independently produce the target, s/he was prompted to do so: "Copy what I do" or "Do what I do". The assessor followed this model to say and point to each of the remaining letters in the word in turn (/e/ and /p/), and then to say each of the three letters in sequence (*t-e-p*) and finally the whole word (*tep*), with the child imitating the assessor each time.

Post-test stage: Immediately after the instruction stage, the post-test was administered. This consisted of the assessor showing the children the same words used in the pre-test and instruction stages, presented in a different order, along with the two new pseudowords. The children were asked to read the words. During this stage, the assessor did not give the children any assistance; however, where necessary, some encouraging prompts were used if a child did not respond or make any attempt to read the words. Examples include prompts such as "Remember how we read the words together", "What do you think this word says?", and so on.

An audio-visual recording was made of the entire dynamic testing session (all three stages) to allow for review of the session to support accurate scoring, and to allow the assessor enough reflection time to make judgements regarding the child's learning behaviours.

Dynamic assessment scoring

The scoring procedures used for the dynamic assessment were adapted from those used by Petersen and colleagues (2015, 2016) and included scores for the total number of correct sounds and correct words at pre-test and post-test, gain scores for sounds and words, as well as strategy, learning, and modifiability scores (refer to Appendix A for details of inter-rater reliability established during pilot testing).

Total number of correct sounds: The total number of correct sounds were calculated during both the pre-test and post-test stage of the assessment. Each pseudoword consisted of a total of three sounds, giving a maximum of 18 points possible at pre-test (six pseudowords with three sounds each) and a maximum of 24 points at post-test (eight pseudowords with three sounds each). Separate correct sounds scores were calculated for taught (application) words and for all words (both application and generalisation words). One point was awarded for a correct sound in the correct (initial, medial, final) position. For example, for the word *tep*, if the child said *lap*, they were awarded 1 point (correct /p/ sound in the final position); if the child said *top,* they were given 2 points (correct /t/ in the initial position and correct /p/ in final position).

Sound gain scores: Sound gain scores were calculated by subtracting the total number of correct sounds at pre-test from the total number of correct sounds at post-test. This was done for both taught (application) words and for all words (application and generalisation).

Total number of correct words: The total number of words read correctly was calculated during both the pre-test and post-test stage. A pseudoword was counted as read correctly if all three phonemes were produced in the correct sequence, and without any pauses. The word could either be read immediately (as a sight word) or first sounded out and then blended into a word immediately thereafter. For example, for the pseudoword *tep* the child was awarded three points if they responded "*t – e – p*" and then with, or without prompting produced "*tep*".

One point was awarded for each word read correctly, with a maximum total of 6 points possible at pre-test (taught words) and 2 additional points at post-test (generalisation words). Separate scores for both taught (application) and all words (application and generalisation) were calculated.

Word gain scores: Word gain scores were calculated by subtracting the total number of correct words at pre-test from the total number of correct words at post-test. This was done for both taught (application) words only, and for all words (application and generalisation).

Sound residuum gain scores: Following the procedures used by Petersen and Gillam (2015), a sound residuum gain score was calculated. Petersen and Gillam (2015) found that the sound residuum gain score was a better predictor of reading ability than the sound gain score. The sound residuum gain score is a percentage of the residuum (i.e., difference between the pre-test sound score and the ceiling) and is used to help account for prior knowledge. For example, a child who scored 8 at pre-test would have a sound residuum score of 10 (ceiling minus pre-test sound score: 18 – 8 = 10). If that child achieved a sound gain score of 10 at post-test, this would be a sound residuum gain score of 100% (sound gain/sound residuum score X 100: 10/10 X 100). A child who had a pre-test score of 0, would have a sound residuum score of 18. If, like the previously mentioned child, this child also had a sound gain score of 10, they would only have a residuum gain score of 56% (10/18 X 100). A sound residuum gain score was calculated on the taught (application) words only, as well as on both the application and generalisation words.

Strategy score: Using the Strategy Scale (Figure 3.1), the assessor rated each child from 1 to 5 on their application of the decoding strategy they were taught during the instruction stage (application) and their ability to transfer the learned strategy to new pseudowords (generalisation). The Strategy Scale was used to assess the reading strategies applied by children during the post-test phase, and is a measure of the child's ability to apply a taught decoding (phonological recoding) strategy and/or acquire orthographic representations rapidly with only limited exposure to the target words and phonemes (Ehri, 2014; Share & Shalev, 2004; Stuart et al., 2008).

Figure 3.1: Strategy Scale

5	4	3	2	1
Correct word (blended all three correct sounds in correct sequence, without pauses).	Decoding strategy used with all three correct sounds produced in correct sequence, but with pauses between them (i.e., no blending).	Decoding strategy used with two correct sounds produced in sequence (e.g., for the word *mep*, /m/-/e/, /me/, /e/-/p/, or /ep/) OR whole word with one incorrect sound (e.g., *mip* or *met*].	No evidence of strategy use. Correct sound in isolation or correct single sound(s) in correct position (e.g., for *mep*, /m/, /p/-/m/, /*mat/,* or /p/-/e/-/m/).	No evidence of strategy use. No evidence of initial sound learning. No response OR incorrect whole word guess OR random sound/letter name(s).

The Strategy Scale, adapted from Petersen and colleagues (2015, 2016) assigns different scores for the strategy used by the child during the post-test. Using the word *mep* as an example, a score of 1 was awarded if the child did not show any evidence of using the strategy taught and/or of rapidly forming orthographic representations; this may be evident by the child not responding, incorrectly guessing a whole word (e.g., *nod*), showing no evidence of initial sound learning (e.g., *nod*), or producing random (guessed) sounds and/or letter name(s) (e.g., /s/ or *s*). Where the child demonstrated no evidence of applying the sound-by-sound decoding strategy, but there was however some evidence of the expeditious acquisition of orthographic representations by producing a correct sound in isolation (e.g., /*m*/) or a word with a correct single sound in the correct position (e.g., *mat*), a score of 2 was given. A score of 3 was awarded if there was evidence of decoding strategy use and/or rapid acquisition of orthographic representations in the form of two correct sounds produced in sequence (e.g., /m/-/e/, /me/, /e/-/p/, or /ep/) or a whole word with one incorrect sound (e.g., *mip* or *met*]. If the child applied the decoding strategy and/or showed evidence of rapidly forming orthographic representations to produce all three sounds in the word correctly and in sequence, but without blending (e.g., /m/ - /i/ - /p/), a score of 4 was awarded. The maximum score of 5 was

awarded if the child correctly read the word, with all three sounds correctly produced in sequence and without any pauses between sounds (e.g., *mep*). The total strategy score (between 8 and 40) was divided by the total number of pseudowords (8) to get a mean score. This was used as the strategy score.

Learning score: The Learning Scale shown in Figure 3.2 was used to guide the assessor in assigning each child a learning score from 1 to 5. This score reflects the assessor's overall perception of the child's responsiveness to the standardised input during the instruction stage. The Learning Scale was adapted from a range of sources including Peña and Villareal's (2000) *Modifiability observation form* (reproduced in Mann, Peña, & Morgan, 2015); Gutierrez-Clellen and colleagues' *Learning behaviour scale* (Gutierrez-Clellen, Brown, Conboy, & Robinson-Zañartu, 1998), Petersen and colleagues' *Teaching responsiveness scales* (2016); Feuerstein and colleagues' list of cognitive deficiencies (2010); Vye, Burns, Delclos, and Bransford's list of behaviours used to compare cognitive approaches of children to teaching (as cited in Lidz, 1991); Lidz's *Response to mediation scale* (2003); and literature reviews on response inhibition (particularly Baggetta & Alexander, 2016; Yeager & Yeager, 2013). The assessor rated each child from 1 to 5 in terms of their internal social-emotional behaviours (anxiety, perseverance, motivation); external social-emotional behaviours (attention, tractability, task confidence); and cognitive arousal (task comprehension, errors). The scores were used to provide areas of focus to support the assessor in arriving at their overall judgement of how well the child responded to input (learning score).

Figure 3.2: Learning Scale

	5	4	3	2	1
			Internal social-emotional		
Anxiety	Calm	Some signs of slight nervousness.	Uncomfortable	Distressed	Distraught
Persever-ance	Consistently perseveres and persists with task (even if frustrated).	Generally persists with task, with **minimal** encouragement from assessor.	Some signs of apathy, indifference OR unwillingness to continue with task. However, persists with some encouragement from assessor.	Frequent signs of apathy, indifference OR unwillingness to continue. However, persists with **significant** encouragement from assessor.	Refuses to attempt any aspect of task.
Motivation	Shows interest and willingness to try tasks throughout session.	Willingness to try **most** tasks throughout session (even if experiencing difficulty).	Ambivalent towards tasks.	**Occasional** attempts to end session OR to avoid task(s) OR occasional expression of dislike of tasks.	Refuses to participate OR avoidant behaviours OR consistent expression of dislike of tasks.
			External social-emotional		
Attention	Attentive, focussed, on-task behaviour without prompting.	Maintains attention with minimal prompting.	Distractible but can refocus with prompting and/or repetition.	Distracted, difficult to refocus.	Highly distractible; unable to maintain attention to task; frequent off-task behaviours.
Tractabil-ity	Cooperative and responsive throughout.	Consistently responsive.	Responsive for some tasks and/or hesitant.	Passive noncompliant or uncooperative.	Resistive/refuses to cooperate or respond.
Task confi-dence[2]	Consistently demonstrates high confidence behaviours.	Demonstrates high confidence behaviours much of the time.	Demonstrates a mix of high and low confidence behaviours.	Regularly seeks confirmation from assessor or pauses.	Frequently demonstrates low confidence behaviours.
			Cognitive arousal		
Task compre-hension	Quickly understands all aspects of what is required.	Average comprehension of what is required.	Slow to comprehend, but eventually understands what is required.	Only rudimentary understanding of what is required.	No evidence of task comprehension.
Errors[3]	No errors.	Single error made. May repeat the same error.	Some errors (2 or 3 errors). May repeat the same errors.	Frequently makes errors (4 or 5 errors). May repeat the same errors.	Very frequently makes errors (including repeating the same errors).
			Learning score		
Learning	Easy	Quite easy	Moderate	Quite difficult	Difficult

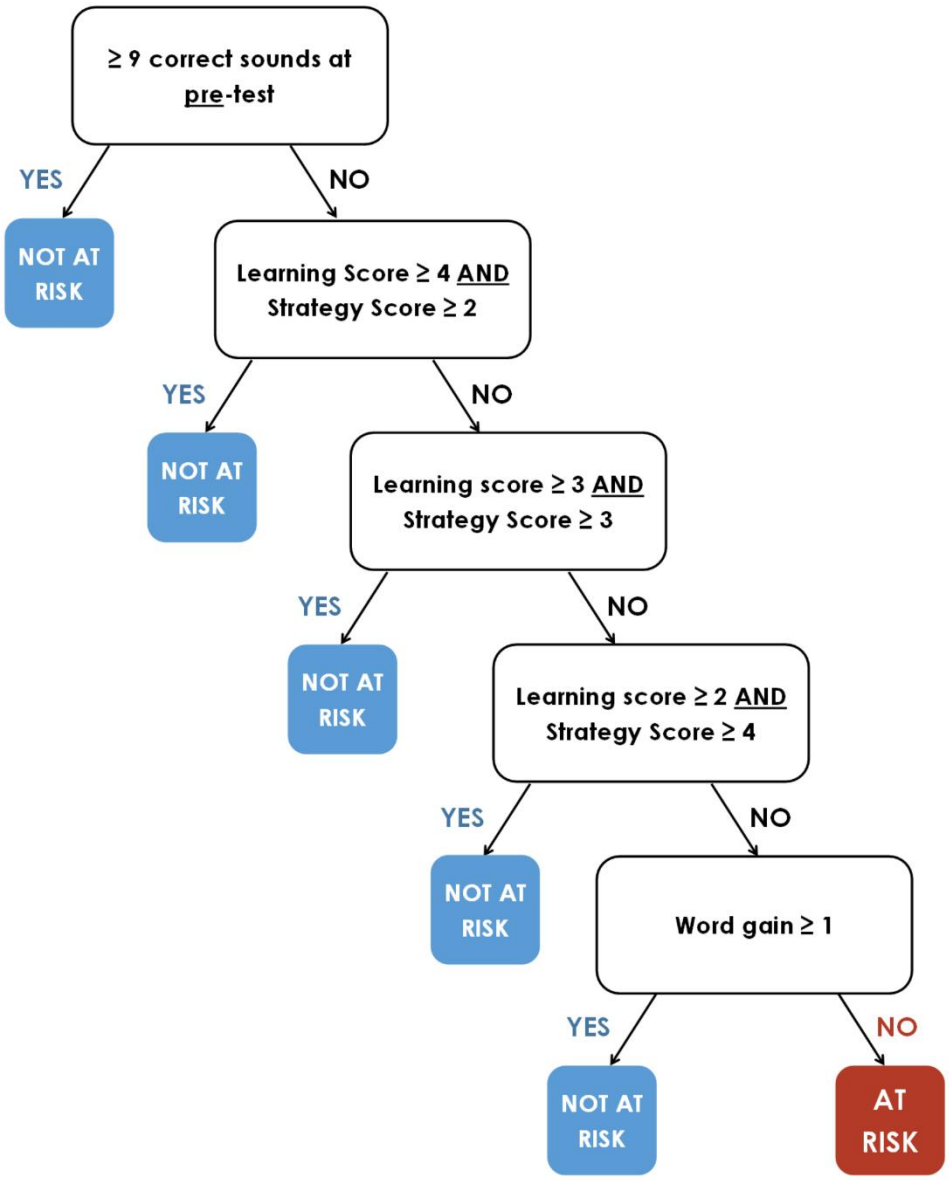
[2] Task confidence behaviours can be verbal or non-verbal. Examples of high-confidence behaviours include smiling and saying, "This is easy", "I can do this", and so on. Examples of low-confidence behaviours include helpless gestures and verbalisations (e.g., "I can't do it!"; "It's too hard"; "I don't know how").

[3] If the assessor models /m/ and the child says /n/, this is an error. If the assessor models "m-o-t" and the child says "m", "mot", "mo-t" or "not", it is an error. If the student self corrects, do not count as an error. However, if the child repeatedly self corrects, a score of 3 (some errors) would be appropriate.

Modifiability score: The learning score and the strategy score were combined to create a composite modifiability score between 2 and 10. This score is hypothesised to provide a representation of the child's overall modifiability (response to input) by combining both measures of the child's ability to apply a taught decoding strategy and/or rapidly learn new orthographic representations (strategy) and the assessor's judgement of how well the child responded to input (learning) (Petersen et al., 2016). Refer to Appendix A (Section A.2) for details of exploratory work completed to determine the method for calculating the modifiability score.

Dynamic assessment dichotomous score (DADS): A dynamic assessment dichotomous score was calculated to supply a risk status classification for each child (*at risk* or *not at risk*). The purpose of this was to allow for an analysis of the sensitivity and specificity of the dynamic assessment. The procedure for calculating the DADS is summarised in Figure 3.3. For example, if a child produces two correct sounds during the pre-test (i.e., < 9), the assessor then looks at the child's learning score and strategy score. If the child has a learning score of 2, the assessor checks whether the child has a strategy score of 4 or more. If the child has a strategy score of 4 or more, the child's dynamic assessment dichotomous score would be *not at risk* (entered as 0 for data analyses). If the strategy score is less than 4, the assessor then checks whether the child had a word gain score of 1 or more. If so, the dichotomous score would be *not at risk*, otherwise, it would be *at risk* (entered as 1 for data analyses). Refer to Appendix A for details of exploratory work completed to determine the method for calculating the DADS.

Figure 3.3: Dynamic Assessment Dichotomous score flow chart



Outcome measures after a year at school (T2)

One year after the children commenced formal schooling, outcome measures were administered to each child individually, in a single test session. These outcome measures are summarised in Table 3.2 and are listed in the order they were administered.

Table 3.2: Summary of after-a-year-at-school (T2) outcome measures

Static measures	
YARC – EWR[a]	The York Assessment of Reading for Comprehension Australian Edition - Early Word Recognition
DIBELS Next[b] • LNF • PSF • NWF	Dynamic Indicators of Basic Early Literacy Skills (DIBELS) • Letter Naming Fluency • Phonemic Segmentation Fluency • Nonsense Word Fluency
School Measures	
OTJ	Overall teacher judgement
Book level	Book level – *Ready to Read* series

[a] Hulme et al., 2012
[b] Good & Kaminski, DIBELS Next assessment manual, 2011

YARC Early Word Recognition

The *York Assessment of Reading for Comprehension (YARC) Australian Edition - Early Word Recognition (EWR)* subtest was used to assess early word reading ability. The test is designed for use with children aged 5 to 7, and the Australian Edition provides norms for students in the age range 5 years 0 months to 7 years 11 months.

The YARC – EWR consists of 30 words graded in difficulty, half of which are phonemically regular and half of which are exception words (phonemically irregular). Each child was asked to read the words and was given as much time as they needed to do so. The child was awarded one point for

each word they read correctly, and a word was counted as read correctly regardless of whether it was first sounded out and then read, or read orthographically (without sounding out). The raw score (out of 30) was converted to an ability score, which in turn was converted to a standard score, using the tables provided in the YARC Test Manual.

Descriptive terms that correspond to the standard scores are also provided, ranging from *severe difficulty* to *excellent* (Hulme et al., 2012). For the purposes of this study, only those children described as having a *severe difficulty* (standard score < 80) were classified as having a reading difficulty (RD); those with a standard score of 80 or more were classified as *no reading difficulty* (NRD).

A high degree of reliability for the YARC EWR was reported (Cronbach's alpha .98). In terms of validity, intercorrelations between the YARC EWR and other YARC Early Reading tests ranged from .57 to .93 (Hulme et al., 2012).

DIBELS Next

DIBELS Next Letter Naming Fluency (LNF) was used to assess alphabet knowledge (letter names). Children are shown a page with upper- and lowercase letters in random order. They need to name as many letters as they can in one minute.

DIBELS Next Phonemic Segmentation Fluency (PSF) was used to assess phonemic awareness skills, in particular, the child's ability to segment three- and four-phoneme words into their individual phonemes. The assessor presents a three- or four-phoneme word and the child is asked to produce the individual phonemes in the word. For example, the assessor says *bed* and the child says "/b/ /e/ /d/" to get three points for the word. If the child says /be/… /d/ or /b/… /ed/, they are awarded two points. No points are awarded if the child repeats the entire word. Once the child has given their answer, the assessor presents the next word. The final score is the number of correct phonemes produced within one minute.

DIBELS Next Nonsense Word Fluency (NWF) was used to assess knowledge of the alphabetic principle and the ability to blend letters into words (i.e., phonologically decode the word). The child is given a page with 50 randomly ordered nonsense words, including vowel-consonant words such as *ec* and consonant-vowel-consonant words such as *jif*. The child is asked to read as many of these words as they can, and if they cannot read a word, to say any sounds they know in the word. The child is given one minute to say as many words or sounds as possible. This test yields two scores: one for correct letter-sound correspondences (*NWF-CLS*) and one for whole words read (*NWR-WWR)*.

The *NWF-CLS* is the number of correct letter-sound correspondences generated during the reading of the nonsense words. For example, for the word *fec*, if the child reads the word as /f/ /e/ /k/, the score is 3. If the child reads *fec* as /fe/ /k/, /f/ /ek/, or *fec*, the score is also 3.

The *NWR-WWR* is the number of nonsense words read correctly as a whole word without first being sounded out. For example, if the child reads *fec* immediately (without sounding out) the score is 1 for *WWR* (as well as 3 points for *CLS)*. However, if the child first sounds out the word (e.g., /f/ /e/ /k/ … *fec*) no points are awarded for *WWR* (but 3 points are awarded for *CLS*).

Alternate form, test-retest, and inter-rater reliabilities ranging from .70 to .99 were reported, indicating strong evidence of reliability for the PSF, NWF-CLS, and NWF-WWR sub-tests. However, the concurrent criterion validity with the GRADE measure ranged from .24 to .40 indicating a small to moderate correlation (Good et al., 2013).

DIBELS Next Composite score: The following measures combined to calculate the Composite score: LNF, PSF, NWF-CLS. Excellent reliabilities were reported for the Composite score, with alternate form, test-retest, and inter-rater reliability coefficients ranging from .94 to .99. A moderate concurrent validity coefficient of .37 was reported between the Composite score and the GRADE criterion measure (Good et al., 2013).

The *DIBELS Next* measures used after one year at school provide *core*, *strategic,* and *intensive* classification categories for PSF, NWF-CLS, and the Composite score. A benchmark goal and cut point for risk is also given for NWF-WWR. However, as the NWF-WWR cut point for risk at this stage is 0, only *core* and *strategic* classification categories are available. For the outcome measures, the start of Grade 1 benchmark goals and cut points for risk were applied, and an *intensive* classification was used as an indication of reading difficulty (RD). As an intensive classification is not available for NWF-WWR at this stage, this score was not dichotomised.

School judgement

Teachers with participant children in their classes were asked to complete a form on which they indicated the following.

Overall Teacher Judgement (OTJ): Based on her/his overall judgement, the teacher was asked to indicate whether s/he thinks think the child has a reading difficulty (e.g., is likely to need additional support). This was a dichotomous response (Yes = Reading difficulty (RD); No = No reading difficulty (NRD).

Book Level: Ready to Read is the core instructional book series that supports reading in The New Zealand Curriculum. Children start reading at Magenta (level 1 and 2), which is classified as 'emergent reader'. During their first year of school, children are expected to progress through Red (levels 3 to 5), Yellow (levels 6 to 8), and Blue (levels 9 to 11), with the expected national standard at the time being that students would be reading at the Green level (levels 12 to 14) after one year at school. Teachers were asked to indicate the current book level for each child.

The Ministry of Education indicates that children still reading at Red or Magenta book level in the *Ready to Read* book series after one year at school require additional support (New Zealand Ministry of Education, n.d. -b). However, in practice, most teachers reported higher book levels (Yellow, and in some cases, Blue) for children they had indicated had a reading difficulty (see Table 3.3).

Table 3.3: Comparison of Overall Teacher Judgement and Book Levels

	Overall Teacher Judgement	
Book Level (Colour)	NRD (%)	RD (%)
Magenta	0	100
Red	0	100
Yellow	37	63
Blue	56	44
Green	100	0
Orange	100	0
Turquoise	100	0
Purple	100	0
Gold	100	0

Based on this, when converting the book level for each child to a dichotomous score (reading difficulty/no reading difficulty), Yellow, Red, and Magenta were used to indicate that additional support is needed (i.e., *reading difficulty* RD).

To offset discrepancies between the Overall Teacher Judgements and Book Levels provided, a dichotomous School score was calculated as follows: If the Overall Teacher Judgement indicated reading difficulty and the book level was Magenta, Red, or Yellow, the School score was *reading difficulty* (scored as 1 for analyses). Otherwise, the School score was 0 (i.e., *no reading difficulty*).

Dichotomous outcome score (T2 dichotomous score)

To allow for an investigation of the classification accuracy of the predictor measures, a dichotomous outcome score of 0 (no reading difficulty) or 1 (reading difficulty) was calculated, following a similar procedure to that used by Petersen and colleagues (2016). Using the dichotomous scores for DIBELS Next PSF, DIBELS Next NWF-CLS, YARC EWR, and the School score, children were classified as having a reading difficulty after a year at school if three or more of these measures had a score of 1 (i.e., reading difficulty).

Statistical analyses

Guided by the review of theory and pertinent prior research, the statistical techniques used to address the research aim for this study included descriptive and inferential analyses. Descriptive analysis included measures of central tendency and variation in scores to describe current reading abilities as assessed by the dynamic and static predictor measures and outcome measures. Inferential results included correlational analyses, logistic regression, receiver operating characteristic (ROC) analyses, and analysis of the classification accuracy and precision of the predictor measures. These analyses were used to explore relationships among the dynamic and static predictors measures and the outcome measures. Details of the statistical analyses are provided in the chapter that follows.

Summary

The aim of the current study was to determine whether measures of emergent and early literacy skills administered at the start of schooling could be used to accurately screen for reading difficulty after one year at school. This chapter outlined the specific areas of investigation and the research design for the study. This included details of the ethical considerations for the study, the research participants and how they were recruited, as well as the assessment measures used and the procedures for the administration, scoring, and analysis.

Chapter 4:
Results

The aim of this study was to investigate the use of a dynamic assessment of decoding, administered to children upon school entry in New Zealand, to predict reading difficulty after a year at school. This chapter presents the results of the data analyses used to investigate whether the results of the static and dynamic tests administered to children at the start of their formal schooling (i.e., predictor measures) were predictive of the early reading tests administered after a year at school (i.e., outcome measures), and which of the predictor measures had superior predictive ability and most precise classification accuracy.

The chapter begins by outlining the statistics software used for data preparation and analyses, the significance level used for all analyses, and the abbreviations used in this chapter. This is followed by a comparison of the children who were tested as both T1 and T2 (longitudinal group), and those who were tested at T1 but were not available for T2 testing (attrition group). After this, the univariate descriptive results are presented for the longitudinal group, to summarise the children's test scores for the static and dynamic predictor measures administered at the start of formal schooling, and those achieved on the outcome measures administered a year later. This is followed by a presentation of the results relevant to the central aim of this study, namely to investigate the use of a dynamic assessment of decoding administered to children upon school entry to predict reading difficulty after a year at school.

The results will be presented based on the following analyses.

1. Correlational analysis to examine the relationship between static and dynamic reading measures administered at the beginning of formal schooling (T1) and reading ability at the end of a year at school (T2).

2. Logistic regression to examine the ability of static and dynamic reading measures administered at the beginning of formal schooling (T1) to predict future reading difficulty. Logistic regression analyses were conducted to determine whether the static and dynamic reading measures administered at the start of formal schooling were able to predict reading difficulty status after a year at school, and to compare the predictive power of these measures.

3. Receiver Operating Characteristic (ROC) analyses and comparisons of sensitivity and specificity were undertaken to examine if there was a difference in predictive classification accuracy of the static measures and dynamic measures in terms of future reading difficulty.

4. Logistic regression and ROC analyses were used to examine if there was an increase in the predictive ability and classification accuracy when the static and dynamic measures were combined versus used on their own.

5. Logistic regression, ROC analysis, and comparisons of sensitivity and specificity were used to examine the dichotomised (*at risk*, *not at risk*) scores for the static and dynamic predictor measures. These dichotomised scores provide information that can readily be understood and easily applied in a school setting.

Statistics software

IBM SPSS Statistics Version 24 (Windows) was used for data file preparation and most of the data analyses. To reduce the possibility of calculation errors, SPSS syntax was used wherever feasible to calculate variables. SPSS does not include a facility to test the difference in the area under the curve (AUC) for two Receiver Operating Characteristic (ROC) curves in order to compare their predictive classification accuracy. To conduct the pairwise comparisons of the ROC curves for different predictor measures, MedCalc 18.5 (Windows) was employed.

Significance level and abbreviations

A significance level of $p < .05$ was used for all statistical analyses. This p level is most commonly used for studies of this nature with similar sample sizes. Box 4.1 provides a key for the abbreviations used for variables in all tables presenting the results of the analyses.

Box 4.1: Key for abbreviations used for predictor and outcome measures

T1 predictor measures	
Static measures	
LNF T1	DIBELS Next Letter Naming Fluency at T1 (kindergarten)
FSF	DIBELS Next First Sound Fluency
COMP T1	DIBELS Next Composite at (kindergarten) – score combining LNF T1 and FSF
RON	CTOPP-2 Rapid Object Naming
Dynamic measures	
PreS	DA Pre-Test Sounds
PreW	DA Pre-Test Words
S Gain (taught)	DA Sound Gain – Taught
S Gain (all)	DA Sound Gain – All
SRG (taught)	DA Sound Residuum Gain – Taught
SRG (all)	DA Sound Residuum Gain – All
W Gain (taught)	DA Word Gain - Taught
W Gain (all)	DA Word Gain - All
DA Learning	DA Learning
DA Strategy	DA Strategy
DA Modifiability	DA Modifiability - composite of DA Learning and DA Strategy
DADS	DA dichotomous score - based on PreS, DA Learning, DA Strategy, and W Gain (all)

T2 outcome measures	
LNF T2	DIBELS Next Letter Naming Fluency at T2 (Grade 1)
PSF	DIBELS Next Phonemic Segmentation Fluency
NWF-CLS	DIBELS Next Nonsense Word Fluency - Correct Letter Sounds
NWF-WWR	DIBELS Next Nonsense Word Fluency – Whole Words Read
COMP T2	DIBEL Next Composite at T2 (Grade 1)
YARC EWR	YARC Early Word Reading
OTJ	Overall Teacher Judgement
Bk Lvl	Book level (colour) of *Ready to Read* Series
T2 dichot	Dichotomised outcome score (1 = reading difficulty; 2 = no reading difficulty)

Note: CTOPP-2 = *Comprehensive Test of Phonological Process* (2nd ed.); DIBELS = *Dynamic Indicators of Basic Early Literacy Skills Next*; DA = Dynamic assessment; YARC = *York Assessment of Reading Comprehension*

Comparison of longitudinal and attrition groups

Table 4.1 summarises the participant characteristics for the children who completed testing at both T1 and T2 (longitudinal group; *n* = 136), and the children who were not available for testing at T2 (attrition group; *n* = 29). The children in the attrition group were slightly older at T1 than those in the longitudinal group and achieved lower scores on all the T1 predictor measures. However, t-tests comparing the attrition and longitudinal groups indicated the differences between the groups were non-significant in terms of age and scores on the T1 static measures. In contrast, on average, the scores for the dynamic assessment measure (PreS) for the attrition group (M = 5.48, SE = 1.393) were significantly lower (M difference = -3.26, 95% CI [-6.185, -0.320], t = -2.19 (163), p = .03) than that for the longitudinal group (M = 8.74, SE = .618). This represented a small effect, *d* = 0.32.

Table 4.1: Participant characteristics – longitudinal and attrition groups

	Longitudinal group				Attrition group			
	n	*M*	*Mdn*	*SD*	*n*	*M*	*Mdn*	*SD*
Age (mos) at T1	136	61.50	61.00	1.58	29	62.28	62.00	2.37
Age (mos) at T2	136	73.29	73	1.72	-	-	-	-
School Decile*	136	87.60	7.00		29	61.45	7.00	
Female	61 (44.9%)				15 (51.7%)			
Male	75 (55.1%)				14 (48.3%)			
LNF T1	136	21.82		18.21	29	15.28		17.87
FSF	136	17.56		15.06	29	12.83		15.20
COMP T1	136	39.38		29.20	29	28.10		29.26
RON	136	10.44		3.16	29	8.17		4.27
PreS	136	8.74		7.21	29	5.48		7.50

* For School Decile, mean rank is reported.

The difference in the numbers of male and female participants in the attrition and longitudinal groups was non-significant (p = .32, Fisher's exact test). Likewise, chi-squared tests showed that differences in terms of composition of ethnicity (χ^2 = 26.50, p = .50, Cramer's V = .40) and language (χ^2 = 7.70, p = .10, Cramer's V = .22) of the two groups were not significant. In terms of school decile, the decile rank for children in the attrition group was significantly lower (mean rank = 61.45) than that of the longitudinal group (mean rank = 87.60, U = 2,597.00, z = 2.76, p = .006). This is in keeping with the higher rate of transient students in lower decile schools that has been reported in other studies (New Zealand Ministry of Education, 2018). However, the effect size for this difference was small (r = .22).

In summary, in the case of age, gender, and scores for the T1 static predictor measures, the differences in the longitudinal and attrition groups were not significant. Although the decile rank was significantly lower for the attrition group, the effect size was small. The DA Modifiability score was also significantly lower for the attrition group, with a medium to large effect. The remainder of this chapter presents the results for the longitudinal group only (i.e., children tested at both T1 and T2).

Descriptive results of predictor and outcome measures

In this section, descriptive results produced by the predictor and outcome measures are presented including distribution trends, and an assessment of outliers (see Table 4.2). All scores shown are raw scores, except for the RON (scale score) and YARC EWR (standard score). For the dynamic assessment, 69 children who could read nine or more sounds at pre-test did not go on to the teaching and post-test phases of the dynamic assessment. Therefore, dynamic assessment post-test scores are only presented for the 67 children who went on to the teaching and post-test stages.

Table 4.2: Descriptive statistics – school entry (T1) static and dynamic predictor measures, and after-a-year-at-school (T2) outcome measures*

	N	*M*	Mode	*SD*	Skew	Kurtosis
Static predictor measures (T1)						
RON (scale score)	136	10.44	12	3.16	-1.22	1.71
LNF T1	136	21.82	3	18.21	0.93	0.47
FSF	136	17.56	0	15.06	0.43	-0.80
COMP T1	136	39.38	3	29.20	0.62	-0.24
Dynamic predictor measures (T1)						
PreS	136	8.74	0	7.21	0.02	-1.72
PreW	136	0.76	0	1.80	2.19	3.23
S Gain (taught)	67	7.61	9	5.46	-0.04	-1.19
S Gain (all)	67	10.88	2	7.08	-0.03	-1.25
SRG (taught) (%)	67	48.17	100	35.16	-0.05	-1.35
SRG (all) (%)	67	50.04	8	32.91	-0.03	-1.34
W Gain (taught)	67	0.30	0	0.89	4.96	30.58
W Gain (all)	67	0.33	0	0.85	4.40	24.58
DA Learning	67	3.25	3	0.82	-0.51	0.40
DA Strategy	67	2.54	3	1.06	0.09	-0.92
DA Modifiability	67	5.79	6	1.68	-0.34	-0.24
Outcome measures (T2)						
LNF T2	136	42.29	47	15.44	0.13	1.09
PSF	136	36.07	49	17.09	-0.47	-0.39
NWF-CLS	136	37.82	33	32.47	1.45	1.78
NWF-WWR	136	8.71	0	11.57	1.77	2.63
COMP T2	136	116.18	31	51.92	0.57	0.97
YARC EWR (standard score)	136	106.43	128	15.37	-0.61	-0.16
OTJ	No reading difficulty: n = 103 (75%)			Reading difficulty: n = 33 (24%)		
Bk lvl	Median = Green			Mode = Green		

* All scores are raw scores, unless otherwise indicated

Static predictor measures (T1)

The static measures administered at T1 all exhibited floor effects, except for the rapid-naming task (CTOPP-2 RON). The DIBELS Next LNF T1 (letter naming) had a mean score of 21.82 (*SD* = 18.21), with a mode of 3 and the DIBELS Next FSF (first sound fluency) subtest had a mean score of 17.56 (*SD* = 15.06) with a mode of 0. According to the DIBELS Next benchmark goals and cut points for risk, an FSF score under 5 is considered *well below benchmark*, and an indicator that the child is at risk of not meeting further reading goals without intensive support. The DIBELS Composite score (which combines the LNF and FSF subtest scores) had a mean of 39.38 (*SD* = 29.20) and a mode of 3. A child who scores below 13 for the Composite score is considered at risk, and is likely to need extra, targeted instructional support. The RON Scale score did not exhibit the same floor effects as the DIBELS measures (mean = 10.44, possible range 1 to 18, *SD* = 3.16; mode = 12).

Response types

Despite a practice phase being included for the first sound fluency (FSF) task, it was clear that some children did not understand what they were required to do. This was even true for some children who appeared to have already developed some early literacy skills. For example, when asked during the practice phase, “What is the first sound you hear in the word ‘man’?”, the child would give the letter name rather than sound and would persist in providing the letter name for all three practice words, and then into the main test itself. In contrast, most children quickly understood what was required for the letter naming (DIBELS LNF) and rapid object naming (CTOPP-2 RON) tasks. In the case of the rapid naming task, even those few children who did not speak English at home and did not know the names of some of the objects at the start of the task, were able to quickly learn these during the practice phase and to use the object names in the task.

Dynamic predictor measures (T1)

The dynamic assessment pre-test scores showed floor effects with a mean of 8.74 (*SD* = 7.21) and a mode of 0 for the sound score and a mean of 0.76 (*SD* = 1.80) and a mode of 0 for the word score. The dynamic assessment pre-test on its own operates as a static test as there is no teaching input from the assessor during this stage: Children are simply shown the pseudowords and asked to read them or tell the assessor any sounds they know in the words. As such, it was not surprising that this test also exhibited floor effects.

In contrast to the dynamic assessment pre-test scores and the DIBELS static measures, the dynamic assessment post-test scores did not show floor effects, the exception being the post-test word scores. The DA Sound Gain (All) score had a mean of 10.88 (actual range -2 to 24), and the mean for the DA Sound Residuum Gain (All) score was 50.04% (actual range -9% to 100%). Scores for the DA Strategy (*M* = 2.54, *SD* = 1.06), DA Learning (*M* = 3.25, *SD* = 0.82), and DA Modifiability, which combines the DA Strategy and DA Learning scores, (*M* = 5.79, *SD* = 1.68) also did not show floor effects.

Response types

As was the case with the DIBELS first sound fluency task, during the dynamic assessment pre-test, several children provided the letter names in words, not understanding that they should be giving the sounds in the word. In most cases, these were the same children who responded by providing letter names rather than sounds in the DIBELS first sound fluency task.

During the teaching phase of the dynamic assessment, it was clear that most children quickly grasped what was required. However, there were several cases where, for a variety of reasons, learners could not or did not apply the instructions given to them. For example, during the teaching phase, when the assessor asked the child to imitate what she did as she sounded out each phoneme, it was evident that some children were not able to distinguish certain sounds, in

particular the plosives (/p/, /t/, and /k/), and nasal consonants (/n/ and /m/). In other cases, children struggled to consistently attend to the task, were distracted (or created distractions), and/or were difficult to refocus on the task. While most children persevered with the task even when they found it challenging, some tried to distract the assessor or tried in other ways to avoid completing the task. In a few cases, children persisted in providing letter names, rather than letter sounds in the pseudoword, both during and after the teaching phase. Information about the aforementioned behaviours (errors, attention, perseverance, etc) was gathered for each child using the Learning Scale, and this helped guide the assessor in assigning an overall DA Learning score for each child.

Outcome measures (T2)

Apart from the DIBELS Next NWR-WWR (whole word reading of nonsense words), the T2 outcome measures did not show floor effects. For the DIBELS Next NWR-CLS (correct letters read in nonsense words), the mean score was 37.82 (mode = 33), which is higher than the benchmark goal of 27 set for this stage. For the LNF T2 (letter naming), the mean score was 42.29 (mode = 47), and for the PSF (phonemic segmentation fluency), the mean score was 36.07 (mode = 49). For the YARC EWR (Early Word Reading), there was evidence of a ceiling effect with a mean score of 106.43 (mode = 128). As the YARC-EWR task is not timed, the children had the opportunity to spend more time trying to sound out words than is the case in the DIBELS Nonsense Word Reading (NWR) task, which is timed and only allows children one minute to read (or provide sounds for) a list of nonsense words. For the NWF-WWR (whole nonsense word read) there were clear floor effects (*M* = 8.71; mode = 0), with half the children scoring fewer than 5 correct words (out of a possible of 50). These floor effects are expected for this measure, with the cut point for risk set at 0.

With regards to the data provided by teachers about each child's reading ability after a year at school, the mode for the Book Level was Green. This is not surprising given that at the time of

testing, the expected national standard was that children would be reading at Green (levels 12 to 14) after a year at school. According to the overall teacher judgement (OTJ), 24% (33 children) were identified by teachers as having a reading difficulty.

The distribution of all the static measures, at both T1 and T2, exhibited significant skew (i.e., non-normal distribution of scores). For the DIBELS measures (LNF T1, FSF, COMP T1, NWF-CLS, NWF-WWR and COMP T2), the skew was significantly positive. In contrast, for the CTOPP-2 RON, DIBELS PSF, and YARC EWR, the skew was significantly negative. For the T1 dynamic measures, the Pre-Word and Word Gain scores were all significantly positively skewed, reflecting the difficulty most of children had reading whole pseudowords (both at pre-test and post-test), while there were a few children who were able to read most or all the pseudowords.

The distribution of nonsense word reading scores at T1 (DA Pre-Word, Word Gain (taught), Word Gain (all)) and at T2 (NWF-WWR) were all leptokurtic (i.e., peaked distribution with thicker than normal tails). Inspection of the data showed that this kurtosis was indicative of the abilities of a small number of children who were able to read the nonsense words, while the majority were unable to read any nonsense words. Inspection of the histogram for the DA Pre-Sound measure indicated a clear bimodal distribution (i.e., two peaks in the histogram), which reflected the fact that while many children (21%) were unable to name any letter sounds, there were also several children (17%) who were able to correctly name all the sounds in the pseudowords.

A range of tools were used to identify outliers, including the use of visual checks (using histograms and boxplots), as well as using z-scores. All extreme outliers were investigated by checking each outlier case manually; in all cases the outliers were found to be genuine outliers and not, for example, a result of a data entry error. For the T1 measures, the outliers for the DIBELS LNF T1 and COMP, and the DA Pre- and Post-Word tasks, were children who performed significantly better than their peers, and in most cases, the same children were outliers in all the

aforementioned tasks. For the CTOPP-2 RON and DA Learning, children who were outliers performed significantly more poorly than the rest of the cohort, and in most cases the same children were outliers for both measures. In the case of the DIBELS nonsense word task at T2 (NWF-CLS and NWF-WWR scores), the same children were outliers and could read many more sounds and words than the other children. On the other hand, for the early word reading task at T2 (YARC EWR), outliers were children who read far fewer words than their peers. With the DIBELS letter naming task at T2 (LNF T2) and the DIBELS composite score at T2, the pattern of outliers was more complex, with outliers who performed both significantly better and significantly worse than their cohort. In most cases, the same children who did much better than their peers on the NWF-CLS and NWF-WWR, also performed much better on the LNF T2 task and achieved higher DIBELS composite scores; and the same children who performed much worse than their peers on the early word reading task (YARC EWR), were also outliers (with far lower scores), for the DIBELS LNF T2 and COMP T2. Despite these outliers, analysis of the z-scores found that they did not have an impact on the overall normal distribution of the measures. This was supported by comparing the mean and 5% trimmed mean values, which showed negligible differences in the mean if the outliers were removed. As a result of this, and the fact that all the outliers were genuine outliers, it was decided not to remove or transform any outliers.

Response types

Unlike at T1, most children quickly understood what was required in the LNF, PSF, and YARC-EWR tasks. For the YARC-EWR task, most children were able to read most or all of the words as sight words and others could, when prompted, first sound out and then read words they could not read as sight words. On the other hand, for the DIBELS Nonsense Word Fluency task, it was evident that some children had difficulty understanding what was required. The short instruction and practice phase appeared to be insufficient for these children. Other children who appeared to grasp what was required, nevertheless found the task challenging and took a long time to

provide the individual letter sounds in words. The task was timed and therefore even though they generally named the sounds correctly, because they took a long time to do so, they did not manage to name many sounds. As was the case with the dynamic assessment test at T1, there were several children who persisted in providing letter names rather than sounds in the nonsense words. In some cases, children who were clearly competent readers at this age (e.g., they read all words as sight words in the YARC) tried to read the nonsense words as if they were real words. Common examples included reading *fij* as *fiji*, *wav* as *wave*, *pek* as *peek*, *fiv* as *five*, and so on. As a result, these children would be awarded points for correct sounds (e.g., 2 out of 3 correct sounds for *wave* instead of *wav*), but no point for a correct word (WWR).

Summary

The descriptive analyses of the children's reading abilities at the start of formal schooling as assessed by the predictor measures of emergent literacy skills, indicated the clear presence of floor effects and positive skew for most of the static measures. This includes those for the dynamic assessment pre-test, which operates as a static test of the child's ability to phonologically decode nonsense words at the start of the dynamic assessment. These floor effects and positive skew are indicative of the fact that most children had difficulty correctly completing the test activities. In contrast to the static measures, floor effects and positive skew were not present in the dynamic assessment post-test scores (i.e., DA post-test sound scores, DA Strategy, DA Learning, and DA Modifiability scores). Most of the outcome measures of early reading also did not exhibit floor effects, however, all exhibited significant skew (i.e., non-normal distribution of scores). All outliers were investigated and found to be genuine outliers. Furthermore, as these outliers did not have a significant impact on the distribution of the data, the outliers were retained, untransformed.

Relationships among predictor and outcome measures

After the completion of descriptive analyses, correlation analyses were conducted to examine whether static and dynamic reading measures administered at the beginning of formal schooling correlated with reading ability at the end of a year at school. Pearson's product moment correlations were used for correlations between all continuous measures; Spearman's rank-order was used for correlations between the Book Level (ordinal) and the continuous measures; and a biserial correlation was conducted to determine intercorrelations between the overall teacher judgement (score of 1 = *reading difficulty*, 0 = *no reading difficulty*) and the continuous and ordinal variables. Give the non-normal distribution of many of the measures, the correlational analyses were conducted using Bias Corrected and accelerated (BCa) bootstrapping (1000 samples). This automates random sampling with replacement to produce a more reliable estimate of the statistic (Field, 2015).

As shown in Table 4.3, almost all the dynamic predictor measures were significantly correlated with the outcome measures at the $p < .05$ level. These correlations were positive, except in the case of correlations between the predictor measures and the OTJ outcome measure, where the correlations were negative (OTJ of 0 = *no reading difficulty*; 1 = *reading difficulty*). In general, the DA Modifiability score correlated most strongly with all the T2 outcome measures, with correlations ranging from .34 (with PSF) to .81 (with the Overall Teacher Judgement). The strongest correlations overall were between the Overall Teacher Judgement (OTJ) at T2 and the DA Modifiability (.81), and DA Learning (.76) scores. Except for the letter naming task (LNF T2), the correlations between the DA Pre-Sound score and the T2 outcome measures were non-significant.

Table 4.3: Intercorrelations between predictor (T1) and outcome (T2) measures

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. LNF T1																
2. FSF	.21															
3. COMP T1	.79**	.77**														
4. RON	.24	.15	.25*													
5. PreS	.29*	.22	.33**	.02												
6. S Gain - All	.55**	.35**	.58**	.24*	.07											
7. W Gain – All	.28*	.50**	.50**	.04	.16	.29*										
8. DA Learning	.47**	.35**	.53**	.47**	.24	.57**	.22									
9. DA Strategy	.62**	.38**	.64**	.21	.32**	.90**	.42**	.57**								
10. DA Modifiability	.62**	.41**	.67**	.36**	.32**	.85**	.37**	.85**	.91**							
11. LNF T2	.51**	.23	.48**	.31*	.26*	.48**	.19	.52**	.53**	.59**						
12. PSF	.19	.30*	.31*	.19	.06	.27*	.15	.31*	.29*	.34**	.45**					
13. NWF-CLS	.43**	.39**	.53**	.17	.15	.46**	.34**	.33**	.48**	.47**	.68**	.32**				
14. COMP T2	.47**	.40**	.55**	.26*	.14	.50**	.30*	.46**	.53**	.57**	.85**	.69**	.87**			
15. YARC EWR	.55**	.28*	.53**	.23	.17	.49**	.25*	.55**	.51**	.59**	.77**	.44**	.75**	.81**		
16. OTJ[a]	-.70**	-.31**	-.65**	-.11	-.18	-.70**	-.23	-.76**	-.69**	-.81**	-.70**	-.58**	-.74**	-.84**	-.93**	
17. Bk Lvl	.70**	.28*	.62**	.22	.10	.54**	.33**	.54**	.57**	.62**	.61**	.50**	.64**	.69**	.69**	.92**

** $p < .01$ * $p < .05$ [a] All correlations are negative as 1 = reading difficulty; 0 = no reading difficulty

There were also significant positive correlations between the majority of the T1 static predictors and the T2 measures at the $p < .05$ level. The exceptions were for the correlations between the T1 static predictors and the OTJ, which were significantly negative. The CTOPP-2 RON did not correlate significantly with the majority of the T2 measures. In general, the correlations between the static predictor measures and T2 measures were weaker than those between the dynamic predictor measures (except for DA Pre-Sound) and T2 measures.

The T2 PSF scores did not correlate significantly with several T1 measures, and for those where the correlation was significant, this relationship was weak. On the other hand, the Overall Teacher Judgement (OTJ) at T2 correlated significantly with all T1 measures except for the CTOPP-2 RON and the DA Pre-Sounds and Word Gain scores, and in general these correlations were stronger than for any of the other T2 measures.

In summary, apart from the DA pre-test scores and the CTOPP-2 RON score, all the dynamic and static predictor measures showed significant correlations with the T2 outcome measures. In general, the dynamic predictor measures were more strongly correlated with the T2 outcome measures than the static predictor measures, with the DA Modifiability score producing the strongest correlation with the T2 outcome measures overall.

Prediction of reading difficulty

Logistic regression was employed to determine whether the static and dynamic reading measures administered at the start of formal schooling were able to significantly predict reading difficulty status after one year at school. Binomial (or binary) logistic regression was chosen because reading difficulty status (i.e., outcome measure at T2) was specified as a dichotomous variable: *reading difficulty* (RD) or *no reading difficulty* (NRD). Whereas linear regression is used to describe the linear dependence of a continuous outcome variable on one or more predictor variables, logistic regression provides information about the probability that a specific outcome will occur (e.g., reading difficulty/no reading difficulty) given certain conditions (e.g., scores on the T1 measures).

Logistic regression is used to determine whether a discrete outcome (e.g., reading difficulty/no reading difficulty) for individual cases can be predicted from a set of continuous, categorical (nominal or ordinal), and/or dichotomous predictors (commonly referred to as *covariates* in logistic regression). Therefore, unlike simple/multiple linear regression where the outcome variable must be continuous, binomial logistic regression is used when the outcome variable is dichotomous with a value of 1 assigned to the outcome of interest, and 0 assigned to the other outcome (Hosmer et al., 2013; Tabachnick & Fidel, 2012). As with linear regression, there are several assumptions and conditions that apply to binomial logistic regression. These assumptions and conditions, as well as how they were met for this study, are summarised in Appendix E.

For the logistic regression analyses, the T2 dichotomous score was used as the outcome variable and was calculated using the DIBELS Next PSF and NWF-CLS scores, the YARC EWR score, and the School score (combination of the OTJ and Book Level scores). If two or more of the aforementioned measures had a score of 1 (i.e., *reading difficulty*), then the overall T2 dichotomous score was coded 1 (*reading difficul*ty); otherwise, the score was coded 0 (*no reading difficulty*). Using the aforementioned criteria, 16 children (12%) were classified as having a reading difficulty.

A binomial logistic regression analysis was conducted using each of the predictor variables separately. The analyses for the continuous predictor variables included only those children who completed the full dynamic assessment (n = 67). This is because children who could correctly read three or more of the first four pseudowords in the dynamic assessment pre-test, or nine or more of the phonemes within these words, did not continue to the teaching and post-test stages of the dynamic assessment, and therefore did not have scores for DA Learning, DA Strategy, or DA Modifiability (n = 69). The results of these analyses are given in Table 4.4. The table shows the chi-square statistic (χ^2), Nagelkerke's R^2 (R^2_N), McFadden's (Likelihood Ratio) R^2 (R^2_L), and the Wald static (Wald) for each model. The χ^2 for each model compares the -2LL for the model containing the predictor variable, and that of a null model.

The -2LL statistic (also referred to as the log likelihood ratio) is used to compare the goodness of fit of two statistical models: in this case, a null model (with only the constant) and an alternative model (containing the predictor variable of interest).

Table 4.4: Logistic regression analyses: Reading difficulty status (T2 dichotomous) by individual continuous predictor variables[a]

	χ^2	R^2_N	R^2_L	Wald
LNF T1	21.08*	.42	.31	7.94**
FSF	4.74**	.11	.07	3.48
COMP T1	15.06 *	.31	.22	7.99**
RON	0.98	.02	.01	1.01
DA Learning	23.62*	.46	.34	12.48*
DA Strategy	20.63*	.41	.30	12.25*
DA Modifiability	27.03*	.52	.39	15.07*

[a] Only the 67 children who completed the full dynamic assessment are included in these analyses

* $p < .001$ ** $p < .05$ Note: all *df*s = 1

As can be seen in Table 4.4, the χ^2 statistic for models including the covariates, except for the CTOPP-2 RON, were all significant. This indicates that, apart from the CTOPP-2 RON, each of the individual static and dynamic measures significantly predicted reading difficulty (T2 dichotomous score) after a year at school.

The model χ^2 statistics for the significant static predictor measures ranged from 4.74 (FSF) to 21.08 (LNF T1). For the three dynamic predictor measures, the χ^2 statistics ranged from 20.63 (DA Strategy) to 27.03 (DA Modifiability). These model χ^2 findings are generally supported by the Wald statistic (z-statistic), which is analogous to the t-test in linear regression (Field, 2015), and which is significant for all models which included a covariate (all ps <.01), except for the DIBELS Next FSF and CTOPP-2 RON.

A comparison was made of the model fit of the individual predictors by comparing the differences in the model χ^2 values for the individual predictor variables (see Table 4.5). In the case of the static measures, the differences between the LNF T1 model and all three other models (FSF χ^2 = 4.74; COMP χ^2 = 15.06, RON χ^2 = 0.98) were statistically significant at $p \leq .01$, indicating the LNF model has a significantly better model fit than any of the other measures. For the dynamic measures, the difference in χ^2 values between DA Modifiability and DA Strategy (6.40) models were significant at p = .01. However, the differences between the DA Modifiability and DA Learning (3.41) scores were non-significant (p = .06). A pairwise comparison of the model fit of the static letter naming measure (LNF T1) and that of the DA Modifiability measure, indicated that the DA Modifiability score provided a significantly better model fit (χ^2 model difference = 5.95, p = .01).

Table 4.5: Comparison of model fit – individual predictors*

	χ^2	χ^2 diff	p
Comparison of:			
LNF T1	21.08		
FSF	16.34	4.74	.00
COMP T1	6.02	15.06	.01
RON	20.10	0.98	.00
Comparison of:			
DA Modifiability	27.03		
DA Learning	23.62	3.41	.06
DA Strategy	20.63	6.40	.01
Comparison of:			
DA Modifiability	27.03		
LNF T1	21.08	5.95	.01

* Only the 67 children who completed the full dynamic assessment are included in these analyses

An additional logistic regression analysis was conducted to determine whether the combination of the LNF T1 and FSF scores would improve the predictive model, compared to the use of LNF T1 on its own. The difference between the LNF T1 -2LL statistic (χ^2= 21.08) and that of the combined LNF and FSF model (χ^2= 21.53) was non-significant (χ^2 difference = 0.45, p = .50).

Likelihood ratio tests showed that three of the static measures (DIBELS Next LNF T1, FSF, and COMP) and each of the dynamic measures (DA Learning, DA Strategy, and DA Modifiability) all uniquely predicted the outcome measure. For this reason, the predictive power of each of these measures was investigated. Two statistics commonly used as indicators of the relative strength of the effect sizes of logistic regression models are the Nagelkerke and McFadden (likelihood ratio) pseudo-R^2s. Similar to the R^2 statistic in linear regression, pseudo-R^2s summarise the overall strength of the model, with values between 0 (no predictive value) and 1 (perfectly predictive) (Hu, Shao, & Palta, 2006; Pett, 2016). However, the pseudo-R^2 values are not directly comparable to R^2 of linear models and they cannot be interpreted as the percentage of variability in the predictor variable shared by the outcome variable (Field, 2015; Pett, 2016; Tabachnick & Fidel, 2012; Smith & McKenna, 2013). Instead, it has been suggested that pseudo-R^2 statistics be used to compare competing models for the same data, with the model that has the largest R^2 value considered superior (IBM, n.d.). Others suggest using these pseudo-R^2 statistics as estimates of effect sizes using different criteria than for ordinary least squares (linear) regression R^2 (Pett, 2016).

The pseudo-R^2 values for each of the static and dynamic predictor measures are given in Table 4.4 (R^2_N = Nagelkerke's R^2; R^2_L = McFadden's (Likelihood Ratio) R^2). Using the criteria suggested by Pett (2016), the effect for the DA Modifiability measure was moderate-strong (R^2_N = .52; R^2_L =.39), while the other DA measures, as well as the DIBELS Next LNF T1 demonstrated a moderate effect (range: .30 to .46). The pseudo-R^2 values indicated a low-moderate effect for the DIBELS Next Composite measure (range: .22 to .31), and a weak effect for the DIBELS Next FSF (range: .07 to .11).

In summary, all the T1 predictor measures, except for the rapid naming task (CTOPP-2 RON) were significant predictors of reading status at T2 (reading difficulty/no reading difficulty). For the *static* measures, the letter naming task (DIBELS Next LNF) produced a significantly better model fit than that using the DIBELS Next FSF, or DIBELS COMP. A combination of the LNF and FSF measures did not significantly improve the model fit over that of the LNF-only model. For the dynamic measures, a model using the DA Modifiability score produced a significantly better model fit than that using the DA Learning score, but the difference between the DA Modifiability and DA Strategy models was non-significant. A comparison of the DA Modifiability and LNF models indicated that the DA Modifiability score produced a significantly better model fit than that of the LNF model. Comparisons of the pseudo- R^2 values for the static and dynamic predictor measures indicated that the DA Modifiability measure had the strongest predictive power of all the measures, followed by that of the other dynamic assessment measures (DA Learning and Strategy), and the DIBELS Next LNF T1 and COMP static measures.

Predictive classification accuracy

This section presents the results comparing the predictive classification accuracy of the static and dynamic measures in terms of future reading difficulty. The section begins with a summary of key terminology related to predictive classification accuracy and commonly used descriptors of classification accuracy. Following this, the results comparing the predictive classification accuracy of the static and dynamic predictor measures are presented.

Predictive classification accuracy refers to how well a statistically significant model can classify cases with a known outcome. The predictive classification accuracy of a measure is contingent upon how well the measure correctly identifies group membership; in this research, this means how well a test administered at the start of schooling can correctly identify those children who,

after a year at school, are classified as having a reading difficulty, and those who are classified as not having a reading difficulty. The most commonly reported descriptors of predictive classification accuracy are: sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve. Each of these were used in the present study.

Sensitivity is the ability of a measure to correctly predict a condition (i.e., true positives); it is the proportion of true positives in a total group of individuals with the target condition (reading difficulty). In the current study, sensitivity can be understood as answering the following question: Out of the entire group of children classified as having a reading difficulty at T2, what percentage were correctly identified as at risk (true positive) using a specific predictor measure?

Specificity is the ability of a test to correctly predict the absence of a condition (i.e., true negatives). In the current study, specificity refers to the proportion of children correctly identified as *not at risk* by a specific predictor measure, out of the total number of children classified as not having a reading difficulty after a year at school.

A *receiver operating characteristic* (ROC) curve provides a visual representation of the true positive rate (sensitivity) and the false positive rate (specificity). The overall accuracy of a particular predictor is measured by the area under the ROC curve (AUC), with an area of 1 representing a perfectly predictive measure, and an area of .5 representing no-better-than-chance predictive accuracy.

Table 4.6 summarises the classification accuracy for each of the individual predictor measures in terms of the AUC (including the standard error and confidence intervals for the AUC), sensitivity, and specificity.

Table 4.6: Classification accuracy of individual predictor measures*

	AUC	*SE* (AUC)	95% CI	Sensitivity	Specificity
LNF T1	.87	.05	.76 to .94	86%	66%
FSF	.66	.08	.53 to .77	86%	45%
COMP T1	.83	.06	.72 to .91	86%	62%
DA Learning	.87	.04	.77 to .94	57%	96%
DA Strategy	.85	.05	.74 to .93	93%	62%
DA Modifiability	.91	.04	.81 to .97	93%	77%

* Only the 67 children who completed the full dynamic assessment are included in these analyses

ROC analyses were conducted to establish the classification accuracy of each of the individual static and dynamic predictor measures. The AUCs for the *static* measures ranged from .66 (FSF) to .87 (LNF). Using the example of the AUC for the FSF measure, the AUC can be interpreted as follows: Given two children (one classified as having a reading difficulty after a year at school, and one classified as not having a reading difficulty), being randomly selected from the study cohort, there would be a 66% probability that the child who would go on to be classified as having a reading difficulty would be correctly identified as *at risk* using the FSF test upon school entry (T1). For the dynamic measures, the DA Modifiability score had the largest AUC (.91), with the DA Learning and DA Strategy scores having AUCs of .87 and .85 respectively.

Pairwise comparisons of the ROC curves (see Table 4.7) revealed that in most cases the difference in the AUCs for the measures were non-significant. The exceptions being for FSF, where the AUC was significantly smaller than for all the other measures, and for the DA Strategy score, which had a significantly smaller AUC than the DA Modifiability score (which combines the DA Learning and DA Strategy scores).

Table 4.7: Pairwise comparisons of ROC curves*

	AUC difference	z	p
LNF T1 ~ FSF	.208	2.76	.006
LNF T1 ~ COMP T1	.039	0.94	.345
LNF T1 ~ DA Learning	.006	0.12	.901
LNF T1 ~ DA Strategy	.017	0.40	.689
LNF T1 ~ DA Modifiability	.041	1.13	.259
FSF ~ COMP T1	.170	3.72	.000
FSF ~ DA Learning	.214	2.92	.004
FSF ~ DA Strategy	.191	2.35	.019
FSF ~ DA Modifiability	.250	3.25	.001
COMP T1 ~ DA Learning	.045	0.86	.392
COMP T1 ~ DA Strategy	.022	0.40	.691
COMP T1 ~ DA Modifiability	.080	1.60	.110
DA Learning ~ DA Strategy	.023	0.55	.582
DA Learning ~ DA Modifiability	.036	1.31	.192
DA Strategy ~ DA Modifiability	.059	2.68	.007

* Only the 67 children who completed the full dynamic assessment are included in these analyses

Although the AUC provides useful information to compare the overall classification accuracy of a test, it does not provide information about the sensitivity and specificity of the test; two tests with very similar or even identical AUCs can have very different sensitivity and specificity values. Therefore, the sensitivity and specificity of the different measures were also compared to identify those with the best balance of sensitivity and specificity (refer to Table 4.6). In terms of the *static* measures, LNF had the best balance of sensitivity and specificity (86% and 66% respectively). For the *dynamic* measures, the balance of sensitivity and specificity was better for the DA Modifiability score (sensitivity = 93%; specificity = 77%) than the other dynamic assessment measures. Although the DA Learning score had higher specificity (96%) than the DA Modifiability score, this was at the expense of sensitivity (57%).

In summary, the DA Modifiability measure produced the best overall classification accuracy (Sensitivity = 93%; Specificity = 77%; AUC = .91) of all the measures investigated. For the static measures, DIBELS LNF T1 produced the best overall classification accuracy (Sensitivity = 86%, Specificity = 66%; AUC = .87). Table F.1 and Table F.2 in Appendix F provide full summaries of the fitted models using the T2 outcome variable and the dynamic and static predictor variables that provided the best model fit and balance of sensitivity and specificity, namely DA Modifiability and DIBELS Next LNF T1.

Comparison of individual and combined predictors

To determine whether the combination of static and dynamic measures produced a statistically significant better model fit than the static or dynamic measures on their own, the model likelihood ratio (deviance) and model chi-square (χ^2) for each model were compared. As mentioned earlier, the likelihood ratio (also referred to as -2LL or -2 log likelihood) is a measure of the probability that the observed values of the outcome variable can be predicted from the observed values of the predictor variables, and is used to test the goodness of fit of a given set of logistic regression models (Pett, 2016). As discussed earlier, of the *dynamic* measures, the DA Modifiability measure produced a better model fit than the DA Strategy and DA Learning measures. Therefore, it was decided to use only the DA Modifiability score in analyses that investigated whether the combination of dynamic and static measures significantly improved the model fit. The model χ^2 statistics for the model using only the DA Modifiability score, and ones where the DA Modifiability score was combined with static measures were compared to see if there was a significant improvement in model fit. Figure 3.4 summarises the models tested, and the results of these analyses are provided in Table 4.8.

Figure 3.4: Model comparisons (single predictor compared to combined predictors)

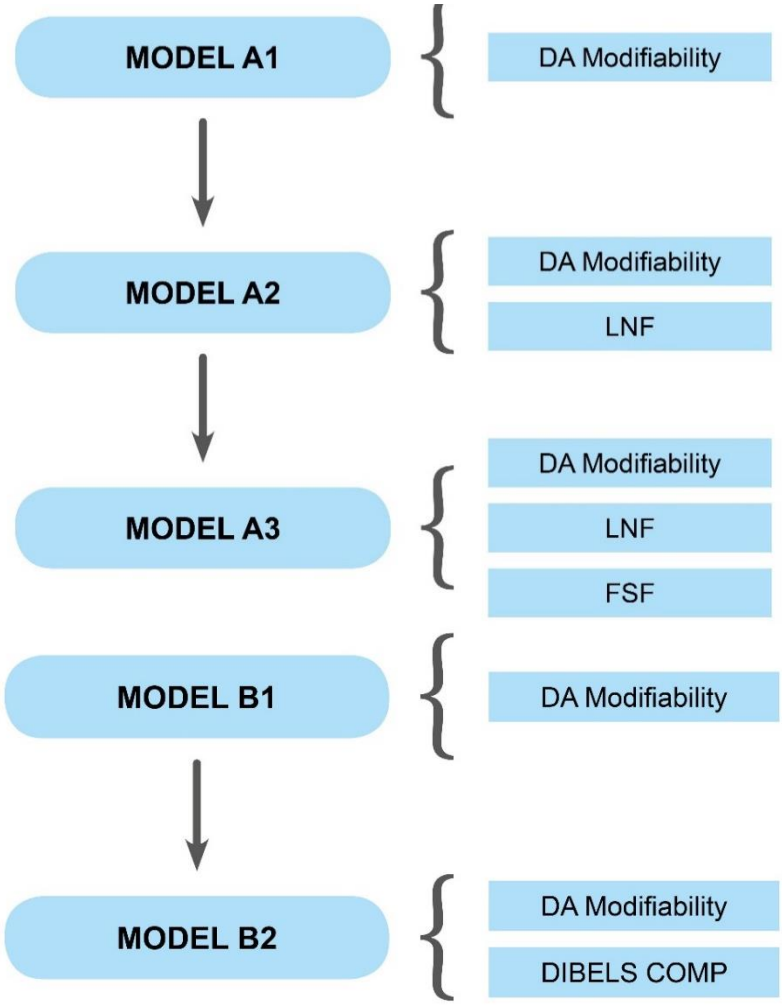


Table 4.8: Model comparisons (single predictor compared to combined predictors)*

	χ^2	χ^2 diff	df	p	Sensitivity	Specificity	AUC
A1. DA Modifiability	27.03				93%	77%	.91
A2. DA Modifiability, LNF	29.15	2.13	1	.15	86%	85%	.92
A3. DA Modifiability, LNF, FSF	29.34	2.31	2	.32	86%	85%	.92
B1. DA Modifiability	27.03				93%	77%	.91
B2. DA Modifiability, COMP	27.31	0.29	1	.59	79%	81%	.92

* Only the 67 children who completed the full dynamic assessment are included in these analyses

The combination of the DA Modifiability score with LNF, both LNF and FSF, or with the DIBELS COMP did not result in a significantly better model fit than that produced by the DA Modifiability measure on its own. However, combining the DA Modifiability measure with the static measures did improve the specificity, at the expense of somewhat reduced sensitivity. The sensitivity and specificity were the same for the model which included the DA Modifiability measure along with both the LNF and FSF, and that which combined the DA Modifiability measure and the LNF only. The addition of the static measures to the DA Modifiability measure resulted in a small non-significant increase in the AUC from .91 to .92.

In summary, when the results of these tests are taken together, combining the DA Modifiability measure with the static measures does not improve the model fit compared to that of the DA Modifiability measure alone. However, there appears to be some support for a model that combines both DA Modifiability and LNF measures to improve the level of specificity, but not sensitivity, compared to the dynamic measure on its own. This is discussed in greater detail in Chapter 5.

Analysis and classification for an applied setting

The present research focused on exploring the predictive validity and classification accuracy of the dynamic assessment in comparison to typical static assessments of emergent literacy skills. Overall, while both the dynamic predictor measures and the majority of the static predictor measures were significantly correlated with reading ability after a year at school, and were able to predict future reading status (i.e., reading difficulty/no reading difficulty), the dynamic measure of modifiability (DA Modifiability) produced the strongest correlations, best model fit, strongest predictive power, and best overall classification accuracy of all the static and dynamic measures administered at the start of schooling. This result was based on continuous scoring of the static and dynamic predictor measures at the start of school. While this approach provides more information and precision for evaluating these measures, interpretation of continuous scoring may not be as common in schools, and it can be more difficult to interpret. On the other hand, the use of a dichotomous risk score is easier for teachers to interpret as a way to determine whether a child is identified as at risk or not at risk for developing reading difficulties.

Therefore, to provide information that can be meaningfully understood and used in an applied setting (e.g., by a teacher or other assessor in a school setting), the dichotomised (*at risk*/*not-at-risk*) scores for the start-of-school static and dynamic predictor measures were also investigated and compared in terms of their predictive validity and classification accuracy. The DIBELS Next Composite score has been shown to provide a more accurate indication of a learner's reading ability than the DIBELS Next LNF or DIBELS Next FSF scores (Good et al., 2013). For this reason, the dichotomised version of the DIBELS Next Composite score, based on the cut-points-for-risk provided (Good et al., 2013), as well as the Dynamic Assessment Dichotomous score (DADS), were used for these analyses (refer to Chapter 3 for a discussion of the determination of the DADS score). For these analyses, the full sample was included (N = 136), as those children for whom testing was discontinued after the dynamic assessment pre-test were coded as *not at risk*.

Table 4.9 summarises the results from the logistic regression and ROC analyses for these dichotomised measures. While the model χ^2 statistics for both the DIBELS Next Composite (χ^2 = 23.15) and DADS (χ^2 = 36.32) indicated that these measures were significantly predictive of reading status, the DADS provided a significantly better model fit than the DIBELS Next COMP (χ^2 model difference = 13.17; p < .001). The pseudo-R^2 statics for the DADS measure demonstrated a moderate effect, while that for the DIBELS Next COMP was low-moderate. Furthermore, the DADS had a superior balance of sensitivity (81%) and specificity (90%) than that for the DIBELS Next COMP (sensitivity = 75%; specificity = 84%). The DADS also had a larger AUC (.86) than the DIBELS Next COMP (.80), although this difference was non-significant (p = .29).

Table 4.9: Model comparisons – dichotomous measures (individual predictors)

	-2LL*	χ^2	R^2_N	R^2_L	Wald	p	Sensitivity	Specificity	AUCs
COMP	75.37	23.15**	.30	.23	19.37	.00	75%	84%	.80
DADS	62.20	36.32**	.46	.37	26.69	.00	81%	90%	.86

* Null Model -2LL = 98.52 ** p < .001 Note: df for all = 1

Additional logistic analyses were conducted to investigate whether a combination of the DADS and COMP dichotomous scores would improve the model fit over that of the DADS alone. The results of these comparisons are given in Table 4.10.

Table 4.10: Model goodness-of-fit comparisons – dichotomous measures (individual predictor compared to combined predictors)

	χ^2	χ^2 diff	p	Sensitivity	Specificity	AUC
DADS	36.32			81%	90%	.86
DADS, COMP	37.61	1.29	.26	81%	90%	.88

Note: df for all = 1

The combination of the dichotomised DADS and DIBELS COMP did not significantly improve the model fit compared to DADS alone (χ^2 difference = 1.29, p = .26), and the sensitivity (81%) and specificity (90%) remained unchanged. Furthermore, the addition of the DIBELS COMP dichotomous measure resulted in a non-significant increase in the AUC from .86 to .88 (p = .37).

In summary, the dynamic measure (DADS) had superior model fit, predictive power, and classification accuracy than the dichotomised DIBELS Next Composite score. Furthermore, a combination of the dynamic measure (DADS) and the dichotomised static DIBELS Next Composite score did not improve the model fit and classification accuracy over that of the DADS measure alone. Table G.1 and Table G.2 in Appendix G provide full summaries of the fitted models using the T2 outcome variable and the dichotomised dynamic (DADS) and static (DIBELS Next Composite) predictor variables.

Summary

Static measures of phonological awareness (DIBELS Next First Sound Fluency) and alphabet knowledge (DIBELS Next Letter Naming Fluency) administered to children upon school entry proved to be predictive of their reading ability after a year at school, but these measures demonstrated floor effects. The static measure of rapid object naming (CTOPP-2 RON) was not significantly correlated with reading ability after a year at school. All post-test scores for the dynamic assessment measures administered at the start of schooling were predictive of reading ability after a year at school. With the exception of the scores from the dynamic assessment pre-test (which operates as a static assessment), the dynamic assessment measures did not suffer from the same floor effects exhibited by the static measures.

In terms of the ability to predict reading difficulty after a year at school, the continuous static measures of phonological awareness and alphabet knowledge (DIBELS Next First Sound Fluency and Letter Naming Fluency), as well as the composite score of the aforementioned

(DIBELS Next Composite), were able to predict reading status (reading difficulty/no reading difficulty) after a year at school. The continuous dynamic assessment measures of Learning, Strategy, and Modifiability (a combination of the DA Learning and DA Strategy scores) were also predictive of reading status, and the DA Modifiability score provided the greatest predictive power and superior classification accuracy of all the measures administered. The combination of the DA Modifiability score and the DIBELS Letter Naming Fluency predictor measure did not result in a significant improvement of predictive power or classification accuracy over the DA Modifiability measure on its own, and although there was an improvement in specificity, this was at the expense of sensitivity.

The dichotomised versions of the static and dynamic measures were also investigated in terms of their ability to predict reading status and provide an accurate classification of reading difficulty risk. This was done to provide information that would be most useful in an applied setting such as a school. The results indicated that the dichotomised dynamic assessment measures (DADS) demonstrated superior predictive ability and classification accuracy to that of the dichotomised static measure (DIBELS Next Composite score). In addition, combining the dynamic measure (DADS) and dichotomised static measure did not improve the model fit and classification accuracy over that of the dynamic measure alone.

Chapter 5:
Discussion

The aim of this study was to investigate whether a dynamic assessment of decoding, administered to New Zealand children when they started formal schooling, was able to predict reading difficulty after a year at school. Being able to identify, upon school entry, those children at risk of developing a reading difficulty makes it is possible to provide early intervention to prevent or ameliorate such a difficulty. To achieve this aim, the predictive validity and classification accuracy of commonly used static measures of emergent literacy and a dynamic measure of phonological decoding, administered to children upon entry to school, were examined and compared. The results of these analyses were outlined in the previous chapter and indicate that a dynamic assessment of decoding can predict reading difficulty and can do so with greater accuracy than static measures of letter naming fluency (DIBELS LNF), phonemic (first sound) awareness (DIBELS FSF), and rapid object naming (CTOPP-2 RON). In this chapter, the key findings are discussed and contextualised in terms of the extant literature related to the early identification of children at risk of reading difficulties. The discussion begins by reviewing the relative ability of the static and dynamic measures administered upon school entry to predict reading outcomes after a year at school. This is followed by a discussion of possible explanations for the superior predictive ability of the dynamic measure. The discussion ends with a review of key features of an effective screening tool, extending beyond the predictive ability and classification accuracy, to consider how the measures used in this study may exhibit these other important features.

Predictive ability of reading measures administered upon school entry

At the start of formal schooling (T1), children were administered static assessments of letter naming (DIBELS Next LNF), phonemic awareness (first sound) (DIBELS Next FSF), and rapid object naming (CTOPP-2 RON). During the same testing session, children were also administered a dynamic assessment of decoding (pseudowords). The dynamic assessment employed a protocol similar to that used by Petersen and colleagues (2016), and consisted of a pre-test which, similar to a static assessment, measured the child's current ability to decode pseudowords independently, without input from the assessor; a teaching stage in which the assessor taught the child a sound-by-sound strategy to decode the pseudowords; and a post-test, where the child was asked to apply, independently, the strategy they had to been taught to decode the same pseudowords, as well as two novel pseudowords consisting of the same letters in a different order. After a year at school (T2), the same children were tested again using measures of letter naming, phonemic segmentation, nonsense word decoding, and sight-word reading skills.

To investigate and examine the predictive ability of the static and dynamic measures administered upon school entry, correlational analyses, logistic regression, and predictive classification accuracy analyses were undertaken. The correlational analyses examined the relationship between scores on the early reading measures and reading ability after a year at school, and the logistic regression and classification analyses were used to examine the ability of these measures to accurately predict reading difficulty status after a year at school. In addition to examining the predictive ability of individual static and dynamic measures, hierarchical logistic regression analyses were undertaken to determine if combining the static and dynamic measures resulted in improved predictive ability. Finally, to provide information that can be easily interpreted and used in a school setting, the dichotomised scores (at risk/not at risk) for

the T1 static and dynamic predictor measures were investigated in terms of their predictive validity and classification accuracy. In this section, the results from these analyses are discussed.

Predictive ability of static measures

Static measures used in the present study were selected because of their widespread use and because previous research has found that tests of the emergent literacy skills of alphabet knowledge, phonemic awareness, and rapid naming, administered even before formal reading instruction has started, are predictive of future reading outcomes (Bishop & League, 2006; Catts et al., 2014; Compton et al., 2006; Kantor et al., 2011; O'Connor & Jenkins, 1999; Schatschneider et al., 2004). Consistent with past research, the findings from the present study showed that both the letter naming (DIBELS Next LNF) and initial sound identification tasks (DIBELS Next FSF) were able to predict reading ability, as well as reading status, after a year at school. The DIBELS Next Composite score, which combines the scores for the letter naming (DIBELS Next LNF T1) and initial phoneme identification (DIBELS Next FSF) task, had the strongest correlations across all the T2 outcome measures. This is in line with the findings of the developers of the DIBELS Next test who indicate that the Composite score provides the best overall estimate of children's emergent literacy skills (Dynamic Measurement Group Inc, 2010).

Of the individual static measures, the T1 letter naming task (DIBELS Next – LNF) had the strongest correlations with the T2 outcome measures. It also was also the best static predictor of reading difficulty after a year at school. Furthermore, combining the letter naming and first sound fluency measures did not significantly improve prediction over that of the LNF measure on its own. This aligns with the findings of numerous other studies that have found that letter naming ability at the start of formal schooling is significantly predictive of future reading outcomes and that letter naming speed, in particular, is the strongest single predictor of future early reading outcomes (e.g., Burke, Crowder, Hagan-Burke, & Zou, 2009; Catts et al, 2015;

National Early Literacy Panel, 2008; Schatschneider et al., 2004). The predictive ability of letter naming speed may lie in the fact that the rapid and efficient access of sub-lexical information such as letter sounds and letter names facilitates word reading, which in turn is the basis of fluent reading (Clemens, Lai, Burke, & Wu, 2017; Schatschneider et al., 2004). Being able to correctly identify the sounds of letters is clearly fundamental to phonological decoding ability, and this in turn forms the basis of orthographic learning (Ehri, 2014; Share, 1999). While the ability to rapidly access letter names may not be directly involved in phonological decoding in the same way, it may instead be an indicator of the child's ability to correctly and rapidly access orthographic information, as retrieving a letter name is comparable to retrieving a word (or syllable) from memory (Clemens et al., 2017). This information is important if one considers that children with reading difficulties have difficulty acquiring orthographic information (Suárez-Coalla et al., 2014).

In addition to being a significant predictor of reading difficulty after a year at school, the letter naming task produced the best overall classification accuracy of the static predictor measures, and a sensitivity level (i.e., proportion of true positives) above that of the recommended minimum of 80% (Bridges & Catts, 2011; Compton et al., 2006; Petersen et al., 2016). However, despite producing the best specificity level (i.e., proportion of true negatives) of the static predictor measures, the 66% specificity still fell well short of the recommended minimum specificity value of 80%. This result means that 34% of children who did not go on to have a reading difficulty at T2 would be identified as *at risk* using the letter naming measure at T1. This mirrors the results of numerous other studies which found that static measures tend to over-identify risk of reading difficulty (e.g., Bridges & Catts, 2011; McCardle et al., 2001; O'Connor & Jenkins, 1999; Petersen et al., 2016).

Unlike some studies which found rapid naming of objects to be a signficant predictor of future reading status (e.g., Catts et al., 2015; Cronin, 2011; Norton & Wolf, 2012), in this study the CTOPP-2 RON was not significantly predictive of future reading ability or reading difficulty status. This may be because, in contrast to all the other static measures administered at T1, the rapid object naming task (CTOPP-2 RON) exhibited a ceiling effect which may indicate that this task was too easy for most of the children. Like floor effects, ceiling effects limit variability in the data for a predictor variable and can therefore weaken the correlation between that variable and an outcome variable. Almost all the children quickly grasped what was required in the rapid object naming task and even those children who did not speak English as their first language, were able to quickly learn the names of objects with which they were unfamiliar (during the practice phase of the test). Although the rapid object naming task was not predictive in this study, this does not discount entirely the role of rapid naming, in general, in the prediction of reading status. Rapid naming can be tested in a variety of ways including the rapid naming of familiar objects, colours, letters, and numbers. At the start of formal schooling children are still learning letters and numbers and are therefore not able to name these as quickly as familiar objects or colours. In this study, the decision was made to include a measure of rapid naming of objects because (a) the children were at the start of formal schooling when they are more likely able to complete object or colour naming tasks than naming letters or numbers; (b) letter naming fluency was already included as part of the testing battery (DIBELS Next LNF); and (c) colour naming tasks may be problematic as an estimated 6% of New Zealand males have some form of colour deficiency (New Zealand Health Technology Assessment Clearing House, 1998).

Despite the fact that familiar objects or colours are more commonly used with children at the start of schooling, previous research has shown that rapid naming tasks involving symbolic items such as letters and numbers are more predictive of future reading than that of non-symbolic items such as colours or objects (Wagner et al., 2013). The results of this study seem to indicate that a letter naming task that includes a rapid naming element (e.g., the DIBELS Next LNF task), does indeed offer superior predictive ability than rapid naming of objects.

Predictive ability of dynamic measures

A dynamic assessment of decoding was examined in the present study because it has the potential to be a better predictor of reading difficulty risk than traditional static assessments (e.g., Bridges & Catts, 2011; Cho & Compton, 2015; Fuchs et al., 2011; Petersen & Gillam, 2015). The importance of early reading measures to be able to accurately predict future reading outcomes should not be underestimated. On the one hand, providing the targeted, intensive intervention needed to improve outcomes for children identified as at risk for a reading difficulty is resource-heavy. Over-prediction could result in unnecessary allocation of these costly resources to children who would have been able to achieve age-appropriate reading outcomes with high quality instruction in a general classroom setting. On the other hand, failing to correctly identify those children who are at risk, means that these children may not get the help they need until much later when reading failure becomes evident. Not only could this have lasting negative consequences for the child's reading and academic development, but later intervention to remediate an existing difficulty is significantly more difficult and resource-heavy than early intervention aimed at preventing a reading difficulty from developing (Compton et al., 2010; Kantor et al., 2011; Torgesen, 2002).

Results showed that all three of the dynamic assessment post-test scores (Learning, Strategy, and Modifiability) were predictive of future reading ability and reading difficulty status. This supports the findings of several other studies that a dynamic assessment of decoding is able to predict future reading ability and reading difficulty status (e.g., Bridges & Catts, 2011; Elbro et al, 2012; Petersen et al., 2016; Petersen & Gillam, 2015; Petersen et al., 2018). All the dynamic measures, except for the DA Pre-Test scores, correlated significantly with the T2 outcome measures, and in general, these correlations were stronger than those between the static predictor measures. All three of the dynamic assessment post-test scores proved to be significantly predictive of reading difficulty status and produced greater predictive power than any of the static measures. The DA Modifiability score had the strongest predictive power of the dynamic measures and was also a better predictor of reading status than any of the static measures.

In addition to predictive power, the DA Strategy and DA Modifiability measures were both more sensitive than any of the static measures, and the DA Modifiability score also produced a specificity level superior to that of any of the static measures. This superior ability of dynamic measures over static measures to correctly identify children who are not at risk (i.e., specificity) supports the findings of numerous other studies (e.g., Bridges & Catts, 2011; Petersen et al., 2016; Petersen & Gillam, 2015; Spector, 1992). Of all the measures administered at the start of formal schooling, the DA Modifiability measure produced the best overall classification accuracy, with the best balance of sensitivity (93%) and specificity (77%) of all the measures. Despite the sensitivity for the Modifiability measure falling just slightly short of the generally recommended value of 80%, it was still meaningfully higher than that achieved by any of the static measures (45% to 66%). This result means that only 23% of children who did not go on to have a reading difficulty at T2 would be identified as *at risk* using the measure at T1, compared to 34% for the best performing static measure – letter naming. Furthermore, because of the small number of children identified as having a reading difficulty in this study, the specificity values need to be

interpreted with caution as a small difference in the proportion of false positives to true positives can have a substantial impact on these values.

Overall, comparative results suggest the dynamic measures provided better overall classification accuracy than the static measures. In 91% of cases the DA Modifiability measure correctly assigned a higher probably of reading difficulty to children who went on to have a reading difficulty at the end of a year at school, whereas for the static measures only between 66% and 87% of cases were correctly identified. The dynamic measures also demonstrated superior sensitivity and specificity over the static measures in the present study.

Comparison of individual and combined predictors

In practice, combining a battery of measures to predict future reading difficulty may not be time or resource efficient, however, it is useful to investigate the outcomes of combined measures to understand whether the combination of measures results in optimal prediction. In the present study, analyses were conducted to investigate whether the combination of the static and dynamic measures would improve the predictive ability over that of the DA Modifiability measure on its own. Combining the DA Modifiability score with the DIBELS Next LNF, both the DIBELS Next LNF and the DIBELS Next FSF, or with the DIBELS Next Composite, did not improve prediction compared to that of the Modifiability measure alone. These findings are in line with that of other researchers who have found some support for the use of dynamic assessment as the primary measure to predict risk of reading difficulty, rather than as supplementary to traditional static assessments of emergent literacy skills (e.g., Bridges & Catts, 2011; Petersen et al., 2016; Petersen & Gillam, 2015). However, the addition of the DIBELS Next LNF to the DA Modifiability measure did result in an improvement in specificity (from 77% to 85%), moving this value to above the suggested 80% minimum. The improvement in specificity was at the expense of sensitivity (from 93% to 86%), but the sensitivity level was still within acceptable bounds. As adding the DIBELS Next LNF to the DA Modifiability measure improved the specificity, this may

indicate that including the LNF in a two-stage process could be helpful in improving the overall classification accuracy, Decisions regarding whether it is better to use the dynamic assessment measure on its own or together with the LNF will also need to take into consideration whether this slight improvement in prediction warrants the additional resources and time needed to administer them.

Analysis and classification for an applied setting

To provide information that can be easily interpreted and used in a school setting, the dichotomised scores (at risk/not at risk) for the T1 static and dynamic predictor measures were investigated in terms of their predictive validity and classification accuracy. For the static measure, children categorised as *intensive* (i.e., likely to need intensive intervention such as small group or individual instruction) for the DIBELS Composite score were classified as *at risk* and given a dichotomised score of 1 (0 = *not at risk*). The at risk/not at risk scores for the dynamic assessment (Dynamic Assessment Dichotomous score - DADS) were calculated using the procedure outlined in Chapter 3. The results partly mirrored those achieved using the continuous static and dynamic measures. While both the static and dynamic measures were predictive of reading difficulty, the dynamic measure dichotomous score proved to be the better of the two in terms of both predictive power and classification accuracy, with a better balance of sensitivity and specificity than the dichotomised DIBELS Next Composite score. In terms of what is considered acceptable minimum levels of sensitivity and specificity (80%), the DADS produced adequate sensitivity (81%) and good specificity (90%). Although this classification accuracy is acceptable, higher levels of sensitivity than specificity would generally be preferred particularly because, using the DADS, 19% of children who went on to be classified as having a reading difficulty after a year at school, were identified as *not at risk* at T1. Greater sensitivity may be possible to achieve if slightly different cut points were used for the calculation of the DADS; however, the difficulty is that improved levels of sensitivity frequently results in reduced

specificity. The level of sensitivity for the dichotomised DIBELS Next Composite (75%) was lower than what is generally considered a minimum acceptable level of sensitivity for a screening tool but showed adequate specificity (84%). This pattern of higher specificity than sensitivity for the DIBELS Next Composite measure is in line with findings of Dewey and colleagues (2015), but other studies have shown a different pattern of classification accuracy, with sensitivity levels being higher than those for specificity (e.g., Petersen & Gillam, 2015; Petersen et al., 2016).

The sensitivity and specificity values achieved by the DADS are similar to those in the study by Petersen and colleagues (2016) which employed a similar dynamic assessment of decoding (sensitivity = 92%; specificity = 83%). Petersen and colleagues (2016) suggested that the classification accuracy of the dynamic measure may be improved by increasing the number of pseudowords in the test and by obtaining a measure of generalisability by using pseudowords at post-test that have the same letters in a different order (e.g., *pog* … *gop*). Both these suggestions were incorporated into the dynamic assessment used in this study: There were six pseudowords in the main pre-test and teaching phases (compared to the four used by Petersen and colleagues), and two new pseudowords were included at post-test. The results from this study do not provide support for these recommended changes to the procedures used by Petersen and colleagues (2016) because, although a higher level of specificity (90%) was achieved in the current study, the sensitivity was notably lower (81%) than that achieved by Petersen and colleagues.

The combination of the dynamic dichotomised score and static dichotomised score was also investigated and results indicated that a combination of the dynamic and static measures did not improve prediction over that of the DADS alone. Given this and the fact that the results using the continuous measures showed only a negligible improvement in classification accuracy when the dynamic measure was combined with the letter naming measure (DIBELS Next LNF), this would lend some support to the use of the DADS on its own, as a tool to predict reading difficulty in an applied (school) setting.

Why did the dynamic measures more accurately predict reading outcomes?

Although almost all the static and dynamic measures proved predictive of reading ability and status after a year at school, in general, the T1 dynamic measures had superior predictive ability than the T1 static measures. One possible explanation for this is that the static measures exhibited floor effects, a finding consistent with numerous other studies with children this age (e.g., Aaron et al., 2008; Anthony & Francis, 2005; Bridges & Catts, 2011; Torppa et al., 2007). Floor effects can have a negative impact on the predictive ability of measures as they result in weaker correlations between children's scores on the predictor measures and later outcome measures (Catts et al., 2009). Often these floor effects occur because children lack the experience or prior learning needed to be able to perform the tasks, especially when administered to children at the start of their formal schooling (Catts et al., 2009). In such a case it becomes difficult to distinguish between children who are not able to perform the task because they are at risk of reading difficulty, and those who are not at risk and would, if given some support, be able to perform the task. Furthermore, the way in which static assessments are administered means that a child's inability to perform a task within the test context may not be as a result of their lack of skill, but rather because they are unable to understand the test instructions (Bridges & Catts, 2011; Petersen et al., 2016). Static assessments aim to ensure standardisation and avoid assessor impact on the results and therefore interactions between the assessor and child being assessed are restricted to a basic instruction and perhaps a limited practice of the task to be completed. Under these conditions, very young children and those who have limited educational and testing experience, may be particularly likely to struggle to understand what they are required to do (Caffrey et al., 2008; Spector, 1992). As floor effects are frequently a result of young children lacking the experience or prior learning needed to successfully complete the tasks or understand the test instructions,

the presence of floor effects can prove problematic for a universal screening tool administered at school entry; the tool is likely to lack specificity, with many children being incorrectly identified as at risk of reading difficulty (e.g., Catts et al., 2009).

In this study, a lack of understanding of the task requirements was particularly evident in the case of the first sound fluency task (DIBELS Next – FSF), indicating that this is likely one contributing factor to the floor effect for this measure. For example, when asked during the practice phase, "What is the first sound you hear in the word 'dog'?", some children would give the letter name rather than the sound and would continue to say letter names for the remaining practice words and the words in the main test itself. This may be because the children did not have the necessary metalinguistic skills needed to comprehend the meaning of 'sound' as opposed to 'letter name', or because the bulk of their prior learning experiences had been in naming letters of the alphabet, rather than phonemic awareness tasks such as identifying the first sound in a word. In contrast to the first sound fluency task, most children quickly understood what was required for the letter naming (DIBELS Next LNF) task. Therefore, floor effects in the case of the letter naming task are more likely linked to lack of ability to name the letters as opposed to difficulty understanding the task instructions. A quarter of the children could only read six or fewer letters, and in most cases, the letters these children could read were the same letters repeated in the test (e.g., *s* repeated both in upper and lowercase form). Letters identified were generally those that look almost identical whether presented in uppercase or lowercase (e.g., *s*, *v*, *x*, *w*) or which they might recognise from their own name or that of a family member.

In the case of the dynamic assessment pre-test, floor effects appeared to occur because of both an inability to understand the task instructions as well as an inability to perform the task. Some children who clearly understood the task instructions (e.g., when asked to read the words, or to identify any sounds they knew, they would reply with, "I don't know how to read

any words/sounds. I only know how to read this sound because it is in my name", etc), were unable to perform the task. Others, who later went on to have no difficulty learning how to decode the pseudowords, misunderstood what was required in the task: When asked to read the words, several children provided the letter names in the words and did not understand that they should instead say the sounds. There were also a few children who tried to read the pseudowords as a real word (e.g., *nep* read as *nip*).

In contrast, the dynamic assessment post-test measures for post-test sounds, learning, strategy, and modifiability did not suffer floor effects, and were normally distributed. This is in keeping with several earlier studies that demonstrated that, contrary to static assessments, dynamic assessment did not display floor effects when administered to young children (Bridges & Catts, 2011; O'Connor & Jenkins, 1999; Petersen et al., 2016; Petersen & Gillam, 2015). It seems likely that the inclusion of a teaching phase in the dynamic assessment helped to avert these floor effects. This is possibly because the input given during the assessment helped children understand what was required in the task. For example, Spector (1992) theorised that a possible reason why dynamic assessment shows greater sensitivity than static assessment is that it is a 'cleaner' measure of ability, as static assessments may place demands on ancillary cognitive skills. By including a teaching phase where the assessor not only gave verbal instructions for what was required, but also demonstrated how to perform the required task, difficulties around understanding the requirements of the task were likely mitigated. As Catts and colleagues (2009) point out, if floor effects are indeed the result of experiential shortcomings in the young child, then the additional experience or knowledge the child gains during the instruction phase of the dynamic assessment may be sufficient to reduce these floor effects. Furthermore, the inclusion of the teaching phase meant that the dynamic assessment was able to differentiate between those children who were unable to perform the skill during the pre-test phase: It helped distinguish between those children who were already close to being able to perform the skill and who could

master the skill with only minimal input, and those who would need intensive input, or who may not be able to perform the skill even with extensive assistance.

Petersen and Gillam (2015) offer another possible explanation for why the dynamic assessment was better able to predict future reading outcomes, namely that it allows an actual word reading task (pseudoword decoding) to be used, rather than testing an emergent literacy skill such as alphabet knowledge or phonological awareness. Research has shown that pseudoword decoding is the best predictor of a child's decoding skill at the start of their schooling and beyond (National Early Literacy Panel, 2008). Furthermore, numerous studies have shown that children with reading difficulties performed poorly in pseudoword decoding tasks (e.g., Herrmann, Matyas, & Pratt, 2006; National Early Literacy Panel, 2008; Rack, Snowling, & Olson, 1992). Unfortunately, most children of this age do not have the phonological decoding knowledge to sound out pseudowords, and therefore measures of pseudoword decoding knowledge are not considered suitable for this age group. Instead, emergent literacy skills that are precursors of phonological decoding (alphabet knowledge and phonological awareness) are generally tested at this stage. However, because it includes a teaching phase, a dynamic assessment is able to assess the child's ability to master a skill they were not capable of independently performing but which is within their *Zone of Proximal Development*; in this case, the application of a sound-by-sound strategy to phonologically decode pseudowords. Furthermore, given that the scoring for the dynamic assessment included points for individual letter sounds produced correctly, it is likely that the dynamic assessment is also tapping letter sound knowledge which, along with, letter naming skill, is a primary aspect of the alphabetic principle (Clemens et al., 2017). The ability to apply the alphabetic principle, along with phonemic awareness, provides the foundation for being able to phonologically decode words which, in turn, affords the child the opportunity to acquire the orthographic representations required for sight word reading (Ehri, 2014; Share, 1999; Share, 2004).

Of the different scores produced by the dynamic assessment of decoding, it was the DA Modifiability score which proved to have superior predictive ability. Not only did it correlate most strongly with all the T2 outcome measures, but it also showed superior ability to correctly predict reading difficulty status after a year at school. The DA Modifiability score is a combination of the DA Learning and DA Strategy scores, capturing both how well the child responds to input and the child's ability to apply a taught decoding strategy and/or rapidly learn new orthographic representations. It therefore acts as a measure of the child's overall modifiability or response to input. By adding the subjective score (DA Learning) to the more objective DA Strategy score, the specificity was improved over that of the DA Strategy score on its own but did nothing to reduce the sensitivity. In other words, adding the assessor's judgement on learning behaviours related to successful learning of reading skills (i.e., the DA Learning score) helped to distinguish the true-positives from false-positives (i.e., improving specificity). This mirrors the findings by Petersen and colleagues (2016), who found that the dynamic assessment measure which combined the child's learning and strategy scores correlated more strongly with the outcome measures than any of the other dynamic assessment measures. This would provide some support for the hypothesis that a measure of the construct of responsiveness to instruction correlates more strongly with future reading ability than does a change in ability to perform a task after input has been given (i.e., pre-test-post-test-gain). This may be because the modifiability score is not only capturing how well the child is able to apply a taught strategy or acquire new orthographic representations, but also their learning behaviours and how these impact on the child's ability to respond to instruction.

Additional characteristics of an effective screening tool

In the context of the prediction of risk for future reading difficulties, predictive ability and classification accuracy are not the only criteria that can impact on the utility of a universal screening tool. Other important criteria include that the tool should focus on providing information for remediation (not only categorisation), and that it should be quick, easy, and cost-effective to use within the target environment. A tool that is able to meet these criteria is likely to have higher acceptability with teachers and schools. While these aspects were not specifically investigated in this study, and therefore empirical results in this regard are not available, some preliminary observations on these features can be offered based on the experiences of using the static and dynamic assessments in this present study.

Providing information for targeted intervention

The main purpose of early prediction of reading difficulties is to be able to provide early, targeted intervention. Therefore, the aim of a prediction tool should be to provide data that will be useful for such intervention. For example, while the level of mother's education has been shown to be a good predictor of future reading difficulties for their child (e.g., Catts, Fey, Zhang, & Tomblin, 2001) this information is potentially only useful in the categorisation of the child in terms of their risk status; it does not provide any useful information in terms of interventions for the child, nor is it a malleable characteristic that teachers can influence or intervene directly on. This is where, in addition to its potential to provide superior predictive power and classification accuracy than traditional static measures, dynamic assessment might be more useful than static assessments in terms of the type of information it renders. While the static measures employed in this study provided an indication of the child's letter naming and phonemic awareness skills, the dynamic assessment provided this information, as well as additional information which may be able to inform intervention. From the post-test sound and word scores, as well as the strategy score,

information may be gleaned as to the child's ability to rapidly acquire orthographic representations or to apply a strategy they have been taught to phonologically decode pseudowords, a skill which has shown to be the best predictor of future decoding ability. This ability to decode pseudowords can, in turn, be seen as the foundation of future orthographic learning and reading ability (Ehri, 2014; Share, 1999). The strategy score, as conceptualised in this study, also includes a measure of the child's ability to apply the strategy they have learnt to an analogous situation (i.e., a pseudoword consisting of the same letters as a word they have been taught, but in a different order). This provides information regarding the child's ability to both learn and apply the strategy they have learnt, not just their ability to repeat the letters or words they have been taught (possibly relying on short-term memory in some cases).

Furthermore, the DA Learning score reveals a range of information about the way in which the child responds to input in terms of specific behaviours exhibited during the teaching phase that could hinder or promote the child's learning. The assessor assigned a DA Learning score from 1 to 5 to the child, guided by a Learning Scale used to focus the assessor on a range of behaviours exhibited by the child during the teaching phase of the assessment. These behaviours include internal social-emotional behaviours (anxiety, perseverance, motivation); external social-emotional behaviours (attention, tractability, task confidence); and cognitive arousal (task comprehension, errors). By assigning scores for each of these areas, the assessor can get a more detailed indication of the specific areas where the child is experiencing any difficulty, and therefore, areas to focus on when providing the child with intervention, or additional support in standard classroom instruction. For example, in some cases, a child demonstrated a lack of ability to attend to the task (low attention score), despite an apparent interest and desire to participate in the task (high motivation score), being cooperative, responsive, and confident (high tractability and task confidence scores), quickly understanding the task (high task comprehension score), and being able to correctly imitate the phonological decoding strategy demonstrated by the assessor (high

errors score). This is an indicator of a possible difficulty around attention rather than, for example, a phonological deficit or other reading-specific difficulty which may be hindering the child's ability to learn how to decode the pseudowords. Conversely there were some children who showed a clear lack of ability to discern and/or replicate the sounds produce by the assessor in the teaching stage (low Errors score on the Learning Scale), despite their interest in the task, ability to attend to the task, cooperativeness, task confidence, etc. These difficulties were particularly evident in terms of an inability to distinguish between plosive sounds (/p/, /t/, and /k/), and nasal consonants (/n/ and /m/) and may point to a difficulty in phonological awareness that would require targeted input in this area. The combined focus on domain specific intervention (e.g., early reading skills such as phonological awareness) and behavioural support strategies leads to greater gains in reading than reading-only interventions (Lane, Menzies, Oakes, & Kalberg, 2012; Stewart, Benner, Martella, & Marchand-Martella, 2007). By incorporating both behavioural assessment and domain specific reading assessment, dynamic assessment may enable reading and behaviour needs to be identified, to more fully inform intervention.

In addition to providing information about an area(s) in which the child may have a specific difficulty, the dynamic assessment also yields information about specific strengths such as strong motivation or ability to attend to input that the child may have, and which could be leveraged to support their learning (Brooks & Weeks, 1998; New Zealand Ministry of Education, 2015). This valuable information is not only available for children identified as *at risk*, but also those who may be on the 'boundary' of reading difficulty risk (i.e., those classified as *not at risk*, but who nevertheless needed to complete the input and post-test phases of the dynamic assessment). These boundary cases are the children most likely to be incorrectly classified in terms of reading difficulty risk, and thus having this information available for those boundary cases classified as *not at risk* would enable close monitoring of these children, particularly in the areas that they have demonstrated potential difficulties.

Quick and straightforward test implementation

Cost, training, time, and specialized equipment involved in assessment are important to consider in terms of the acceptability of a particular tool. Tests that may be superior in terms of predictive validity and classification accuracy under research conditions, may not be practical or palatable for use in the real world of the classroom and school context. For example, many research-based assessments are time-consuming, conducted by highly trained individuals, and use specialized equipment such as computer software. As a result, while these tests may perform well in terms of validity and classification accuracy, they may be impractical, and therefore not used, in the school environment. In the case of the present study, although both the static and dynamic measures need to be individually administered, this can usually be easily accommodated within the New Zealand school system, where children enter school throughout the year, on or close to their fifth birthday. The current rolling enrolments of students into the reception classroom will likely make it convenient for the teacher to individually test a child shortly before or after they have begun their formal schooling.

The static measures selected for use in this study, are all quick and easy to administer, and produce results that are easy to interpret. For the present study, the static measures each took around 1½ to 2 minutes to administer (including instruction and short practice of test items), and the administration, using a standardised protocol, was straightforward. Marking of the test items was also very quick and easy.

The dynamic measure was also easy to administer, score, and interpret within the school context, and would likely require only brief training of teachers prior to use. While the dynamic assessment took a bit longer to administer (between 5 and 8 minutes), as mentioned earlier, it allowed for more information to be gathered regarding the child's reading and behavioural strengths and weaknesses (profile). The dynamic assessment also uses a standardised protocol,

making it quick and easy for the assessor to learn and apply. Marking was slightly more complex than that for the static assessment and judgement was required on the part of the assessor to determine the DA Learning score. Even though the assessor for this study was not a new-entrant teacher, the dynamic assessment produced good predictive classification accuracy. As new-entrant teachers may be superior judges of the children's learning behaviours, if the test was administered by teachers themselves, classification accuracy of the dynamic assessment may be improved. Furthermore, although the dynamic assessment took slightly longer to administer, and was slightly more complex to mark, it has the potential to provide very useful information to support decisions regarding the most effective intervention(s) for those children at risk of developing a reading difficulty. Furthermore, it provided a much more accurate prediction of future reading difficulty than did the static measures.

Summary

The findings of this study provide support for the use of a dynamic assessment of decoding, administered to children upon school entry within the New Zealand context, as an effective universal screening tool for future reading difficulty. The results indicated that a dynamic assessment of decoding administered at the start of formal schooling can predict reading ability and whether a child will have a reading difficulty after a year at school. Static assessments of letter naming fluency and phonemic awareness (first sound) were also able to predict reading ability and reading difficulty status, however, unlike the dynamic assessment, these measures suffered from floor effects. Furthermore, the dynamic assessment showed superior predictive power and classification accuracy to that of the static measures. In particular, the DA Modifiability score which captures both the child ability to apply a taught decoding strategy or to acquire orthographic representations, as well as their ability to respond to instruction, proved to have superior predictive validity and classification accuracy to all other measures. This mirrors the research of Petersen and colleagues (2016) who found that a modifiability score had superior predictive ability to a range of static measures of emergent literacy, as well as other dynamic assessment measures. Possible explanations for the dynamic measure's superior ability to predict future reading outcomes are that it does not suffer from floor effects in the same way that static measures administered at this early stage do and that it is able to assess an actual reading skill (phonological decoding) rather than a precursor emergent literacy skill. The reduction of floor effects and the ability to assess an actual reading skill are both made possible due to the inclusion of a teaching stage within the dynamic assessment.

Comparisons of models which included only the dynamic assessment measure and those which included a combination of the dynamic assessment and static measures indicated that the addition of the static measures did not improve prediction over that of the dynamic assessment on its own. This is in keeping with Petersen and colleagues (2016) who found that combining the dynamic assessment with static measures did not improve the classification accuracy over that of the dynamic assessment on its own. However, there was some indication that the addition of the static letter naming task (DIBELS Next LNF) to the dynamic assessment may result in a slightly improved specificity (but not sensitivity).

In addition to its superior ability to predict reading ability and accurately classify children at risk of reading difficulty, there is also some indication that the dynamic assessment of phonological decoding may be able to demonstrate additional important characteristics of an effective screening tool, namely to provide information that can usefully inform and support targeted intervention for children at risk of a reading difficulty, and to be practical to administer within the New Zealand school context. Implications for the selection and use of measures and recommendations for future research will be addressed in the next chapter.

Chapter 6:
Conclusion

The importance of providing children at risk of reading difficulty with early, intensive, and targeted intervention, is widely acknowledged (Bishop & League, 2006; Bridges & Catts, 2011; Huang et al., 2014; Lonigan et al., 2013). As earlier intervention has been found to be more effective and less resource-heavy than later intervention, the identification of reading difficulty risk needs to occur as early as possible, preferably before the child starts formal reading instruction (Kantor et al., 2011; Morlini et al., 2014; Tunmer & Greaney, 2010). To provide this early intervention, children who are at risk of developing a reading difficulty need to be accurately identified. A universal screening tool needs to be able to accurately predict future reading ability and correctly classify children in terms of their risk of developing a future reading difficulty. To do so, the screening tool should consist of measures of the skills most likely to predict future reading ability. Key factors in determining these skills are the age and reading development stage of the children to be tested. In the case of novice readers, the ability to phonologically decode pseudowords has been found to be the best predictor of a child's later word reading ability (e.g., National Early Literacy Panel, 2008; Rack et al., 1992), and deficits in pseudoword reading have been found to be related to reading difficulty (e.g., Herrmann et al., 2006). However, in the case of children who have not yet begun formal reading instruction, assessment of pseudoword decoding using traditional static assessments is not generally considered feasible because most children have little or no decoding ability at this stage. For this reason, emergent literacy skills such as alphabet knowledge and phonological awareness are most commonly assessed. There is abundant evidence that traditional static tests of these skills can predict future reading ability. However, when administered to very young children, these tests are typically plagued by floor effects and as a

result are less effective in accurately distinguishing between those children at risk of developing a future reading difficulty, and those who are not. One possible reason for this is that the results on static measures may be influenced by a young child's inability to understand what is required of them as very little or no assessor support is provided during the test. In contrast, dynamic assessment includes a teaching stage in which the assessor supports the child by teaching them a particular reading skill, which the child is asked to apply. Thus, unlike static assessment, dynamic assessment focuses on the child's learning potential, rather than only on their current ability. Furthermore, because dynamic assessment includes a component of instruction within the test, it is possible to assess a skill not yet acquired by a child, but which is within their ability to master with assistance (i.e., in Vygotskian terms, within their *Zone of Proximal Development*). This means that it is possible to use a dynamic assessment of pseudoword decoding with children at the start of their formal schooling.

Although there has been growing interest in the use of dynamic assessment to screen for future reading difficulty, much of the research conducted to date has taken place in the United States of America and Europe, and with children who have been at school for some time. This current research study sought to investigate whether a dynamic assessment of decoding could be used in the New Zealand context to screen children at the start of their formal schooling for the risk of developing a reading difficulty.

In this chapter, the research findings and related conclusions are presented. This is followed by a discussion of the contribution made by this research, as well as the limitations and recommendations for future research. Finally, the implications of this research for current assessment practices in New Zealand, and in particular regarding the early identification of reading difficulty risk, are explored.

Summary of the research findings

This longitudinal study compared the predictive validity and classification accuracy of a dynamic assessment of decoding with that of commonly used static measures of emergent literacy skills by using a combination of correlational analyses, logistic regression, and predictive classification accuracy analyses. The dynamic and static measures were administered to New Zealand children at the start of their formal schooling, with follow-up assessment after a year at school. The decision to do the follow-up assessment after a year at school was made because it coincides with the time at which children in New Zealand are commonly referred for remedial support. Correlational analyses of the data revealed information regarding the relationship between the measures administered at the start of school, and those administered a year later. The ability of the start-of-school measures to accurately predict reading status after a year at school was examined using the logistic regression and classification accuracy analyses. This study also employed hierarchical logistic regression and classification accuracy analyses to investigate whether a combination of the dynamic and static measures resulted in superior predictive validity and classification accuracy than the dynamic measure on its own.

The results of the study provide support for the use of a dynamic assessment of phonological decoding for the prediction of reading difficulty with children at the start of their formal schooling. The dynamic assessment of decoding provided superior predictive ability and classification accuracy than the static measures and combining the dynamic measure with static measures did not improve the overall ability of the dynamic measure alone to predict future reading difficulty. There is also some indication that the dynamic measure may be able to fulfil other key characteristics of an effective screening tool, namely that a screening tool should provide data to inform and support preventative interventions for those children at risk of a reading difficulty, and that it should be quick and easy to administer.

Research contribution

While in other countries (particularly the United States of America and Scandinavian countries), there have been several studies looking at the use of dynamic assessment as a screening tool for children at-risk of reading difficulty, there is a dearth of such research in New Zealand. Furthermore, at the time of writing, only three other studies, all by Petersen and colleagues (2015, 2016, 2018), focussed specifically on the use of a dynamic assessment of decoding using English-like pseudowords consisting of real letters, administered to children at the start of their formal schooling. The aforementioned research by Petersen and colleagues was conducted in the United States of America with children for whom English and/or Spanish was their main language. The realities of the New Zealand school system and population are very different to those in the United States of America and to date there has not been any research into the use of dynamic assessment to identify risk of reading difficulty within the New Zealand context. This study set out to address this gap and to add to the growing body of research into the use of dynamic assessment as a tool for predicting reading difficulty risk, thereby supporting the implementation of effective interventions for those children who need it most.

Furthermore, although the procedure used in this study for the dynamic assessment was based on that used by Petersen and colleagues (2016), there were several key differences in the dynamic assessment itself and the scales used to inform the DA Strategy and DA Learning scores respectively. Firstly, in an attempt to improve the predictive ability of the assessment, two additional pseudowords were added to the pre-test and teaching stages of the assessment (Petersen and colleagues used four words; the current study used six), and two new pseudowords were added at post-test, as a measure of generalisability (i.e., ability to apply the decoding strategy to pseudowords consisting of the same letters as those the child had been taught, but in a different order). Secondly the Strategy Scale was refined to allow for a score

out of 5 rather than 3. Finally, the Learning Scale was significantly expanded to focus on additional learning behaviours. This was done based on several studies into modifiability and responsiveness to teaching which identified key traits linked to effective and efficient learning (Baggetta & Alexander, 2016; Feuerstein et al., 2010; Guitierrez-Clellen et al., 1998; Lidz, 2003; Mann et al., 2015; Lidz, 1987; Yeager & Yeager, 2013).

The findings of this study did not provide support for the additional pseudowords used, as the results produced were not superior to those achieved by Petersen and colleagues (2016). The changes to the Learning and Strategy Scale also did not appear to improve prediction, however, the use of a more comprehensive Learning Scale did mean that more detailed information was provided on the learning behaviours of the children, including information about the specific areas of difficulty that could benefit from targeted input, as well as about specific strengths the child may have which could be leveraged to support their learning.

Limitations

A limitation of this study was that it was restricted to children in schools in the Auckland North and West regions, and as such, may not be generalisable to the entire New Zealand population of children. In particular, the Auckland North and West regions have a significantly larger proportion of Asian children (whose first languages include Mandarin, Korean, and other Asian languages), than any other area in New Zealand. Furthermore, although the study aimed to include children from across the socio-economic spectrum, it proved difficult to secure the participation of lower decile schools, resulting in the study being somewhat skewed to a higher social-economic demographic.

A further limitation of this study relates to sample size. A total of 165 children participated in the study at T1, with 135 of those children being available for follow-up testing at T2. Only 16 of these children were identified as having a reading difficulty at T2. This represents 12% of the total sample, which is in keeping with most estimates of reading difficulty prevalence. However, small numbers of positive outcomes (in this case reading difficulty) can result in the differences in sensitivity being inflated because a small difference in the proportion of false positives to true positives can have a disproportionately large impact on the sensitivity values.

One of the key advantages attributed to dynamic assessment is its superior ability over that of static assessments to correctly predict reading difficulty in children who belong to a cultural/linguistic minority (Gutiérrez-Clellen & Peña, 2001; Petersen et al., 2018). However, in the current study there were insufficient numbers of children who did not speak English as their main language to investigate this aspect of dynamic assessment. In particular, of the 36 children who spoke a language other than English as their main language, or who were bilingual speakers of English and another language, only one was identified as having a reading difficulty after a year at school. This did not allow for any meaningful analyses of the predictive validity and classification accuracy of the dynamic assessment in terms of children who belong to a linguistic minority in New Zealand.

The duration of the study was an additional limitation. The current study covered a period of only a year, with children tested using the predictor measures at the start of formal schooling, and then a year later, using the outcome measures. A multiyear study investigating the predictive validity and classification accuracy of dynamic and static measures administered to children upon school entry and after two, three and, four years at school, may give a more accurate indication of the actual predictive validity and classification accuracy of these tools in terms of children's ongoing reading development across key skill areas (e.g., reading fluency, accuracy, and comprehension).

Recommendations for future research

The results of this study indicated that the combination of the dynamic assessment of decoding and the static predictor measures does not meaningfully improve the predictive power or classification accuracy of the dynamic assessment alone. Therefore, if improvements to the classification accuracy of the dynamic assessment are desired, this would need to be accomplished through making changes to how the dynamic assessment is administered or scored, or to use different dynamic or static measures. In this regard, below are some suggestions for adaptations to the dynamic assessment which could be considered.

Using different criteria for the discontinuation rule of the dynamic assessment may be one way to improve sensitivity of the tool. Following a procedure similar to that used by Petersen and colleagues (2015, 2016), children who were able to read nine sounds correctly from the first four words did not go on to the teaching phase of the dynamic assessment, but were classified as *not at risk*. As a result, the discontinuation rule was applied to two children who went on to be classified as having a reading difficulty. It is possible that if these two children had gone through the entire process of being taught and then re-tested, the outcome of the test would have been different.

Changes to the teaching phase could also be considered as a possible way to improve the predictive validity and classification accuracy of the dynamic assessment. For example, rather than teaching the individual phonemes within the pseudoword, children could be taught the individual phonemes in isolation (e.g., flash cards with each of the phonemes in isolation), then shown the individual phonemes together (e.g., line up the phonemes to create the pseudoword), and then taught how to blend them. This may help to reduce cognitive load when compared to showing the child the entire pseudoword from the start, as was done in the current study.

Ideally, identification of risk of reading difficulty should take place as early as possible. In the current study, children were assessed for risk of reading difficulty at the start of formal schooling, before formal reading instruction had begun. This was done to overcome the logistical difficulties involved in trying to follow children from their early childhood centres to primary schools. For two key reasons, even this early time in the child's literacy instruction should be considered later than desired. Firstly, to maximise the impact that targeted intervention could have on a child at risk of a reading difficulty, any intervention should take place in the earliest possible stages of the child's reading development when the child is acquiring the emergent literacy skills of phonological awareness and knowledge of phoneme-letter correspondences. This begins before the start of formal schooling, when the child first starts to develop phonological awareness (FPG Child Development Institute, University of North Carolina at Chapel Hill, 2007; Lonigan et al., 2013; Macaruso & Rodman, 2011). Secondly, even at the early stage children were tested in this study, there is some indication that prior learning may have been a confounding factor in the case of some children. For example, during the phonological decoding task, a few children who were clearly very familiar with the letters of the alphabet persisted in providing letter names rather than letter sounds. This likely hindered their ability to learn how to correctly complete the task. In an attempt to address these issues, future research could investigate the use of a dynamic assessment of phonological awareness with younger children (e.g., around three years of age).

Another useful area for further research would be evaluating the other characteristics of a screening tool not specifically investigated in this study. These characteristics include the need for the tool to be quick, easy, and cost-effective to use in the target environment, and to provide information for remediation (not only for reading difficulty risk categorisation). In this regard it would be useful to investigate how easy it would be to train teachers to administer the dynamic assessment, and their feedback on how quick and easy they found it to

implement within their particular contexts. It would also be useful to investigate whether similar or better results could be achieved when the dynamic assessment is administered by teachers, rather than by a researcher. Furthermore, research regarding the usefulness of the information provided by the dynamic assessment in terms of addressing children's specific learning needs and strengths would help to determine whether the dynamic assessment in its current form is providing meaningful information to teachers and schools.

Implications for Dynamic Assessment

Based on comparisons of the continuous measures, there was some indication that including both the dynamic assessment and a letter naming fluency task may reduce the number of false positives when compared to use of the dynamic assessment alone. However, the advantage of improved specificity in a two-stage process needs to be weighed against the additional time and other resources that would be required. Furthermore, a comparison of the dichotomised versions of the dynamic measure and that of the DIBELS Composite score, did not result in an improvement in the overall classification accuracy. Cut points for risk for the DIBELS Next Letter Naming Fluency (LNF) task are not provided, and therefore, further research would be needed to establish whether the combination of the dynamic assessment dichotomous score and a dichotomised version of the DIBELS Next LNF score would result in improved predictive ability. If it did, then inclusion of the letter naming task may well be worthwhile, as the task is very quick and easy to administer. The test itself takes only one minute, and with minimal practice most teachers should be able to complete marking of the test at the same time as it is administered. Furthermore, using a computer spreadsheet containing a relevant formula to automatically indicate risk status using both the dynamic assessment and DIBELS Next LNF task, would mean that the total additional time needed to include the letter naming task in the testing battery would be minimal.

Implications for practice

Numerous studies have shown the importance of providing children at risk of reading difficulty with intensive, evidence-based early intervention (Bishop & League, 2006; Bridges & Catts, 2011; Huang et al., 2014; Lonigan et al., 2013; O'Connor & Jenkins, 1999). As such it could be argued that the current "wait-to-fail at the end of Year 1" approach in New Zealand schools is doing our children a disservice. Instead, screening for reading difficulty risk should take place at the start of formal schooling to allow for early targeted intervention. While most schools complete some form of assessment shortly after a child starts school, generally this is not used to screen for reading difficulties, but instead to get an overall impression of the child's current literacy skills (Education Review Office, 2018). Formal assessment of learners to identify reading difficulties generally only takes place after children have been at school for a year and is frequently in the form of the Observation Survey of Early Literacy Achievement (Clay, 2013). Although teachers carry out ongoing assessment throughout the first year of schooling (often in the form of running records), even if they think a child may have a reading difficulty, most do not attempt to address this during the first year, but instead wait to refer the child for Reading Recovery after a year at school (Education Review Office, 2018).

The tendency to wait until the child has been at school for a year before formally testing and providing intervention, is likely linked to the view that testing at the start of schooling is too early because children and their literacy skills change so much over the first year, and that children need to have the opportunity to "settle in to schooling". Added to this, the lack of specificity of traditional static assessments administered at the start of formal schooling means that they do not reliably predict which children will go on to have a reading difficulty (Caffrey et al., 2008; Kantor et al., 2011; McAlenney & Coyne, 2015). This may lead many schools to believe it is better to wait before formally testing children. This inability of static assessments to accurately

predict future reading status is likely because the results may be negatively impacted by a young child's inability to understand what is required of them in the static assessment, because they lack the knowledge to complete the task itself (Catts et al., 2009), or because static assessments are only able to test emergent literacy skills, rather than measures such as pseudoword reading, which have been shown to more closely relate to later reading outcomes (Petersen & Gillam, 2015). In contrast, a dynamic assessment includes a teaching stage in which the assessor supports the child by teaching them a particular reading skill they are asked to apply. This allows for a more accurate assessment of the child's current and potential literacy skill that is independent of confounding factors such as prior learning. However, overcoming beliefs and current practices for assessment may be a challenge to adoption of dynamic assessment. Other potential barriers would include awareness and training for staff.

To address these challenges, the present study needs to be replicated using a larger sample that is more closely representative of the population of New Zealand as a whole (refer to Limitations section). This would lend stronger support to the use of a dynamic assessment of decoding administered to all children at the start of formal schooling. In addition, advocacy and education for education stakeholders may help clarify why early screening is an important initial step in addressing the issue of reading difficulties in young children.

Factors supporting the adoption of dynamic assessment include the efficiency of time, as the measure appeared to provide a quick and easy way in which to accurately identify children who were at risk. Furthermore, it provided information regarding the specific aspects of the child's responsiveness to reading instruction which could hinder or support their ability to learn to read. Thus, the dynamic assessment can identify children at risk for reading difficulties and provide important information to individualise the type of targeted intervention needed. A further potential advantage of dynamic assessment is that it may help identify other areas for further assessment to gain more in-depth understanding of the child's specific needs, for example, if a

possible difficulty in the area of attention or behavioural issues is identified. While this dynamic assessment of decoding focusses mainly on the skills underlying word decoding (phonological awareness and orthographic learning), it is acknowledged that reading is a complex process and that difficulties in other areas (in particular oral language comprehension skills) can also contribute to difficulties in reading. However, as was discussed in Chapter 2 (Literature Review), during the first years of learning to read, decoding is a better predictor of reading ability than oral language skills (e.g., Vellutino & Fletcher, 2005; National Early Literacy Panel, 2008). So, while the dynamic assessment of decoding may not be able to identify all educational difficulties children may experience, the results of the present study suggest that it is an effective and acceptable way to screen children at the start of schooling for risk of reading difficulties and that it can provide important information for intervention.

While not currently an approach used within New Zealand, adoption of dynamic assessment of decoding could be utilised as part of a wider response to intervention (RTI) approach. In this type of model, the dynamic assessment could be used as the universal screening tool to determine reading difficulty risk at the start of school, and iterative progress monitoring thereafter would provide data about all learner's responsiveness to instruction and/or intervention. One of the current criticisms levelled against an RTI approach used in other countries is the length of time involved in identifying reading difficulty. This is because RTI relies on periodic screening and progress monitoring over several weeks to determine learner's responsiveness to classroom instruction. In contrast, using a dynamic assessment as a universal screening tool at the start of formal schooling would allow the immediate identification of children at risk of reading difficulty so that intervention can be implemented without delay.

While further and larger scale research about dynamic assessment needs to be done, preventing or ameliorating reading difficulties in young children should be a top priority for all schools in New Zealand. The negative and associated effects of reading difficulty, including the adverse impacts on academic achievement, employment prospects, and the child's psychological and emotional well-being (Burden, 2008; Lyon, 2003; McCardle et al., 2001; Torgesen, 2002), are too great to wait to take measures to address this issue. Ideally each child's teacher will assess them upon or soon after entry to school, using the dynamic assessment of decoding. This will allow the teacher to observe, first hand, the learning behaviours exhibited by the child and identify the child's relative strengths and weakness in relation to responding to instruction. To simplify and expedite the calculation of a final dynamic assessment dichotomous score, a simple spreadsheet containing a formula could be used to record the child's different scores on the dynamic assessment.

The information provided by the dynamic assessment can then be used to provide targeted intervention for those children who need it. Depending on the specific aspect of responsiveness to reading instruction identified as being weak for the child, this could take the form of individual intensive intervention targeted at developing phonological awareness and phoneme-grapheme mapping skills, and/or strategies implemented to support the child's sensory integration, thinking skills, task engagement, and so on; both in an individual or small group context, and within the full class environment as well. Better understanding of children's risk of reading difficulty and learning skills and strategies would be valuable information for new entrant teachers to help support children's best start at school.

Concluding statement

An assessment able to predict future reading difficulty with perfect sensitivity and specificity, and which is practical and cost-effective to implement within the New Zealand school system is unlikely to ever exist. However, because it is so important to be able to accurately identify those at risk, and to intervene as early as possible to potentially stave off or at least ameliorate reading difficulties, research into the area of assessment for reading difficulty screening in vital. The deleterious consequences of a reading difficulty are wide-ranging and start to affect children very early in their schooling. Therefore, action needs to be taken as early as possible to accurately identify children at risk and provide information that can be used to support these children as soon as possible. It is hoped that this research has added to the existing body of research into the early identification of reading difficulty risk by revealing how a dynamic assessment of decoding may be used at the start of formal schooling to accurately predict future reading status. There is an indication that a dynamic assessment of decoding can accurately identify those children who will go on to have a reading difficulty after a year at school, and those who will not; allowing schools to confidently make decisions regarding the children who need support immediately and take action during their first year of schooling.

References

Aaron, P. G., Joshi, R. M., & Quatroche, D. J. (2008). *Becoming a professional reading teacher.* Baltimore, MD: Paul H. Brookes Publications.

Adolf, S. M., Catts, H. W., & Lee, J. (2010). Kindergarten predictors of second versus eighth grade reading comprehension impairments. *Journal of Learning Disabilities, 43*(4), 332-345. https://doi.org/10.1177/0022219410369067

Al Otaiba, S., Connor, C. M., Folsom, J. S., Wansek, J., Greulich, L., Schatschneider, C., & Wagner, R. K. (2014). To wait in Tier 1 or intervene immediately: A randomized experiment examining first-grade response to intervention in reading. *Exceptional Children, 18*(1), 11-27. https://doi.org/10.1177/0014402914532234

Allison, P. D. (2014). *Measures of fit for logistic regression*. Retrieved from https://statisticalhorizons.com/wp-content/uploads/GOFForLogisticRegression-Paper.pdf

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science, 14*(5), 255-259. https://doi.org/10.1111/j.0963-7214.2005.00376.x

Arnell, K. M., Klein, R. M., Joanisse, M. F., Busseri, M. A., & Tannock, R. (2009). Decomposing the relation between Rapid Automatized Naming (RAN) and reading ability. *Canadian Journal of Experimental Psychology, 63*(3), 173-184. https://doi.org/10.1037/a0015721

Baggetta, P., & Alexander, P. A. (2016). Conceptualization and operationalization of executive function. *Mind, Brain, and Education, 10*(1), 10-33. https://doi.org/10.1111/mbe.12100

Bailet, L. L., Repper, K., Murphy, S., Piasta, S., & Zettler-Greeley, C. (2011). Emergent literacy intervention for prekindergarteners at risk for reading failure: Years 2 and 3 of a multiyear study. *Journal of Learning Disabilities, 46*(2), 133-153. https://doi.org/10.1177/0022219411407925

Bishop, A. G., & League, M. B. (2006, Fall). Identifying a multivariate screening model to predict reading difficulties at the onset of kindergarten: A longitudinal analysis. *Learning Disability Quarterly, 29*(4), 235-252. https://doi.org/10.2307/30035552

Boan, C. H., Aydlett, L., & Multunas, N. (2007). Early childhood screening and readiness assessment. In B. A. Bracken, & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 49-68). Mahwah, NJ: Lawrence Erlbaum Associates.

Bodrova, E., & Leong, D. J. (2007). *Tools of the mind: The Vygotskian approach to early childhood education* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Boets, B., Vandermosten, M., Poelmans, H., Luts, H., Wouters, J., & Ghesquière, P. (2011). Preschool impairments in auditory processing and speech perception uniquely predict future reading problems. *Research in Development Disabilities, 32*(2), 560-570. https://doi.org/10.1016/j.ridd.2010.12.020

Boscardin, C. K., Muthén, B., Francis, D. J., & Baker, E. L. (2008). Early identification of reading difficulties using heterogeneous developmental trajectories. *Journal of Educational Psychology, 100*(1), 192-208. https://doi.org/10.1037/0022-0663.100.1.192

Bosse, M.l., Chaves, N., Largy, P., & Valdois, S. (2015). Orthographic learning during reading: The role of whole-word visual processing. *Journal of Research in Reading, 38*(2), 141-158. https://doi.org/10.1111/j.1467-9817.2012.01551.x

Bowey, J. A., & Muller, D. (2005). Phonological recoding and rapid orthographic learning in third-graders' silent reading: A critical test of the self-teaching hypothesis. *Journal of Experimental Child Psychology, 92*(3), 203-219. https://doi.org/10.1016/j.jecp.2005.06.005

Bridges, M. S., & Catts, H. W. (2011). The use of dynamic screening of phonological awareness to predict risk for reading disabilities in kindergarten children. *Journal of Learning Disabilities, 44*(4), 330-338. https://doi.org/10.1177/0022219411407863

Brooks, P. L., & Weeks, S. A. (1998). A comparison of the responses of dyslexic, slow learning and control children to different strategies for teaching spellings. *Dyslexia, 4*(4), 212-222. https://doi.org/10.1002/(SICI)1099-0909(1998120)4:4<212::AID-DYS120>3.0.CO;2-Q

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Budoff, M. (1987). Measures for assessing learning potential. In C. S. Lidz (Ed.), *Dynamic Assessment: An interactional approach to evaluating learning potential* (pp. 53-81). New York, NY: Guilford Press.

Burden, R. (2008, August 1). Is dyslexia necessarily associated with negative feelings of self-worth? A review and implications for future research. *Dyslexia, 14*(3), 188-196. https://doi.org/10.1002/dys.371

Burke, M. D., Crowder, W., Hagan-Burke, S., & Zou, Y. (2009). A comparison of two path models for predicting reading fluency. *Remedial and Special Education, 30*(2), 84-95. https://doi.org/10.1177/0741932508315047

Butterworth, B., & Kovas, Y. (2013, April 19). Understanding neurocognitive developmental disorders can improve education for all. *Science, 340*(6130), 300-305. https://doi.org/10.1126/science.1231022

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008, Winter). The predictive validity of dynamic assessment: A review. *The Journal of Special Education, 41*(4), 254-270. https://doi.org/10.1177/0022466907310366

Camilleri, B., & Botting, N. (2013, Sept/Oct). Beyond static assessment of children's receptive vocabulary: The dynamic assessment of word learning (DAWL). *International Journal of Language & Communication Disorders, 48*(5), 565-581. https://doi.org/10.1111/1460-6984.12033

Campione, J. C. (1989, March). Assisted assessment: A taxonomy of approaches and an outline of strengths and weaknesses. *Journal of Learning Disabilities, 22*(3), 151-165. https://doi.org/10.1177/002221948902200303

Campione, J. C., & Brown, A. L. (1990). Guided learning and transfer: Implications for approaches to assessment. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto, *Diagnostic monitoring of skill and knowledge acquisition* (pp. 141-172). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carlson, J. S., & Wiedl, K. H. (1979, Oct-Dec). Toward a differential testing approach: Testing-the-limits employing the Raven matrices. *Intelligence, 3*(4), 323-344. https://doi.org/10.1016/0160-2896(79)90002-3

Carran, D. T., & Scott, K. G. (1992, June). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education, 12*(2), 196-211. https://doi.org/10.1177/027112149201200205

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006, April). Language deficits in poor comprehenders: A case of the simple view of reading. *Journal of Speech, Language, and Hearing Research, 49*(2), 278-293. https://doi.org/10.1044/1092-4388(2006/023)

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, B. J. (2001, January). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech & Hearing Services in Schools, 32*(1), 38-50. https://doi.org/10.1044/0161-1461(2001/004)

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *Journal of Learning Disabilities, 48*(3), 281-297. https://doi.org/10.1177/0022219413498115

Catts, H. W., Nielsen, D. C., Bridges, S. M., & Liu, Y.S. (2014). Early identification of reading comprehension difficulties. *Journal of Learning Disabilities*, 49(5), 1-15. https://doi.org/10.1177/0022219414556121

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009, Mar-Apr). Floor effects associated with universal screening and their impact on early identification of reading disabilities. *Journal of Learning Disabilities, 42*(2), 163-176. https://doi.org/10.1177/0022219408326219

Chamberlain, M. (July 2013). *PIRLS 2010/11 in New Zealand: An overview of findings from the third cycle of the Progress in International Reading Literacy Study (PIRLS).* Wellington, New Zealand: New Zealand Ministry of Education. Retrieved from http://www.educationcounts.govt.nz/publications/series/2539/114981/125051

Cho, E., & Compton, D. L. (2015). Construct and incremental validity of dynamic assessment of decoding within and across domains. *Learning and Individual Differences, 37*, 183-196. https://doi.org/10.1016/j.lindif.2014.10.004

Cho, E., Compton, D. L., Fuchs, D., Fuchs, L., & Bouton, B. (2014). Examining the predictive validity of a dynamic assessment of decoding to forecast response to Tier 2 intervention. *Journal of Learning Disabilities, 47*(5), 409-423. https://doi.org/10.1177/0022219412466703

Clay, M. M. (2013). *An Observation Survey of Early Literacy Achievement* (3rd ed.). Portsmouth, NH: Heinemann.

Clemens, N., Lai, M. H., Burke, M., & Wu, J.Y. (2017). Interrelations of growth in letter naming and sound fluency in kindergarten and implications for subsequent reading fluency. *School Psychology Review, 46*(3), 272-287. https://doi.org/10.17105/SPR-2017-0032.V46-3

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. (2006, June). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*(2), 394-409. https://doi.org/10.1037/0022-0663.98.2.394

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., . . . Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false

positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327-340. https://doi.org/10.1037/a0018448

Compton, D. L., Gilbert, J. K., Jenkins, J. R., Fuchs, D., Fuchs, L. S., Cho, E., . . . Bouton, B. (2012). Accelerating chronically unresponsive children to Tier 3 instruction: What level of data is necessary to ensure selection accuracy? *Journal of Learning Disabilities, 45*(3), 204-216. https://doi.org/10.1177/0022219412442151

Conners, F. A., Loveall, S. J., Moore, M. S., Hume, L. E., & Maddox, C. D. (2011). An individual differences analysis of the self-teaching hypothesis. *Journal of Experimental Child Psychology, 108*(2), 402-410. https://doi.org/10.1016/j.jecp.2010.09.009

Corriveau, K. H., Goswami, U., & Thomson, J. M. (2010). Auditory processing and early literacy skills in preschool and kindergarten population. *Journal of Learning Disabilities, 43*(4), 369-382. https://doi.org/10.1177/0022219410369071

Cotton, S., & Crewther, S. (2012). Developmental dyslexia: A conceptual and measurement quandary. In M. L. Falese (Ed.), *Encyclopedia of education research* (pp. 119-153). New York, NY: Nova Science Publishers

Coventry, W. L., Byrne, B., Olson, R. K., Corley, R., & Samuelsson, S. (2011). Dynamic and static assessment of phonological awareness in preschool: A behavior-genetic study. *Journal of Learning Disabilities, 44*(4), 322-329. https://doi.org/10.1177/0022219411407862

Cronin, V. S. (2011). RAN and Double-Deficit Theory. *Journal of Learning Disabilities, 46*(2), 182-190. https://doi.org/10.1177/0022219411413544

de Beer, M. (2010). A modern assessment psychometric approach to dynamic assessment. *Journal of Psychology in Africa, 20*(2), 241-246. https://doi.org/10.1080/14330237.2010.10820372

de Jong, P. F. (2011). What discrete and serial rapid automatized naming can reveal about reading. *Scientific Studies of Reading, 15*(4), 314-337. https://doi.org/10.1080/10888438.2010.485624

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988, Sept). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837-845. https://doi.org/10.2307/2531595

Dewey, E. N., Powell-Smith, K. A., Good, R. H., & Kaminski, R. A. (2015). *DIBELS Next technical adequacy brief.* Eugene, OR: Dynamic Measurement Group. Retrieved from https://dibels.org/?ourl=/papers/DIBELSNextTechnicalAdequacy.pdf

Dörfler, T., Golke, S., & Artelt, C. (2009). Dynamic assessment and its potential for the assessment of reading competence. *Studies in Educational Evaluation, 35*(2), 77-82. https://doi.org/10.1016/j.stueduc.2009.10.005

Dufva, M., Niemi, P., & Voeten, M. J. (2001). The role of phonological memory, word recognition, and comprehension skills in reading development: From preschool to grade 2. *Reading and Writing: An Interdisciplinary Journal, 14*(1), 91-117. https://doi.org/10.1023/A:1008186801932

Dynamic Measurement Group Inc. (2010, December 1). *DIBELS Next benchmark goals and composite score.* Retrieved from https://dibels.uoregon.edu/docs/DIBELSNextFormerBenchmarkGoals.pdf

Dyslexia Foundation of New Zealand. (2015). *Dyslexia Foundation of New Zealand - About.* Retrieved from http://www.dyslexiafoundation.org.nz/about.html

Edele, A., & Stanat, P. (2015, July 20). The role for first-language listening comprehension in second-language reading comprehension. *Journal of Educational Psychology*, 108(2), 1-18. https://doi.org/10.1037/edu0000060

Education Review Office. (2018). *Evaluation at a glance: A decade of assessment in New Zealand primary schools - practice and trends.* Wellington, New Zealand: Education Review Office.

Ehri, L. C. (2005a). Development of sight word reading: Phases and findings. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook.* Malden, MA: Blackwell Publishing.

Ehri, L. C. (2005b). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9*(2), 167-188. https://doi.org/10.1207/s1532799xssr0902_4

Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading, 18*(1), 5-21. https://doi.org/10.1080/10888438.2013.819356

Ehri, L. C., & Saltmarsh, J. (1995). Beginning readers outperform older disabled readers in learning to read words by sight. *Reading and Writing: An Interdisciplinary Journal, 7*(3), 295-326. https://doi.org/10.1007/BF03162082

Elbro, C., Daugaard, H. T., & Gellert, A. S. (2012). Dyslexia in a second language? A dynamic test of reading acquisition may provide a fair answer. *Annuals of Dyslexia, 62*(3), 172-185. https://doi.org/10.1007/s11881-012-0071-7

Feuerstein, R., Feuerstein, R. S., & Falik, L. H. (2010). *Beyond smarter: Mediated learning and the brain's capacity for change.* New York, NY: Teachers College Press.

Field, A. (2015). *Discovering statistics using IBM SPSS statistics.* London, UK: SAGE Publications.

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for differrent types of alphabetic orthographies? *Educational Psychology Review, 23*(4), 553-576. https://doi.org/10.1007/s10648-011-9175-6

FPG Child Development Institute, University of North Carolina at Chapel Hill. (2007, Spring). *RTI Goes to Pre-K: An early intervening system called Recognition & Response.* Retrieved from https://fpg.unc.edu/sites/fpg.unc.edu/files/resources/early-developments/FPG_EarlyDevelopments_v11n1.pdf

Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*(1), 93-99. https://doi.org/10.1598/RRQ.41.1.4

Fuchs, D., Compton, D. L., Fuchs, L. S., Bouton, B., & Caffrey, E. (2011). The construct and predictive validity of dynamic assessment of young children learning to read: Implications for RTI frameworks. *Journal of Learning Disabilities, 44*(4), 339-347. https://doi.org/10.1177/0022219411407864

Fuchs, D., Fuchs, L. S., Compton, D. L., Bouton, B., Caffrey, E., & Hill, L. (2007, May/June). Dynamic assessment as responsiveness to intervention: A scripted protocol to identify young at-risk readers. *Teaching Exceptional Children*, *39*(5), 58-63. https://doi.org/10.1177/004005990703900508

Furnes, B., & Samuelsson, S. (2011). Phonological awareness and rapid automatized naming predict early development in reading and spelling: Results for a cross-linguistic longitudinal study. *Learning and Individual Differences, 21*(1), 85-95. https://doi.org/10.1016/j.lindif.2010.10.005

Gabrieli, J. D. (2009, July 17). Dyslexia: a new synergy between education and cognitive neuroscience. *Science, 325*(5938), 280-283. https://doi.org/10.1126/science.1171999

Gallagher, E. J. (2003, August). The problem with sensitivity and specificity ... *Annals of Emergency Medicine*, *42*(2), 298-303. https://doi.org/10.1067/mem.2003.273

Gellert, A. A., & Elbro, C. (2018). Predicting reading disablities using dynamic assessment of decoding before and after the onset of reading instruction: A longitudinal study from kindergarten through grade 2. *Annuals of Dyslexia, 68*(2), 126-144. https://doi.org/10.1007/s11881-018-0159-9

Good, R. H., & Kaminski, R. A. (2011). *DIBELS Next assessment manual.* Retrieved from http://www.dibels.org/

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A., & Latimer, R. J. (2013). *DIBELS Next technical manual.* Retrieved from http://www.dibels.org

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6-10. https://doi.org/10.1177/074193258600700104

Grigorenko, E. L. (2009, Mar/Apr). Dynamic Assessment and Response to Intervention: Two sides of one coin. *Journal of Learning Disabilities, 42*(2), 111-132. https://doi.org/10.1177/0022219408326207

Gustafson, S., Svensson, I., & Fälth, L. (2014). Response to Intervention and Dynamic Assessment: Implementing systematic, dynamic and individualised interventions in primary school. *International Journal of Disability, Development and Education, 61*(1), 27-43. https://doi.org/10.1080/1034912X.2014.878538

Gutierrez-Clellen, V. F., Brown, S., Conboy, B., & Robinson-Zañartu, C. (1998). Modifiability: A dynamic approach to assessment immediate language change. *Journal of Children's Communication Development, 19*(2), 31-42. https://doi.org/10.1177/152574019801900204

Gutiérrez-Clellen, V. F., & Peña, E. (2001, October). Dynamic assessment of diverse children: A tutorial. *Language, Speech, and Hearing Services in Schools, 32*(4), 212-224. https://doi.org/10.1044/0161-1461(2001/019)

Guthke, J., & Stein, H. (1996). Are learning tests the better version of intelligence tests? *European Journal of Psychological Assessment, 12*, 1-13. https://doi.org/10.1027/1015-5759.12.1.1

Herrmann, J. A., Matyas, T., & Pratt, C. (2006). Meta-analysis of the nonword reading deficit in specific reading disorder. *Dyslexia, 12*(3), 195-221. https://doi.org/10.1002/dys.324

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*(2), 127-160. https://doi.org/10.1007/BF00401799

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Hoti, A. U., Heinzmann, S., Müller, M., Oliveira, M., Wicki, W., & Werlen, E. (2011). Introducing a second foreign language in Swiss primary schools: The effect of L2 listening and reading skills on L3 acquisition. *International Journal of Multilingualism, 8*(2), 98-116. https://doi.org/10.1080/14790718.2010.527006

Hu, B., Shao, J., & Palta, M. (2006). Pseudo-R^2 in logistic regression model. *Statistica Sinica*, 847-860. Retrieved from http://www3.stat.sinica.edu.tw/statistica/oldpdf/a16n39.pdf

Huang, F. L., Moon, T. R., & Boren, R. (2014). Are the reading rich getting richer? Testing for the prescence of the Matthew Effect. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 30*(2), 95-115. https://doi.org/10.1080/10573569.2013.789784

Hulme, C., & Snowling, M. (2009). *Developmental disorders of language learning and cognition.* Hoboken, NJ: Wiley-Blackwell.

Hulme, C., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., & Snowling, M. (2012). *York Assessment of Reading for Comprehension. Early Reading. Australian Edition Manual.* London: GL Assessment.

IBM. (n.d.). *Pseudo R-squared measures*. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/tutorials/plum_germcr_rsquare.html

Jeltova, I., Birney, D., Fredine, N., Jarvin, L., Sternberg, R. J., & Grigorenko, E. L. (2011). Making instruction and assessment responsive to diverse students' progress: Group-administered dynamic assessment in teaching mathematics. *Journal of Learning Disabilities, 44*(4), 381-395. https://doi.org/10.1177/0022219411407868

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a Response to Intervention framework. *School Psychology Review, 36*(4), 582-600. Retrieved from http://eds.a.ebscohost.com.ezproxy.massey.ac.nz/eds/pdfviewer/pdfviewer?vid=3&sid=fd231650-def1-4fae-8e0c-1db84b793a2d%40sessionmgr4006

Jones, M. W., Branigan, H. P., & Kelly, M. L. (2009). Dyslexia and nondyslexic reading fluency: Rapid automatized naming and the importance of continuous lists. *Psychonomic Bulletin and Review, 16*(3), 567-572. https://doi.org/10.3758/PBR.16.3.567

Joshi, M. R., & Aaron, P. G. (2012). Componential model of reading (CMR): Validation studies. *Journal of Learning Disabilities, 45*(5), 387-390. https://doi.org/10.1177/0022219411431240

Kaminski, R. A., & Powell-Smith, K. A. (2017). Early literacy intervention for preschoolers who need Tier 3 support. *Topics in Early Childhood Special Education, 36*(4), 205-217. https://doi.org/10.1177/0271121416642454

Kantor, P. T., Wagner, R. K., Torgesen, J. K., & Rashotte, C. (2011). Comparing two forms of dynamic assessment and traditional assessment of preschool phonological awareness. *Journal of Learning Disabilities, 44*(4), 313-321. https://doi.org/10.1177/0022219411407861

Kendeou, P., Savage, R., & van den Broek, P. (2009). Revisiting the simple view of reading. *British Journal of Educational Psychology, 79*(2), 353-370. https://doi.org/10.1348/978185408X369020

Koda, K. (2007, June). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning, 57*(1), 1-44. https://doi.org/10.1111/0023-8333.101997010-i1

Lane, K. L., Menzies, H. M., Oakes, W. P., & Kalberg, J. R. (2012). *Systematic screenings of behavior to support instruction.* New York, NY: The Guilford Press.

Lervåg, A., & Hulme, C. (2014). Rapid Automatized Naming (RAN) taps a mechanism that places constraints on the development of early reading fluency. *Psychological Science, 20*(8), 1040-1048. https://doi.org/10.1111/j.1467-9280.2009.02405.x

Lidz, C.S. (1991) *Practitioner's guide to dynamic assessment.* New York, NY: The Guilford Press.

Lidz, C. S. (2003). *Early childhood assessment.* Hoboken, NJ: John Wiley & Sons.

Lidz, C. S., & Gindis, B. (2003). Dynamic assessment of the evolving cognitive functions in children. In A. Kozulin, *Vygotsky's educational theory in cultural context* (pp. 99-116). Cambridge, United Kingdom: Cambridge University Press.

Lidz, C. S., & Haywood, H. C. (2014). From dynamic assessment to intervention: Can we get there from here? *Transylvanian Journal of Psychology*, 81-108. Retrieved from http://eds.a.ebscohost.com.ezproxy.massey.ac.nz/eds/pdfviewer/pdfviewer?vid=6&sid=fd231650-def1-4fae-8e0c-1db84b793a2d%40sessionmgr4006

Lidz, C. S., & Peña, E. D. (2009). Response to intervention and dynamic assessment: Do we just appear to be speaking the same language? *Seminars in Speech and Language, 30*(2), 121-133. https://doi.org/10.1055/s-0029-1215719

Lonigan, C. J., Purpura, D. J., Wilson, S. B., Walker, P. M., & Clancy-Menchetti, J. (2013). Evaluating the components of an emergent literacy intervention for preschool children at risk of reading difficulties. *Journal of Experimental Child Psychology, 114*(1), 111-130. https://doi.org/10.1016/j.jecp.2012.08.010

Loveall, S. J., & Conner, F. A. (2013). Individuals with intellectual disability can self-teach in reading. *American Journal on Intellectual and Developmental Disabilities, 118*(2), 108-123. https://doi.org/10.1352/1944-7558-118.2.108

Lundberg, I., Frost, J., & Petersen, O.P. (1988, Summer). Effects of an extensive program for stimulating phonological awareness in preschool. *Reading Research Quarterly, 23*(3), 263-284. Retrieved from http://www.jstor.org.ezproxy.massey.ac.nz/stable/748042

Lyon, G. R. (2003). Reading disabilities: Why do some children have difficulty learning to read? What can be done about it? *Perspectives, 29*(2). Retrieved from www.interdys.org

Macaruso, P., & Rodman, A. (2011). Efficacy of computer-assisted instruction for the development of early literacy skills in young children. *Reading Psychology, 32*(2), 172-196. https://doi.org/10.1080/02702711003608071

Manis, F. R., Seidenberg, M. S., & Doi, L. M. (1999). See Dick RAN: Rapid naming and the longitudinal prediction of reading subskills in first and second graders. *Scientific Studies of Reading, 3*(2), 129-157. https://doi.org/10.1207/s1532799xssr0302_3

Mann, W., Peña, E. D., & Morgan, G. (2015, August). Child modifiability as a predictor of language abilities in deaf children who use American Sign Language. *American Journal of Speech-Language Pathology, 24*(3), 374-385. https://doi.org/10.1044/2015_AJSLP-14-0072

Massey University, New Zealand. (2010). *Code of ethical conduct for research, teaching and evaluations involving human participants.* Palmerston North, New Zealand: Massey University.

Mather, N., Wending, B. J., & Kaufman, A. S. (2011). *Essentials of dyslexia assessment and intervention.* Hoboken, NJ: John Wiley & Sons.

McAlenney, A. L., & Coyne, M. D. (2015). Addressing false positives in early reading assessment using intervention response data. *Learning Disability Quarterly, 38*(1), 53-65. https://doi.org/10.1177/0731948713514057

McCardle, P., Scarborough, H. S., & Catts, H. W. (2001). Predicting, explaining and preventing children's reading difficulties. *Learning Disabilities Research and Practice, 16*(4), 230-239. https://doi.org/10.1111/0938-8982.00023

Mislevy, J. L., & Rupp, A. A. (2010). Predictive validity. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1076-1078). Thousand Oaks, CA: SAGE Publications.

Moll, K., Loff, A., & Snowling, M. J. (2013). Cognitive endophenotypes of dyslexia. *Scientific Studies of Reading, 17*(6), 385-397. https://doi.org/10.1080/10888438.2012.736439

Morlini, I., Stella, G., & Scorza, M. (2014). A new procedure to measure children's reading speed and accuracy in Italian. *Dyslexia, 20*(1), 54-73. https://doi.org/10.1002/dys.1462

Moyle, M. J., Heilmann, J., & Berman, S. S. (2013). Assessment of early developing phonological awareness skills: A comparison of the Preschool Individual Growth and Development Indicators and the Phonological Awareness and Literacy Screening - PreK. *Early Education and Development, 24*, 668-686. https://doi.org/10.1080/10409289.2012.725620

Nation, K., Angell, P., & Castles, A. (2007). Orthographic learning via self-teaching in children learning to read English: Effects of exposure, durability, and context. *Journal of Experimental Child Psychology, 96*(1), 71-84. https://doi.org/10.1016/j.jecp.2006.06.004

National Centre for Research Methods. (2011, July 22). *Using statistical regression methods in educational research*. Retrieved from http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod4/9/index.html

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel. A scientific synthesis of early literacy development and implications for intervention.* National Institute for Literacy. Retrieved from http://lincs.ed.gov/publications/pdf/NELPReport09.pdf

New Zealand Health Technology Assessment Clearing House. (1998, October). *Colour vision screening: A critical appraisal of the literature.* Retrieved from http://nzhta.chmeds.ac.nz/publications/nzhta7.pdf

New Zealand Ministry of Education. (2007). *Literature review: An international perspective on dyslexia.* Wellington, New Zealand: Crown Publishing.

New Zealand Ministry of Education. (2008). *About dyslexia: Teacher resource.* Wellington, New Zealand: Ministry of Education.

New Zealand Ministry of Education. (2015). *Implementing an inclusive curriculum.* Retrieved from https://nzcurriculum.tki.org.nz/content/download/161528/1194589/file/Implementing%20an%20Inclusive%20Curriculum_2017.pdf

New Zealand Ministry of Education. (2015, October 8). *School deciles*. Retrieved from http://www.education.govt.nz/school/running-a-school/resourcing/operational-funding/school-decile-ratings/

New Zealand Ministry of Education. (2016, July). *Ambitious for New Zealand - The Ministry of Education Four Year Plan 2016-2020.* Retrieved from https://www.education.govt.nz/assets/Uploads/4-Year-Plan-2016-WEB.pdf

New Zealand Ministry of Education. (2018). *Transient students.* Retrieved from https://www.educationcounts.govt.nz/statistics/indicators/main/student-engagement-participation/transient-students

New Zealand Ministry of Education. (n.d. -a). *Schools rolls*. Retrieved from https://www.educationcounts.govt.nz/statistics/schooling/student-numbers/6028

New Zealand Ministry of Education. (n.d. -b). *Year by year progress and achievement in reading against National Standards, the colour wheel and curriculum levels*. Retrieved from http://assessment.tki.org.nz/Reporting-to-parents-whanau/Examples-and-templates/Illustrating-progress-in-foundations-for-learning/Year-by-year-reading-against-NS-colour-wheel-and-curr-levels

Nijakowska, J. (2010). *Second language acquisition: Dyslexia in the foreign language classroom.* Bristol, United Kingdom: Multilingual Matters.

Norton, E. S., & Wolf, M. (2012). Rapid Automatized Naming (RAN) and reading fluency: Implications for understanding and treatment of reading disabilities. *The Annual Review of Psychology, 63*(1), 427-252. https://doi.org/10.1146/annurev-psych-120710-100431

O'Connor, R. E., & Jenkins, J. R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3*(2), 159-197. https://doi.org/10.1207/s1532799xssr0302_4

O'Donnell, P. S., & Miller, D. N. (2011). Identifying students with specific learning disabilities: School psychologists' acceptability of the discrepancy model versus Response to Intervention. *Journal of Disability Policy Studies, 22*(2), 83-97. https://doi.org/10.1177/1044207310395724

Ortiz, M., Folsom, J. S., Al Otaiba, S., Greulich, L., Thomas-Tate, S., & Connor, C. M. (2012). The componential model of reading: Predicting first grade reading performance of culturally diverse students from ecological, psychological, and cognitive factors assessed at kindergarten entry. *Journal of Learning Disabilities, 45*(5), 406-417. https://doi.org/10.1177/0022219411431242

Ouellette, G., & Fraser, J. R. (2009). What exactly is a *yait* anyway: The role of semantics in orthographic learning. *Journal of Experimental Psychology, 104*(2), 239-251. https://doi.org/10.1016/j.jecp.2009.05.001

Petersen, D. B., & Gillam, R. B. (2015). Predicting reading ability for bilingual Latino children using dynamic assessment. *Journal of Learning Disabilities, 48*(1), 3-21. https://doi.org/10.1177/0022219413486930

Petersen, D. B., Allen, M. M., & Spencer, T. D. (2016). Predicting reading difficulty in first grade using dynamic assessment of decoding in early kindergarten: A large-scale longitudinal study. *Journal of Learning Disabilities*, *49*(2), 1-16. https://doi.org/10.1177/0022219414538518

Petersen, D. B., Gragg, S. L., & Spencer, T. D. (2018, October). Predicting reading problems 6 years into the future: Dynamic Assessment reduces bias and increases classification accuracy. *Language, Speech, and Hearing Services in Schools, 49*(4), 875-888. https://doi.org/10.1044/2018_LSHSS-DYSLC-18-0021

Pett, M. A. (2016). *Nonparametric statistics for health care research: Statistics for small sample and unusual distributions.* Thousand Oaks, CA: Sage Publishing.

Plante, E., & Vance, R. (1994, January). Selection of preschool language tests: A data based approach. *Language, Speech, and Hearing Services in Schools, 25*(1), 15-24. https://doi.org/10.1044/0161-1461.2501.15

Poehner, M. E. (2008). *Dynamic Assessment: A Vygotskian approach to understanding and promoting L2 development.* Berlin, Germany: Springer Science+Business Media.

Proctor, E., Silmere, H., Raghaven, R., Hovmand, P., Aarons, G., Bunger, A., . . . Hensley, M. (2011, March). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(2), 65-76. https://doi.org/10.1007/s10488-010-0319-7

Rack, J. P., Snowling, M. J., & Olson, R. K. (1992, Winter). The nonword reading deficit in developmental dyslexia: a review. *Reading Research Quarterly, 27*(1), 28-53. https://doi.org/10.2307/747832

Ricketts, J., Bishop, D. V., Nation, K., & Pimperton, H. (2011). The role of self-teaching in learning orthographic and semantic aspects of new words. *Scientific Studies of Reading, 15*(1), 47-70. https://doi.org/10.1080/10888438.2011.536129

Roberts, G., Mohammed, S. S., & Vaugh, S. (2010). Reading achievement across three language groups: Growth estimates for overall reading and reading subskills obtained with the early childhood longitudinal survey. *Journal of Educational Psychology, 102*(3), 668-686. https://doi.org/10.1037/a0018983

Roberts, J. A., & Scott, K. A. (2006). The simple view of reading: Assessment and intervention. *Top Language Disorders, 26*(2), 127-143. https://doi.org/10.1097/00011363-200604000-00005

Robinson-Zañartu, C., & Carlson, J. (2013). Dynamic assessment. In K. F. Geisinger (Ed.), *APA handbooks in psychology. APA handbook of testing and assessment in psychology, Vol. 3. Testing and assessment in school psychology and education* (pp. 149-367). American Psychological Association. https://doi.org/10.1037/14049-007

Safa, M. A., & Beheshti, S. (2018). Interactionist and interventionist Group Dynamic Assessment (GDA) and EFL learners' listening comprehension development. *Iranian Journal of Language Teaching Research, 6*(3), 37-56. Retrieved from https://pdfs.semanticscholar.org/421b/219a97fcb623deb7572bd28b6800352732ba.pdf?_ga=2.81327703.1557566774.1565059468-488870835.1565059468

Scanlon, D. M., Vellutino, F. R., Small, S. G., Fanuele, D. P., & Sweeney, J. M. (2005). Severe reading difficulties - can they be prevented? A comparison of prevention and intervention approaches. *Exceptionality, 13*(4), 209-227. https://doi.org/10.1207/s15327035ex1304_3

Schatschneider, C., Francis, D. J., Carlson, C. D., Fletcher, J. M., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology, 96*(2), 265-282. https://doi.org/10.1037/0022-0663.96.2.265

Seidenberg, M. S. (2007). Connectionist models of reading. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 235-350). Oxford, United Kingdom: Oxford University Press.

Share, D. L. (1995). Phonological recoding and self-teaching: *Sine qua non* of reading acquisition. *Cognition, 55*(2), 151-218. https://doi.org/10.1016/0010-0277(94)00645-2

Share, D. L. (1999). Phonological recoding and orthographic learning: A direct test of the Self-Teaching Hypothesis. *Journal of Experimental Child Psychology, 72*(2), 95-129. https://doi.org/10.1006/jecp.1998.2481

Share, D. L. (2004). Orthographic learning at a glance: On the time course and developmental onset of self-teaching. *Journal of Experimental Child Psychology, 87*(4), 267-298. https://doi.org/10.1016/j.jecp.2004.01.001

Share, D. L., & Shalev, C. (2004). Self-teaching in normal and disabled readers. *Reading and Writing: An Interdisciplinary Journal, 17*(7), 769-800. https://doi.org/10.1007/s11145-004-2658-9

Share, D. L., & Stanovich, K. E. (1995). Cognitive processes in early reading development: Accommodating individual differences into a model of acquisition. *Issues in Education: Contributions from Educational Psychology, 1*, 1-57. Retrieved from http://www.keithstanovich.com/Site/Research_on_Reading_files/Share_Stanovich_IIE_1995.doc.

Shaywitz, B. A., Lyon, G. R., & Shaywitz, S. E. (2006). The role of functional magnetic resonance imaging in understanding reading and dyslexia. *Developmental Neuropsychology, 30*(1), 613-632. https://doi.org/10.1207/s15326942dn3001_5

Shaywitz, S. (2005). *Overcoming dyslexia: A new and complete science-based program for reading problems at any level*. New York, NY: Vintage Books.

Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R^2 indices. *Multiple Linear Regression Viewpoints, 36*(2), 17-26. Retrieved from http://www.glmj.org/archives/articles/Smith_v39n2.pdf

Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2012, June). Do L1 reading achievement and L1 print exposure contribute to the prediction of L2 proficiency? *Language Learning, 62*(2), 473-505. https://doi.org/10.1111/j.1467-9922.2012.00694.x

Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology, 84*(3), 353-363. https://doi.org/10.1037/0022-0663.84.3.353

Sprenger-Charolles, L., Siegel, L. S., Béchennec, D., & Serniclaes, W. (2003). Development of phonological and orthographic processing in reading aloud, in silent reading, and in spelling: A four-year longitudinal study. *Journal of Experimental Child Psychology, 84*(3), 194-217. https://doi.org/10.1016/S0022-0965(03)00024-9

Statistics New Zealand. (2013, December). *2013 Census ethnic group profiles.* Retrieved from http://www.stats.govt.nz/Census/2013-census/profile-and-summary-reports/ethnic-profiles.aspx

Sternberg, R. J. (Ed.). (2003). *International handbook of intelligence.* Cambridge, United Kingdom: Cambridge University Press.

Sternberg, R. J., & Grigorenko, E. L. (2001). All testing is dynamic testing. *Issues in Education, 7*(2), 138-171. Retrieved from http://search.ebscohost.com.ezproxy.massey.ac.nz/login.aspx?direct=true&db=aph&AN=6435343&site=eds-live&scope=site

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential.* Cambridge, United Kingdom: Cambridge University Press.

Stevenson, C. E., Bergwerff, C. E., Heiser, W. J., & Resing, W. C. (2014). Working memory and dynamic measures of analogical reasoning as predictors of children's reading and math achievement. *Infant and Child Development, 23(1)*, 51-66. https://doi.org/10.1002/icd.1833

Stewart, R. M., Benner, G. J., Martella, R. C., & Marchand-Martella, N. E. (2007, Fall). Three-tier models of reading and behavior: A research review. *Journal of positive behavior interventions, 9*(4), 239-253. https://doi.org/10.1177/10983007070090040601

Stoiber, K., & Gettinger, M. (2016). Multi-tiered systems of support and evidence-based practices. In S. Jimerson, M. Burns, & A. Van der Heyden (Eds.), *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (pp. 121-141). Boston, MA: Springer.

Stuart, M., Stainthorp, R., & Snowling, M. (2008). Literacy as a complex activity: Deconstructing the simple view of reading. *Literacy, 42*(2), 59-66. https://doi.org/10.1111/j.1741-4369.2008.00490.x

Suárez-Coalla, P., Ramos, S., Álvarez-Cañizo, M., & Cuetos, F. (2014). Orthological learning in dyslexia Spanish children. *Annual of Dyslexia, 64*(2), 166-181. https://doi.org/10.1007/s11881-014-0092-5

Swanson, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology, 84*(4), 473-488. https://doi.org/10.1037/0022-0663.84.4.473

Swanson, H. L., & Howard, C. B. (2005, Winter). Children with reading disabilities: Does dynamic assessment help in the classification? *Learning Disability Quarterly*, 28(1), 17-34. https://doi.org/10.2307/4126971

Szumilas, M. (2010, Aug). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry, 19*(3), 227-229. Retrieved from https://www-ncbi-nlm-nih-gov.ezproxy.massey.ac.nz/pmc/articles/PMC2938757/

Tabachnick, B. G., & Fidel, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson Education.

Tissink, J., Hamers, J., & van Luit, J. (1993). Learning potential tests with domain-general and domain-specific tasks. In J. Hamers, K. Sijtsma, & A. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 243-266). Lisse, Netherlands: Swets & Zeitlinger Publishers.

Torgesen, J. K. (2002, January-February). The prevention of reading difficulties. *Journal of School Psychology, 40*(1), 7-26. https://doi.org/10.1016/S0022-4405(01)00092-9

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Burgess, S., & Hecht, S. (1997). Contributions of phonological awareness and rapid automatic naming ability to the growth of word-reading skills in second-to-fifth-grade children. *Scientific Studies of Reading, 1*(2), 161-185. https://doi.org/10.1207/s1532799xssr0102_4

Torppa, M., Poikkeus, A.M., Laakso, M.L., Tolvanen, A., Leskinen, E., Leppanen, P. H., . . . Lyytinen, H. (2007). Modeling the early paths of phonological awareness and factors supporting its development in children with and without familial risk of dyslexia. *Scientific Studies of Reading, 11*(2), 72-103. https://doi.org/10.1080/10888430709336554

Tunmer, W. E., Chapman, J. W., Greaney, K. T., Prochnow, J. E., & Arrow, A. W. (2013). *Why the New Zealand National Literacy Strategy has failed and what can be done about it. Evidence from the Progress in International Literacy Study (PIRLS) 2011 and Reading Recovery monitoring reports.* Massey University Institute of Education. Auckland: Massey University Institute of Education. Retrieved from http://www.massey.ac.nz/massey/fms/Massey%20News/2013/8/docs/Report-National-Literacy-Strategy-2013.pdf

Tunmer, W., & Greaney, K. (2010). Defining dyslexia. *Journal of Learning Disabilities, 43*(3), 229-243. https://doi.org/10.1177/0022219409345009

Tzuriel, D. (2001). *Dynamic assessment of young children.* New York, NY: Springer Science+Business Media. https://doi.org/10.1007/978-1-4615-1255-4

Tzuriel, D. (2014, September). Mediated Learning Experience (MLE) and cognitive modifiability: Theoretical aspects and research applications. *Transylvanian Journal of Psychology*, 15-49. Retrieved from http://search.ebscohost.com.ezproxy.massey.ac.nz/login.aspx?direct=true&db=aph&AN=100190089&site=eds-live&scope=site

van der Vleuten, C. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*(1), 41-67. https://doi.org/10.1007/BF00596229

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice, 18*(3), 137-146. https://doi.org/10.1111/1540-5826.00070

Vellutino, F. R., & Fletcher, J. M. (2005). Developmental dyslexia. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 362-378). Malden, MA: Blackwell Publishing. https://doi-org.ezproxy.massey.ac.nz/10.1002/9780470757642.ch19

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry, 45*(1), 2-40. https://doi.org/10.1046/j.0021-9630.2003.00305.x

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. A., Pratt, A., Chen, R., & Denckla, M. B. (1996, December 1). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*(4), 601-638. https://doi.org/10.1037/0022-0663.88.4.601

Vellutino, F. R., Scanlon, D. M., Small, S., & Fanuele, D. P. (2006, March/April). Response to intervention as a vehicle for distinguishing between children with and without reading disabilities: Evidence for the role of kindergarten and first-grade intervention. *Journal of Learning Disabilities, 39*(2), 157-169. https://doi.org/10.1177/00222194060390020401

Vellutino, F. R., Scanlon, D. M., Zhang, H., & Schatschneider, C. (2008). Using response to kindergarten and first grade intervention to identify children at-risk for long-term reading difficulties. *Read Write, 21*(4), 437-480. https://doi.org/10.1007/s11145-007-9098-2

Vygotsky, L. S. (1978). *Mind in society: The development of hIgher psychological processes.* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.) Cambridge, MA: Harvard University Press.

Wagner, R., Torgesen, J., Rashotte, C., & Pearson, N. A. (2013). Comprehensive Test of Phonological Processing, Second Edition - CTOPP-2. Austin, TX: PRO-ED.

Wang, J.R., & Chen, S.F. (2016). Development and validation of an online dynamic assessment for raising students' comprehension of science text. *International Journal of Science & Mathematics Education, 14*(3), 373-389. https://doi.org/10.1007/s10763-014-9575-4

Wilson, S. B., & Lonigan, C. J. (2010). Identifying preschool children at risk of later reading difficulties: Evaluation of two emergent literacy screening tools. *Journal of Learning Disabilities, 43*(1), 62-76. https://doi.org/10.1177/0022219409345007

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology, 91*(3), 415-438. https://doi.org/10.1037/0022-0663.91.3.415

Wolff, U. (2014). RAN as a predictor of reading skills, and vice versa: Results from a randomised reading intervention. *Annals of Dyslexia*, 64(2), 1-24. https://doi.org/10.1007/s11881-014-0091-6

Yeager, M., & Yeager, D. (2013). *Executive function and child development.* New York, NY: WW Norton & Company.

Yeong, S. H., Fletcher, J., & Bayliss, D. M. (2014, May 19). Importance of phonological and orthographics skills for English reading and spelling: A comparison of English monolingual and Mandarin-English bilingual children. *Journal of Educational Psychology*, *106*(4), 1-15. https://doi.org/10.1037/a0036927

Ziegler, J. C., Perry, C., & Zorzi, M. (2014). Modelling reading development through phonological decoding and self-teaching: Implications for dyslexia. *Philosophical Transactions of the Royal Society, 369*(1634), 1-9. https://doi.org/10.1098/rstb.2012.0397

Zoccolotti, P., De Luca, M., Marinelli, C. V., & Spinelli, D. (2014, November). Modeling individual difference in text reading fluency: A different pattern of predictors for typically developing and dyslexic readers. *Frontiers in Psychology, 5*, 1-18. https://doi.org/10.3389/fpsyg.2014.01374

Appendix A:
Exploratory work for dynamic assessment scoring

In this Appendix, the exploratory work undertaken to determine the scoring procedure for the dynamic assessment is outlined. This includes the pilot testing conducted to refine the Learning Scale (for the DA Learning score) and Strategy Scale (for the DA Strategy score), as well as the statistical analyses conducted to determine the method for calculating the DA Modifiability and DA Dichotomous scores.

A.1 Pilot testing of Dynamic Assessment of Pseudoword Decoding

Prior to embarking on the main study, pilot testing of the dynamic assessment measure was conducted. This was done for three main reasons. First, the Learning Scale (used to guide the assessor in assigning a DA Learning score), was newly developed by this researcher, and the Strategy Scale (for DA Strategy score) was adapted and expanded from that used by Petersen and colleagues (2015, 2016). The pilot test was used to investigate the use of these scales in determining the DA Learning and DA Strategy scores. Second, pilot testing gave the researcher the opportunity to practise administration and scoring of the dynamic assessment to help ensure accurate and efficient administration and scoring during the main study. Third, inter-rater reliability checks were conducted as part of the pilot testing. Here, one of the researcher's supervisors scored the same children tested and scored by the researcher during the pilot testing, with an inter-rater reliability coefficient of 1 (100% agreement) achieved.

The pilot testing was completed at a decile 4 school on Auckland's North Shore that draws children from a wide range of ethnic, language, and socio-economic backgrounds. The class teacher was asked to give consent forms to the parents of all children who had just begun

formal schooling, and a total of 11 completed consent forms were returned. All 11 of these children were included in the pilot testing.

For the pilot testing, each child was tested individually during school time. Testing took place in a quiet room to reduce the likelihood of external distractions. The test session was on average between five and six minutes long and was video-recorded to allow for later analysis and verification of test scoring.

The results of the pilot testing were used to make refinements to both the Strategy Scale and the Learning Scale. For the Strategy Scale, there were only minor changes made to wording to clarify some of the items on the scale. For example, the descriptor for a score of 3 was changed from "Decoding strategy used with two correct sounds produced in sequence (e.g., for the word *mep*, /m/-/e/ or /me/)" to "Decoding strategy used with two correct sounds produced in sequence (e.g., for the word *mep*, /m/-/e/, /me/, /e/-/p/, or /e*p*/) OR whole word with one incorrect sound (e.g., *mip* or *met*]".

The piloted Learning Scale (shown in Figure A.1) was developed based on a range of sources including Peña and Villareal's (2000) *Modifiability observation form* (reproduced in Mann, Peña, & Morgan, 2015); Gutierrez-Clellen and colleagues' *Learning behaviour scale* (Gutierrez-Clellen, Brown, Conboy, & Robinson-Zañartu, 1998), Petersen and colleagues' *Teaching responsiveness scales* (2016); Feuerstein and colleagues' list of cognitive deficiencies (2010); Vye, Burns, Delclos, and Bransford's list of behaviours used to compare cognitive approaches of children to teaching (as cited in Lidz, 1991); Lidz's *Response to mediation scale* (2003); and literature reviews on response inhibition (particularly Baggetta & Alexander, 2016; Yeager & Yeager, 2013). Analysis of the piloted Learning Scale led to minor refinements of wording, particularly that in the Errors item. In Figure A.1, items where wording was changed are italicised.

Figure A.1: Piloted Learning Scale

	5	4	3	2	1
			Internal social-emotional		
Anxiety	Calm	Some signs of slight nervousness.	Uncomfortable	Distressed	Distraught
Persever-ance	Consistently perseveres and persists with task, even if frustrated.	Generally persists with task, with **minimal** encouragement from assessor.	Some signs of apathy, indifference OR unwillingness to continue with task. However, persists with **some** encouragement from assessor.	Frequent signs of apathy, indifference OR unwillingness to continue. However, persists with **significant** encouragement from assessor.	Refuses to attempt any aspect of task.
Motivation	*Enthusiastic and interested in tasks.*	Willingness to try most tasks throughout session (even if experiencing difficulty).	Ambivalent towards tasks.	Occasional attempts to end session or to avoid task(s) OR occasional expression of dislike of tasks.	Refuses to participate OR avoidant behaviours OR consistent expression of dislike of tasks.
			External social-emotional		
Attention	Attentive, focussed, on-task behaviour without prompting.	Maintains attention with occasional prompting.	Distractible but can refocus with regular prompting and/or repetition.	Distracted, difficult to refocus.	Highly distractible, unable to maintain attention to task, and frequent off-task behaviours.
Tractability	*Enthusiastic, cooperative and responsive throughout.*	Consistently responsive.	*Passive or hesitant and minimally responsive.*	Passive noncompliant or uncooperative.	Resistive/ refuses to cooperate or respond.
Task confidence	Consistently demonstrates high confidence behaviours.[4]	Demonstrates high confidence behaviours much of the time.	*Demonstrates neutral confidence behaviours.*	Regularly seeks confirmation from assessor or pauses.	Frequently demonstrates low confidence behaviours.
			Cognitive arousal		
Task compre-hension	Quickly understands all aspects of what is required.	Average comprehension of what is required.	Slow to comprehend, but eventually understands what is required.	Only rudimentary understanding of what is required.	No evidence of task comprehension.
Errors[5]	*Errors not made OR made very rarely (1 or 2).*	*Seldom makes errors.*	*Sometimes makes errors.*	*Frequently makes errors.*	*Very frequently makes errors.*

[4] Task confidence behaviours can be verbal or non-verbal. Examples of high-confidence behaviours include smiling and saying, "This is easy"; "I can do this"; and so on. Examples of low-confidence behaviours include helpless gestures and verbalisations (e.g., "I can't do it!"; "It's too hard"; "I don't know how").

[5] If the assessor models /m/ and the child says /n/, this is an error. If the assessor models *m-e-p* and the child says *m*, *mep*, *me-p* or *mop*, it is an error. If the student self corrects, do not count as an error. However, if the child repeatedly self corrects a score of 3 (some errors) would be appropriate.

	5	4	3	2	1
			Self-regulation		
Response inhibition[6]	Waits for instruction throughout session; does not speak before assessor has finished doing so.	Waits for instruction most of the time during the session.	Some impulsive responses.	Frequent impulsive responses.	Impulsive responses throughout session; consistently responds/talks before assessor has finished speaking.
Self-reward[7]	Consistent use of self-rewarding behaviours or positive feedback for accurate responses.	Use of self-rewarding behaviours or positive feedback most of the time; may sometimes use these behaviours inappropriately[8]	Some use of self-rewarding behaviours or positive feedback; behaviours may be inappropriate.	Minimal use of self-reward or positive feedback; behaviours may be inappropriate.	No use of self-reward or positive feedback.

More significant changes were also made in the form of removing three items from the piloted scale. The piloted Learning Scale included a sub-scale for Self-Regulation (consisting of scores for Response Inhibition and Self-Reward). Analysis of these scores indicated that the scores for Response Inhibition were not correlated with the overall learning score, and that excluding Response Inhibition improved the score reliability from the scale (if Response Inhibition included $\alpha = .897$; if excluded, $\alpha = .908$). For this reason, the Response Inhibition item was removed from the scale. With regards to the Self-Reward item, it became evident that scoring for the Task Confidence and Self-Reward items were focussing on very similar behaviours exhibited by the children. This was supported by analysis that showed the means scores for these items were identical and that the items were highly correlated ($r = .99$). As a result, it was decided to exclude the Self-Reward item, thereby eliminating the entire Self-Regulation subscale. This helped to simplify the scale, without negatively impacting score reliability.

[6] Response inhibition is a measure of the child's impulsivity. High levels of response inhibition (scores at the higher end of the scale) indicate a lower level of impulsivity; low levels of response inhibition indicate high impulsivity.

[7] Self-reward and positive feedback behaviours could be verbal (e.g., "I did it!"; "I knew I could do it") or non-verbal (e.g., smiling; looking pleased).

[8] An example of inappropriate use of self-reward behaviours would be for the child to reward themselves when an incorrect response has been given.

A.2 DA Modifiability score

The DA Modifiability score is intended to provide a representation of the child's overall modifiability by looking at both how well the child responded to input and the child's ability to apply a taught decoding strategy and/or rapidly learn new orthographic representations. Petersen and Gillam (2015) and Petersen and colleagues (2016) based the calculation of the modifiability score on a combination of the learning score (as an indicator of the child's response to input) and the strategy score (as an indicator of the child's ability to apply the taught decoding strategy). However, in the present study, the researcher hypothesised that instead of the strategy score, the sound gain and sound residuum gain scores (how many new sounds the child learnt), could potentially serve as indicators of the child's ability to apply the decoding strategy and/or their rapid orthographic learning. For this reason, the researcher decided to do an exploratory analysis to determine which combination of variables provided a modifiability score with superior predictive validity. This analysis was done using the larger sample from the main study (not the pilot study).

Petersen and Gillam (2015) found that the sound residuum gain score was a better predictor of reading ability than the sound gain score. For this reason, the researcher decided not to include the combination of the learning score and sound gain score in the analysis. Box A.1 summarises the models explored to determine the DA Modifiability score used in this study.

Box A.1: Models explored to determine DA Modifiability score

Model 1	Learning + Strategy
Model 2	Learning + Sound Residuum Gain - All words (Learning + SRG All)
Model 3	Learning + Sound Residuum Gain - Taught words only (Learning + SRG Taught)

A biserial correlational analysis was conducted to investigate the relationship between the different methods to calculate the DA Modifiability score and the T2 Dichotomous score (reading/no reading difficulty at the end of a year at school). All three models were significantly related to the T2 Dichotomous score with large effect sizes ranging from .745 to .846 (all ps = .000).

Following the correlational analysis, logistic regression, and receiver operating characteristic (ROC) analyses were conducted to examine the predictive validity of each of the models. Procedural information about correlational analysis, logistic regression, and ROC analyses is provided in Chapter 4. Results from the logistic regression indicated that all three models were significant fits to the data (all ps = .000; Model 1 $\chi^2(1) = 27.03$, Nagelkerke $R^2 = .518$; Model 2 $\chi^2(1) = 22.193$, Nagelkerke $R^2 = .440$; Model 3 $\chi^2(1) = 22.93$, Nagelkerke $R^2 = .452$), but that the differences in fit for the three models were non-significant.

Area under the ROC curve (AUC) analyses were conducted to examine the predictive accuracy of the different models. The ROC curve is a visual representation of sensitivity and specificity. The true positive rate (sensitivity) is plotted against the false-positive rate (1 - specificity) for different cut points of a parameter. AUC ranges from 0.5 (no better than chance prediction) to 1.0 (perfectly predictive).

The AUCs for all three models were significant. For Model 1 (Learning + Strategy), the AUC was .91, indicating that in 91% of cases the model correctly assigned a higher probability of reading difficulty to children who went on to have a reading difficulty at the end of year at school. For the other two models (Learning + SRG All and Learning + SRG Taught) the AUC was .88. The significance of the difference between AUC results was calculated using the method of DeLong et al. (1988). There was no significant difference between the AUCs of the three models.

Finally, a comparison was made of the sensitivity and specificity of the different models. Model 1 (Learning + Strategy) produced the best balance of sensitivity and specificity (sensitivity = 93%;

specificity = 77%). For both Model 2 (Learning + SRG All) and Model 3 (Learning + SRG Taught), sensitivity (86%) and specificity (76%) were lower than Model 1.

All three models showed strong correlations with the T2 outcome measure, were good fits to the data, and showed strong predictive accuracy. Therefore, for the present study, Model 1 (Learning + Strategy) was adopted for the calculation of the DA Modifiability score, as it achieved the best balance of specificity and sensitivity.

A.3 DA Dichotomous score (DADS)

Petersen and colleagues' (2016) calculated a Dynamic Assessment Dichotomous score (DADS) based chiefly on the child's learning score and the number of sounds gained from pre-test to post-test. This researcher decided to investigate whether other combinations of variables could produce a DADS with superior predictive ability. This was done for two main reasons. First, the exploratory analysis for the calculation of the DA Modifiability score indicated that the combination of the DA Learning and DA Strategy scores provided the best balance of sensitivity and specificity. Second, the work conducted by Petersen and Gillam (2015) found the sound residuum gain score was a better predictor of reading ability than the sound gain score. Box A.2 summarises the models explored to determine the DADS used in this study.

Box A.2: Models explored to determine the DADS

Model 1	Learning, Sound Gain - All words (adapted from Petersen and colleagues (2016))
Model 2	Learning, Sound Gain - Taught words (adapted from Petersen and colleagues (2016))
Model 3	Learning, Strategy
Model 4	Learning, Sound Residuum Gain - All words (Learning, SRG All)
Model 5	Learning, Sound Residuum Gain - Taught words (Learning, SRG Taught)

Correlational analyses indicated significant correlations between the T2 outcome variable and all five of the methods for calculating the DADS, with Phi coefficients ranging from .52 to .59 (all ps < .001).

As all five models were significantly correlated with the T2 outcome measure, logistic regression was conducted to investigate the fit of each model. Results of the logistic regression analyses indicate that all five models were significant fits to the data (see Table A.1). However, the model using a combination of the learning and strategy scores (Model 3) was a significantly better fit than the other four models (χ^2 = 36.32, p = .000, Nagelkerke R^2 = .46).

Table A.1: Comparison of goodness of fit of models for DADS calculation

Model	χ^2	p	R^2_N
(1) Learning, Sound Gain - All	32.10	.000	.41
(2) Learning, Sound Gain – Taught	29.47	.000	.38
(3) Learning, Strategy	36.32	.000	.46
(4) Learning, SRG All	32.43	.000	.41
(5) Learning, SRG - Taught	30.16	.000	.39

Following the logistic regression, the predictive accuracy of the models was investigated using AUC analyses. The AUCs for all three models were significant and ranged from .82 to .86, with the model using the learning and strategy scores producing the largest AUC (see Table A.2).

Table A.2: AUCs for models for DADS calculation

Model	**AUC**	p
(1) Learning, Sound Gain - All	.83	.000
(2) Learning, Sound Gain – Taught	.82	.000
(3) Learning, Strategy	.86	.000
(4) Learning, SRG All	.84	.000
(5) Learning, SRG - Taught	.84	.000

Finally, the sensitivity and specificity of the five models were compared (see Table A.3). This showed that Model 3 (using the learning and strategy scores) also had the best balance of sensitivity (81%) and specificity (90%).

Table A.3: Comparison of sensitivity and specificity of models for DADS calculation

	T2 Dichotomous score	
Model	Sensitivity (%)	Specificity (%)
(1) Learning, Sound Gain - All	75	91
(2) Learning, Sound Gain – Taught	75	89
(3) Learning, Strategy	81	90
(4) Learning, SRG All	81	88
(5) Learning, SRG - Taught	81	86

Consequently, because of its superior model fit and classification accuracy, the DA Learning score and DA Strategy score model was adopted for the calculation of the DADS.

Appendix B:
Information and consent form for parents/caregivers

Dynamic Assessment as an Early Screening Tool for Risk of Reading Difficulty in New Zealand New Entrants

INFORMATION SHEET FOR PARENTS/CAREGIVERS

Dear Parent

My name is Susan Bisschoff, and I am a doctoral student enrolled for a PhD degree in the Institute of Education at Massey University. I am conducting this research for my thesis on the use of dynamic assessment as an early screening tool for risk of reading difficulty.

I have permission from [Principal's name] (Principal) to conduct this research at [School name]. I would like to invite your child to participate in this research and would be grateful if you would allow him/her to participate. This study involves assessing children at two different times: (1) I would like to assess all new entrants as soon as possible after they have started at the school; (2) One year later, I would like to reassess the same children. At this time, I will also ask teachers to provide me with children's reading level (as determined by the teacher). All new entrants to the school who have signed consent forms from their parent/guardian will be able to participate.

I will assess all children participating in the study individually, using short tests of pre-reading and early reading skills. Each assessment session should take under 15 minutes. If at any point a child becomes too tired or no longer wishes to participate, testing will be stopped. Audio-visual recordings will be made of the assessment sessions to ensure accurate record keeping. These recordings will be for reliability checking of scoring only.

All data gathered as part of this study will remain strictly confidential and only group results will be reported. In addition to being used in my thesis, the data collected may be used in journal articles and/or conference presentations. All data will be securely stored for a period of 7 years, after which it will be destroyed. Once the research has been completed, a summary of the findings will be made available to your child's school. Please note that only summary findings from the study will be available, not individual results for your child.

You are under no obligation to accept this invitation. If you decide to allow your child to participate, you or your child have the right to:

- decline to answer any particular question;

- withdraw your child from the study at any time until [insert date];
- ask any questions about the study at any time during participation;
- provide information on the understanding that your name/your child's name will not be used unless you have given permission to the researcher;
- be given access to a **summary** of the project findings when it is concluded.

If you are willing for your child to participate in this study, please complete the attached consent form and return to it to your child's teacher, or directly to me via email or in the post.

- **Email**: [insert email address]
- **Postal Address**: [insert address]

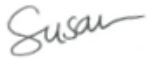
Once you have returned the completed consent form, your child's teacher will be informed that your child may participate in the study so that they are aware that I will be asking the child to participate in the assessment.

Thank you very much for your time and help in making this study possible. If you have any questions or would like further information, please contact me on the email address above, or phone me on [insert mobile number].

My supervisor for this project is Dr Alison Arrow.

Postal Address:	Institute of Education Massey University Manawatū Private Bag 11-222 Palmerston North 4442
Telephone:	+64 (06) 356 9099 ext. 84460
Email:	A.W.Arrow@massey.ac.nz

Kind regards



Susan Bisschoff

Committee Approval Statement

This project has been reviewed and approved by the Massey University Human Ethics Committee: Southern B, Application 16/27. If you have any concerns about the conduct of this research, please contact Dr Rochelle Stewart-Withers, Chair, Massey University Human Ethics Committee: Southern B, telephone 06 356 9099 x 83657, email humanethicsouthb@massey.ac.nz.

Dynamic Assessment as an Early Screening Tool for Risk of Reading Difficulty in New Zealand New Entrants

PARENT/CAREGIVER CONSENT FORM

Researcher: Susan Bisschoff, PhD student, Institute of Education, Massey University

I have read the Information Sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time.

I, as the parent/guardian of ..... [enter child's name], agree for him/her to participate in this research under the conditions set out in the Information Sheet. I give permission for the following.

- My child's reading level (as determined by their teacher after one year at school) to be made available to the researcher for use in this study.
- My child to be audio-visually recorded for response checking, while being assessed as part of this study.

Signature: ..... **Date:** .....

Full Name:
(Please print clearly) .....

Please provide the following information about your child.

IMPORTANT NOTE: Answering these questions is optional. This information will only be used to report group results (ie to describe the population of children generally).

First name: _____ **Family name**: _____

Date of birth: ___ ___ /___ ___ / ___ ___ (D D M M Y Y) **Gender**: ☐ Female ☐ Male

Which ethnic group(s) do you feel your child belongs to? (Please tick next to all that apply)

☐ NZ European/Pākehā ☐ Pasifika (please specify) _____

☐ NZ Māori – Please indicate iwi affiliation ☐ Asian (please specify) _____

_____ ☐ Other (please specify) _____

Main language spoken at home: ☐ English ☐ te reo Māori ☐ Other (please specify)

_____

Appendix C:

Information and consent forms for schools and teachers

Dynamic Assessment as an Early Screening Tool for Risk of Reading Difficulty in New Zealand New Entrants

INFORMATION SHEET FOR SCHOOLS

Dear [insert name]

My name is Susan Bisschoff, and I am a doctoral student enrolled for a PhD degree in the Institute of Education at Massey University. I am conducting this research for my thesis on the use of dynamic assessment as an early screening tool for risk of reading difficulty.

I would like to invite [insert school name] to participate in this research and would be grateful for any assistance you could offer me. This is a longitudinal study which involves assessing children on two separate occasions:

1. I would like to assess all new entrants as soon as feasible after they have started at the school (preferably within their first month at school)
2. Approximately one year later, when the children have been at the school for a full year, I would like to reassess the same children. At this time, I would also ask to be provided with children's reading level (as assessed by their teachers).

Children whose parents/caregivers have given their consent for children to participate will each be individually assessed by myself using a range of short measures of emergent and early literacy skills. On each assessment occasion (ie start of school and after one year at school), assessment should not exceed a total of 15 minutes. If at any point a child becomes too tired or no longer wishes to participate, testing will be stopped. In such a case, an attempt will be made at a later date to assess the child.

Audio-visual recordings will be made of the assessment sessions to ensure accurate record keeping and to allow for analysis at a later time. During testing it is important that distractions to the child are kept to a minimum, and as such I will need a quiet place to assess the children. If possible, it would be ideal if a separate room could be made available for this.

All data gathered as part of this study will remain strictly confidential and only group results will be reported. In addition to being used in my thesis, the data collected may be used in journal articles and/or conference presentations. All data will be securely stored for a period of 7 years, after which it will be destroyed. Once the research has been completed, a research report detailing a summary of the findings will be made available to you.

If you are willing for your school to participate in this study, please complete the attached consent form, signed by yourself or Board or Trustees chair (whoever you feel is most appropriate), and return to me via email or in the post.

- **Email**: [insert email address]
- **Postal Address**: [insert postal address]

Thank you very much for your time and assistance in making this study possible.

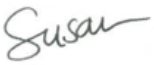
If you have any queries or would like further information, please contact me on the email address above, or phone me on [insert phone number].

My supervisor for this project is Dr Alison Arrow.

Postal Address: Institute of Education
Massey University Manawatū
Private Bag 11-222
Palmerston North 4442

Telephone: (06) 356 9099 ext. 84460
Email: A.W.Arrow@massey.ac.nz

Kind regards



Susan Bisschoff

Committee Approval Statement

This project has been reviewed and approved by the Massey University Human Ethics Committee: Southern B, Application 16/27. If you have any concerns about the conduct of this research, please contact Dr Rochelle Stewart-Withers, Chair, Massey University Human Ethics Committee: Southern B, telephone 06 356 9099 x 83657, email humanethicsouthb@massey.ac.nz.

Dynamic Assessment as an Early Screening Tool for Risk of Reading Difficulty in New Zealand New Entrants

CONSENT FOR SCHOOL TO PARTICIPATE IN RESEARCH

Researcher: Susan Bisschoff, PhD student, Institute of Education, Massey University

School: [insert school name]

I have read the Information Sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time.

I agree to my school's participation in this research under the conditions set out in the Information Sheet.

Signature: ..... **Date:** .....

Full Name:
(Please print clearly) .....

Please provide the name(s) and school contact details of the new entrant teacher(s) at the school.

Name:_____ **Name**:_____

Email:_____ **Email**:_____

Tel:_____ **Tel**: _____

Name:_____ **Name**:_____

Email:_____ **Email**:_____

Tel:_____ **Tel**: _____

Dynamic Assessment as an Early Screening Tool for Risk of Reading Difficulty in New Zealand New Entrants

INFORMATION SHEET FOR TEACHERS

Dear [insert name]

My name is Susan Bisschoff, and I am a doctoral student enrolled for a PhD degree in the Institute of Education at Massey University. I am conducting this research for my thesis on the use of dynamic assessment as an early screening tool for risk of reading difficulty.

I have permission from [insert principal name] to conduct this research at [insert school name], and would be grateful for any assistance you could offer me. This study that involves children on two separate occasions: (1) I would like to assess all new entrants as soon as feasible after they have started at the school (preferably within their first month at school); (2) Approximately one year later, when the children have been at the school for a full year, I would like to reassess the same children. At this time, I would also ask to be provided with children's reading level (as assessed by yourself).

Children whose parents/caregivers have given their consent for children to participate will each be individually assessed by myself using a range of short measures of emergent and early literacy skills. On each assessment occasion (ie start of school and after one year at school), assessment should not take more than a total of 15 minutes. I would be grateful if you would allow me to assess children in your class, and to assist me in communicating with parents to get their consent for their child's participation. This would involve sending information sheets and consent forms home with children (or giving these directly to the parents/caregivers if you were able to do so), and then collecting the completed consent forms.

Audio-visual recordings will be made of the assessment sessions to ensure accurate record keeping and to allow for analysis at a later time. During testing it is important that distractions to the child are kept to a minimum, and as such I will need a quiet place to assess the children.

All data gathered as part of this study will remain strictly confidential and only group results will be reported. In addition to being used in my thesis, the data collected may be used in journal articles and/or conference presentations. All data will be securely stored for a period of 7 years, after which it will be destroyed. Once the research has been completed, a research report detailing a summary of the findings will be made available to the school.

If you are willing to allow me to work with children from your class for this study, please complete the attached consent form, and return to me via email or in the post.

- **Email**: [insert email address]
- **Postal Address**: [insert postal address]

I would then appreciate an opportunity to meet with you to discuss how I can undertake this study in a way that ensures as little as possible disruption to your class or added workload for yourself.

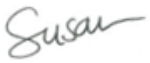
Thank you very much for your time and assistance in making this study possible.

If you have any queries or would like further information, please contact me on the email address above, or phone me on [insert phone number].

My supervisor for this project is Dr Alison Arrow.

Postal Address:	Institute of Education Massey University Manawatū Private Bag 11-222 Palmerston North 4442
Telephone:	(06) 356 9099 ext. 84460
Email:	A.W.Arrow@massey.ac.nz

Kind regards



Susan Bisschoff

Committee Approval Statement

This project has been reviewed and approved by the Massey University Human Ethics Committee: Southern B, Application 16/27. If you have any concerns about the conduct of this research, please contact Dr Rochelle Stewart-Withers, Chair, Massey University Human Ethics Committee: Southern B, telephone 06 356 9099 x 83657, email humanethicsouthb@massey.ac.nz.

Dynamic Assessment as an Early Screening Tool for Risk of Reading Difficulty in New Zealand New Entrants

CONSENT TO PARTICIPATE IN RESEARCH (TEACHERS)

Researcher: Susan Bisschoff, PhD student, Institute of Education, Massey University

School: [insert school name]

I have read the Information Sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time.

I agree to allow children from my class to participate in this research under the conditions set out in the Information Sheet.

Signature: ..... **Date:** .....

Full Name:
(Please print clearly) .....

Appendix D:
Ethics Committee approval

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Date: 06 September 2016

Dear Susan Bisschoff

Re: Ethics Notification - **SOB 16/27 - Dynamic Assessment as an Early Screening Tool for Risk of Reading Difficulty in New Zealand New Entrants.**

Thank you for the above application that was considered by the Massey University Human Ethics Committee: **Human Ethics Southern B Committee** at their meeting held on **Tuesday, 6 September.**

Approval is for three years. If this project has not been completed within three years from the date of this letter, reapproval must be requested.

If the nature, content, location, procedures or personnel of your approved application change, please advise the Secretary of the Committee.

Yours sincerely

Dr Brian Finch
Chair, Human Ethics Chairs' Committee and Director (Research Ethics)

Research Ethics Office, Research and Enterprise
Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand **T** 06 350 5573; 06 350 5575 **F** 06 355 7973
E humanethics@massey.ac.nz **W** http://humanethics.massey.ac.nz

Appendix E:
Assumptions and conditions of binomial logistic regression

As with linear regression, there are several assumptions and conditions that apply to binomial logistic regression. These are summarised in Box E.1.

Box E.1: Assumptions of binomial logistic regression

- Dependent variable must be binary (dichotomous) with two levels coded '1' and '0'
- Absence of outliers
- Linearity of the logit of the outcome variable and any continuous predictors
- Absence of multicollinearity among predictors
- Independence of errors
- Sufficient expected frequencies in each category

For all logistic regression analyses, the dependent variable was dichotomous with *reading difficulty* coded as 1, and *no reading difficulty* coded as 0. Residual statistics (Cook's distance, leverage, standardised residuals, and DFBeta values) were analysed for each of the logistic regression models to identify any substantial outliers or influential cases. Any outliers or influential cases were investigated and were all found to be genuine outliers or influential cases, most frequently when a child identified as having a reading difficulty at T2 was classified as not at risk using the predictor measure in question. In most cases, the same children were outliers for more than one model (i.e., they were classified as having a reading difficulty after a year at school but were identified as not at risk by more than one of the predictor measures).

Analysis of the linearity of the logit with the continuous predictors indicated that the assumption of linearity of the logit had been met for all models. Similarly, tests for multicollinearity in models containing more than one predictor showed an absence of collinearity between the predictor variables.

In education research, a common way in which the assumption of independent errors can be violated is if children are clustered in a hierarchical structure, for example, clustered within classes or schools, with children from the same school and in the same classes tending to be more similar than those in other classes and/or schools (National Centre for Research Methods, 2011). In the current study, participants were drawn from across 10 different schools from a range of deciles. All children beginning school at the participant schools were invited to participate in the study, regardless of the class they were in. When these same children were followed up after a year at school, they were spread across different classes within the school.

Another possible way in which the assumption of independent errors can be violated is if children are tested together and are therefore influenced by, or copy, each other's responses (Pett, 2016). However, for the current study, children were tested individually and generally on the same day so that there was little if any chance for them to discuss the testing process or test items with each other. The type of test items and how the tests were administered (many of them timed and therefore fast-paced) would also make it difficult for children to remember and share information with one another about the test items and their own responses.

Inspection of cross-tabulations for each of the predictor variables by the T2 outcome measure, as well as of the estimated coefficients and estimated standard errors for each of the models, indicated that the assumption of sufficient expected frequencies in each category was met, and that other possible related numerical issues such as overfitting or over-dispersion were not present (Hosmer et al., 2013).

Appendix F:
Logistic regression model summaries (continuous measures)

Table F.1 and Table F.2 provide full summaries of the fitted models using the T2 outcome variable and the dynamic and static predictor variables that provided the best model fit and balance of sensitivity and specificity, namely DA Modifiability and DIBELS Next LNF T1.

Table F.1: DA Modifiability predictor and T2 Dichotomous outcome

Outcome variable	Predictor variable	*b*	*SE (b)*	Wald	*df*	*p*
T2 dichot[a]	DA Mod[b]	-1.2	.32	15.07	1	.000
	Constant	4.9	1.6	10.07	1	.002

[a] T2 dichot: T2 dichotomous outcome (0 = No reading difficulty; 1 = Reading difficulty)

[b] DA Mod: DA Modifiability combing DA Learning and DA Strategy scores (range = 2-10, higher scores indicate better performance on measure).

Table F.2: DIBELS LNF predictor and T2 Dichotomous outcome

Outcome variable	Predictor variable	*b*	*SE (b)*	Wald	*df*	*p*
T2 dichot[a]	DIBELS LNF[b]	-.39	.14	7.94	1	.005
	Constant	.72	.60	1.43	1	.23

[a] T2 dichot: T2 dichotomous outcome (0 = No reading difficulty; 1 = Reading difficulty)

[b] DIBELS LNF: DIBELS Letter Naming Fluency score (actual range = 0-85, higher scores indicate better performance on measure).

Appendix G:
Logistic regression model summaries (dichotomised measures)

Table G.1 and Table G.2 provide full summaries of the fitted models using the T2 outcome variable and the dichotomised dynamic (DADS) and static (DIBELS Next Composite) predictor variables. For each of these measures, an indication is also given of the odds ratio for the model. The odds ratio represents the odds of a specific outcome, given a particular value for the predictor. In this case, it indicates how much more likely or unlikely it is for the child to be classified as having a reading difficulty at T2, given a certain result on the predictor measure. If the odds ratio is greater than 1, this indicates that as the predictor increases, the odds of the outcome also increase (Field, 2015; Szumilas, 2010). For example, in the case of the model including the DADS, this indicates that the odds of a child identified as at risk (using the DADS measure) being classified as having a reading difficulty, is 39 times higher than that for a child who is identified as *not at risk* (see Tables G.1 and G.2). Alternatively, if the odds ratio is smaller than 1, an increase in the predictor means a decrease in the odds of the outcome. Inspection of the 95% confidence intervals for the odds ratio indicates the odds ratio for both the LNF and DA dichotomous measures are both statistically significant at $p < .05$ as they do not cross 1.

Table G.1: Model summary of logistic regression – DADS (dichotomous) predictor and T2 Dichotomous outcome

Outcome variable	Predictor variable	*b*	*SE (b)*	Wald	*df*	*p*	OR[c]	95% CI (OR)	RR[d]	95% CI (RR)
T2 dichot[a]	DADS[b]	3.66	.71	26.69	1	.000	39.00	9.72 to 156.55	8.1	4.52 to 14.60
	Constant	-3.58	.59	37.48	1	.000	0.03			

[a] T2 dichot: T2 dichotomous outcome (0 = No reading difficulty; 1 = Reading difficulty)

[b] DADS: Dynamic Assessment Dichotomous score (0 = Not at risk; 1 = At risk)

[c] OR: Odd ratio

[d] RR: Relative risk

Table G.2: Model summary of logistic regression – DIBELS Next COMP T1 (dichotomous) predictor and T2 Dichotomous outcome

Outcome variable	Predictor variable	*b*	*SE (b)*	Wald	*df*	*p*	OR[c]	95% CI (OR)	RR[d]	95% CI (RR)
T2 dichot[a]	COMP dichot[b]	2.77	.63	19.37	1	.000	15.95	4.65 to 54.73	4.74	2.87 to 7.81
	Constant	-3.23	.51	40.11	1	.000	0.04			

[a] T2 dichot: T2 dichotomous outcome (0 = No reading difficulty; 1 = Reading difficulty)

[b] COMP dichot: DIBELS Next Composite dichotomous score (0 = Not at risk; 1 = At risk) administered at Time 1 (T1).

[c] OR: Odd ratio

[d] RR: Relative risk

In circumstances where the outcome is relatively common (e.g., occurs in more than about 5% of the population), the odds ratio tends to overestimate relative risk, and in such a case it is suggested that the relative risk values should also be reported (Pett, 2016). Another advantage of reporting relative risk values is that, unlike sensitivity and specificity, likelihood ratios are "reproducible test performance characteristics that tend to remain highly stable across

populations" (Gallagher, 2003, p. 300). For these reasons, the relative risk values are also given for DADS and LNF dichotomous models for this study (see Tables G.1 and G.2).

The relative risk indicates the probability of an outcome (in this case reading difficulty) given a certain condition (in this case being identified as *at risk* or *not at risk* using a specific predictor measure). So, for example, the probability of a child classified as *at risk* using the DADS measure going on to be identified as having a reading difficulty is 8.1 times higher than that for a child who is identified as *not at risk* on the DADS. As was the case with the odds ratio for both the LNF and DADS dichotomous measures, the relative risk values are statistically significant as they do not cross 1. Furthermore, the tighter confidence intervals for the relative risk for both gives some indication that the values for the relative risk are more accurate than that for the odds ratio.