MAKING WRONGS ~~W~~RIGHT: IMPROVING EARLY READING EDUCATION THROUGH AUTOMATION AND PERSONALIZATION

Max Emile van der Velde MSc BA

MAKING WRONGS ₩RIGHT: IMPROVING EARLY READING EDUCATION THROUGH AUTOMATION AND PERSONALIZATION

DOCTORAL THESIS
To obtain
the degree of doctor at the University of Twente,
on the authority of the Rector Magnificus,
prof.dr.ir. A. Veldkamp
on account of the decision of the Doctorate Board,
to be publicly defended
on Thursday 25 June 2026 at 14:30 hours

by

Max Emile van der Velde

This dissertation has been approved by:

Promotors
prof.dr.ir. B.P. Veldkamp

Copromotors
prof.dr. R.C.W. Feskens

Layout and cover design: Arul Raja | www.ridderprint.nl
Print: www.ridderprint.nl

ISBN(print): 978-90-365-7184-5
ISBN(digital): 978-90-365-7185-2

URL: https://doi.org/10.3990/1.9789036571852

Manuscript Committee:

Chair / secretary:	prof.dr. T. Bondarouk	University of Twente
Promotors:	prof.dr.ir. B.P. Veldkamp	University of Twente
Copromotors:	prof.dr. R.C.W. Feskens	University of Twente
Committee Members:	dr. M. J. S. Brinkhuis	University of Utrecht
	dr. J.W. Luyten	University of Twente
	prof.dr. A. P. J. van den Bosch	University of Utrecht
	prof.dr. P. C. J. Segers	Radboud University
	prof.dr. S.M. van den Berg	University of Twente

CONTENTS

Chapter 1

INTRODUCTION

Reading is one of the most fundamental skills required to fully participate in everyday society. It finds its importance through facilitating activities such as learning, communication, critical thinking, and the simple enjoyment of reading itself. However, reading performances among primary and second schoolers have been on the decline in many countries the world over, leading to an alarming increase in functional illiteracy (Meelissen et al., 2023; Mullis et al., 2023).

Reversing this worrying trend might be facilitated through improving children's reading fluency skills, as reading fluency is a critical component for attaining literacy (National Institute of Child Health & Human Development, 2000). Reading fluency concerns the ability to read aloud with accuracy, speed and prosody (Kuhn et al., 2010; Pikulski & Chard, 2005). The critical link between reading fluency and literacy is generally described in terms of the established relationship with reading comprehension (e.g. Amendum et al.,2021; Fuchs et al., 2001; Shinn, 1998). Specifically, research has linked fluency to comprehension through automaticity and prosody (Groen et al., 2018; Kuhn et al., 2010).

Here, automaticity reflects the level of mastery a person has obtained for the combination of quick and accurate reading through instruction and practice (Kim et al., 2021; Logan, 1988). Once automaticity is sufficiently obtained, the process of reading requires limited cognitive resources, allowing the person to focus on more complex cognitive tasks, such as comprehension (Aldhanhani & Abu-Ayyash, 2020; Morris & Perney, 2018). This relationship, in which automaticity facilitates comprehension, follows a general characteristic of reading, as described by the Verbal Efficiency Theory (Perfetti, 1985). Perfetti states that lower complexity tasks, such as reading accurately and/or with tempo, need to be mastered before more complex tasks can be mastered. As such, automaticity leads to comprehension through a person's mastery of accurate and swift decoding and word identification, and facilitates it through an increase in the availability of cognitive resources.

Likewise, prosody has been shown to facilitate or enhance the retention of meaning through the focus and attentiveness that follow from its implementation (Kuhn et al.,2010; Miller & Schwanenflugel, 2008; Silva et al., 2021). Here, prosody concerns expressive components of reading, which are generally reflected through aspects such as phrasing, use of pauses, intonation, stress and pitch (Miller & Schwanenflugel, 2008; Share, 2008). Like automaticity, prosody has been established to affect comprehension throughout primary education (Veenendaal et al., 2016), retaining predictive power when controlling for automaticity (Groen et al., 2018; Veenendaal et al., 2015).

In conclusion, improving reading fluency skills is likely to affect a person's current and eventual ability to comprehend text. However, although these conceptual and theoretical relationships are established, practical and instructional issues currently stand in the way of alleviating reading deficiencies through improving children's reading fluency. Followingly, we shall shortly discuss factors that impact reading enjoyment and development in general, after which we dive into how this dissertation aims to contribute towards an improved reading landscape.

1.1 THE READING CRISIS

As discussed, many countries around the world are currently struggling with declines in literacy skills among children. In some cases, such as in the Netherlands, the situation has become so severe that experts are now referring to it as a reading crisis.

While no sole cause exists for this reading crisis, various contributors have been identified by experts. For example, it has been argued that a reduction in the purchasing and renting of books, a reduction in the reading of books by parents and teachers, and a reduction in the availability of language studies at universities, has naturally lead to a reduction in the quality of reading education (Dera, 2023). Others blame the way in which reading comprehension is currently taught, implicating the unnatural and joyless way in which children are taught to read through a vast ocean of reading methods (Knol, 2024). Correspondingly, the Dutch Reading Foundation (Stichting Lezen) has stressed the importance, and increasing lack, of reading motivation as a major contributor to the reading crisis (Stichting Lezen, 2024).

Given the complexity of the situation, solving the reading crisis is not simple, and it will likely not be resolved through any singular approach, nor dissertation. However, we believe considerable ground can be gained by improving primary reading education, as research has demonstrated that early reading proficiency tends to be predictive of children's eventual reading ability (Verhoeven & van Leeuwe, 2008). As such, improving children's early reading proficiency might provide palpable means of improving reading performances throughout formal education and beyond.

Following this reasoning, the Advanced Speech Technology and Learning Analytics for Child Personalized Reading Education (ASTLA) project was given form (Cucchiarini et al., 2020), and its completion has since served as the main purpose of the current dissertation.

1.2 THE ASTLA PROJECT

ASTLA is a project that aims to contribute towards the optimalisation of reading development through improving reading fluency assessment and instruction. Specifically, ASTLA strives to improve primary reading education through the utilization of Automatic Speech Recognition (ASR) and Learning Analytics (LA). Here, ASR concerns the "independent, machine-based process of decoding and transcribing oral speech" (Levis & Suvorov, 2012, p. 1), while LA reflects "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Society for Learning Analytics, 2011). By integrating these approaches within practice, ASTLA aspires to create an innovative, integrated approach to reading fluency assessment, which includes the generation of

objective and detailed reading diagnostics, and the automatic provision of personalized feedback.

ASTLA is split up into three subprojects, which were all funded through a grant from the Dutch Research Council (NWO: NWO 406.20.TW.009). Out of these subprojects, this dissertation fulfilled the third. The other two projects concerned a post-doc and a second PhD project, respectively tackled by Chistian Tejedor Garcia and Wieke Harmsen. These projects focussed on furthering the current state-of-the-art in Dutch child speech recognition, and on the derivation of speech-based measures that provide diagnostic reading information on all relevant reading fluency characteristics (Cucchiarini et al., 2020). As such, these two projects provided the reading diagnostics used to complete the current dissertation. In return, this dissertation facilitated the completion of the other two projects through the collection of the speech data.

Together, ASTLA's subprojects are to result in the construction of a new reading fluency assessment instrument that facilitates its ambitions, the Speech Enabled Reading Diagnostics App (SERDA). In essence, SERDA is as a measurement instrument that collects and analyses children's speech during reading aloud tasks. In its finalized form, SERDA is envisioned to provide diagnostic information on all aspects of reading fluency, in addition to didactically sound suggestions for improvement. Within this framework, SERDA was envisioned to alleviate some of the issues that currently plague the reading landscape, and to further facilitate teachers in remediating reading deficiencies. In the sections that remain, we will investigate the issues that stand at the forefront of this dissertation and describe how SERDA might contribute towards overcoming them.

1.3 READING FLUENCY'S ISSUES

Throughout this dissertation, we primarily focussed on resolving two types of issues. In general, these .issues relate to the incompleteness of many current reading fluency assessments, and the types of information that current instruments provide within practise. Followingly, we will discuss these issues and how SERDA might serve to counteract them.

First, consider one of the most prominently used reading fluency assessment instruments, the Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DORF; University of Oregon, 2020). While this instrument has been shown to provide reliable and valid indications of children's reading speed and accuracy, no prosodic evaluations are included within its administration. Equivalent issues have been identified for similarly popular assessment instruments within the Netherlands, such as the DMT and AVI (i.e., Drie-Minuten-Toets (DMT), Dutch Decoding Test and AVI; van Til et al., 2018a, 2018b), indicating that this does not solely reflect current reality for the English language. Although this lack of prosodic consideration is in conflict with the earlier discussion of

reading fluency conceptualization, it is not surprising from a practical point of view. Oral reading fluency assessment tends to be a relatively time-intensive endeavour (e.g. COTAN; Egberink & Leng, 2024a, 2024b), which is oftentimes outsourced to teachers. In order to include a measure of prosody, teachers would be required to administer an additional assessment instrument, such as the Multidimensional Fluency Scale (MDFS; Rasinski et al., 2006; Zutell & Rasinski, 1991). To further complicate matters, evaluations of prosody tend to provide relatively subjective measures (Kuhn et al., 2010) that require multiple raters to obtain reliable scores, further increasing the testing burden placed on teachers.

The second issue pertains to the information that can currently be obtained from reading fluency assessment instruments. Reading fluency assessment currently generally comes down to placing a teacher and student in a separate room, where the student will read aloud a set of words or passages within a predefined amount of time. For each word, the teacher marks whether the word was read correctly or not, while also keeping track of the amount of time it took the student to read the entire wordlist or passage. Based on these scores, norm-graded scores are provided, based on which students' reading fluency proficiency is classified. While this approach is not problematic in and of itself, it does provide opportunity for teachers' subjective interpretations to impact children's perceived reading performance. Recent research by Harmsen et al. (2026) suggests that this issue might be more problematic than currently believed. Specifically, they found that at least three raters are required to obtain sufficiently reliable word reading estimates for high stakes tests, indicating that differences between raters can be quite prominent. Consequently, it is hard to argue that current assessment practices for reading fluency are not, at least, partially subjective and thereby potentially unfair.

Furthermore, we argue that the informativeness of scores could be improved. While the norm-based grades that follow from current reading fluency assessments allow for the direct comparison of reading performances between children from different and similar grades, they tend to solely provide task-level information. Therefore, these tests tend to provide limited diagnostic value with regard to the individual difficulties and qualities that characterize a child's reading, unless teachers put in a lot of additional work. As a result, teachers are provided with limited resources to individualize or personalize their instruction.

Finally, it should be noted that reading fluency assessment is currently quite rigid. As administrations are still mostly conducted on paper, there is limited potential to vary the content of specific tasks based on student's interests, especially at scale. In addition, limited variability is possible with regard to the timing and frequency of administration, as teachers oftentimes conduct the assessments during school.

In short, reading fluency instruction could be improved by altering existing assessment methodology to incorporate more objective fluency features, including prosodic ones, while simultaneously reducing the testing burden placed on teachers. Additionally, the

value of reading fluency assessment could be enhanced by increasing the availability of diagnostic information, such that it that facilitates a more personalized approach to reading instruction. Throughout the remainder of this introduction, we discuss the solutions proposed to overcome these challenges.

1.4 AUTOMATIC READING FLUENCY ASSESSMENT

The first solution concerns automating the assessment and scoring of oral reading tasks, such that teachers' testing burdens are reduced, while increasing the objectivity, completeness and level of detail of reading fluency diagnostics. To elaborate, when automating the assessment processes, teachers would no longer be required to spend all their time and attention on the scoring itself and might focus on evaluating children's general performance instead. That is, if they are required to be present during administration at all. In addition, an automated approach has the potential to provide a more systematic, complete and objective reflection of reading performance than human evaluations.

A historically popular approach to automate reading assessment concerns the previously mentioned approach called ASR. Attempts at implementing ASR to improve reading fluency date back more than twenty years (Mostow et al., 2003; Reeder et al., 2007). Since then, a lot of work has been conducted to construct and implement ASR-based reading applications (Bolaños et al., 2013; Loukina et al., 2019; Proença et al, 2015; Sabu & Rao, 2018; Silva et al., 2021). However, most of this work has primarily or solely focused on the extraction of reading errors (Nicolao et al., 2018; Yilmaz et al., 2014), while speed and prosody have received less attention.

The first aim for SERDA is to deliver objective and detailed diagnostics on all reading fluency components. In order for such an approach to be feasible, the mode of administration should be changed from pen-and-paper to digital through the use of a tablet, laptop or telephone. This way, audio recordings can be stored, allowing for the implementation of ASR. Specifically, this would allow for the comparison of what a child read aloud to what the child should have read, providing an indication of accuracy. In addition, digital assessment facilitates the logging of responses, providing easily accessible indications of reading speed at the word level. Moreover, it would facilitate the extraction of prosodic features through analysing patterns in children's reading volume, pitch and pauses. Finally, moving from pen-and-paper to digital assessment has the benefit of providing additional flexibility with regard to the timing, variability of content and the frequency and location of administration, which, in combination with the provision of detailed individual diagnostics, greatly contributes towards implementing our second solution.

1.5 PERSONALIZED READING INSTRUCTION

The second objective of SERDA concerns facilitating the personalization of reading fluency instruction, such that teachers are supported in optimizing the reading development trajectories and reading motivation of their students.

To elaborate, personalized learning has increasingly been identified as crucial to facilitate the optimal development of students during their journey throughout formal education (e.g. Connor & Morrison, 2016; Li & Wong, 2021). As a result, great strides have been made to guide the implementation of personalized learning within educational systems (Bray & McClaskey, 2013, 2015, 2016; Evans, 2012; Grant & Basley, 2014). These efforts have proven effective, as evidenced from the great number of studies that have since summarized attempted implementations and conceptualizations (Berge, 2011; Bernacki et al., 2021; Li & Wong, 2021; Scott et al., 2017; van Schoors et al., 2021; Shemshack & Spector, 2020). These efforts have also appeared merited, as personalization has been shown to improve student's achievements, motivation, understanding, satisfaction and learning efficiency (Falcão et al., 2018; Gómez et al., 2014; Zheng et al., 2021).

The implementation of personalized education has been a popular topic for some time now. Early evaluations date back far, as illustrated by Bloom (1984), who investigated the impact of student-to-teacher ratio on learning more than four decennia ago. Lately, technological developments have allowed for a variety of novel implementations (Chrysafiadi & Virvou, 2015; Scott et al., 2017; Xie et al., 2019). Early examples of such technologically enabled personalisation approaches concern mobile and web-based learning systems (Chen and Chung, 2008; Chrysafiadi & Virvou, 2015; Kompen et al., 2019), while interest has recently primarily evolved towards the implementation of AI to enhance personalization (Bhutoria, 2022; Maghsudi et al., 2021; Tapalova & Zhiyenbayeva, 2022).

While the interest in and benefits of personalized education are well documented, opinions vary with regard to its definition (Schoors et al., 2021; Shemshack & Spector, 2020). Most definitions stress that personalization extends beyond adapting or individualizing instruction, distinguishing itself through a focus on student's ownership of their learning (Li & Wong, 2021). This ownership can be enabled through many aspects of the educational system, such as the variety of education, freedom of choice, student-teacher relationships, and the availability and quality of learning devices, learning opportunities and feedback. Incorporating these perspectives, personalized learning is regarded as learning that is tailored towards students' strengths, needs and interests, including students' wishes with regard to content, timing and form, such that sufficient flexibility and support is provided to ensure skill mastery (Patrick et al., 2013)

To facilitate the accurate tailoring of reading fluency instruction to children's needs, it is of the utmost importance to accurately identify children's individual reading strengths and weaknesses. A popular personalisation approach to identify and differentiate between

such characteristics concerns the specification of reading profiles, which are quantifications or summaries of a learner's understanding, competencies, skills and attributes (Barthakur, 2023) with respect to reading. The profiling of reading performances also knows a long history, dating back as far as the eighties (e.g Argyle, 1989). Since then, a lot of research has focussed on profiling reading development and stability (Foorman et al., 2017; Psyridou et al., 2021; Risberg et al., 2024). However, research rarely focusses solely on reading fluency, and does not focus on profiling children based on all aspects of reading fluency. As such, the current dissertation furthers the existing reading profiling literature by placing SERDA's implementary focus specifically on reading fluency diagnostics.

1.6 THIS DISSERTATION

Throughout this section, the structure of the dissertation is described.

Within chapter two, a scoping review is presented. Throughout this review we investigated how reading fluency is defined and assessed within the literature, and how reliable and valid current assessment instruments are. This review resulted in a systematically constructed overview of current and past reading fluency assessment instruments, as well as their qualities and shortcomings, which was used to inform the development of SERDA. Another important result concerned the finding that reading fluency instruments tended to report high reliability and validity, but rarely incorporated any measure of expressiveness. This indicated a gap between reading fluency definitions and assessment procedures, raising concerns about the validity of current approaches.

Throughout chapter three, we describe the framework behind, and development of, SERDA. Specifically, we discuss how utilizing a combination of ASR and learning analytics might improve reading fluency assessment. In addition, we detail how the reading tasks were constructed, and how the SERDA dataset was collected. The study resulted in the construction of SERDA, as well as its administration to a sample of 653 Dutch primary school children. In addition, some preliminary utility, reliability and validity evidence is provided in favour of SERDA's implementation in practice.

Throughout chapter four, we provide an in-depth discussion of the validity of SERDA's accuracy, speed and automaticity metrics. For these purposes, we utilized the Argument-Based-Approach to validation (ABP; Kane, 1992, 2006, 2013). Specifically, we answer the question of whether an oral reading fluency assessment instrument that utilizes ASR can provide valid word decoding and passage reading scores. The results provide evidence that reliable and valid word decoding and passage reading measures can be generated using an ASR-based oral reading fluency assessment instrument.

Throughout chapter five, we examine the potential of using the detailed diagnostics obtained from SERDA to construct distinct, practically and didactically relevant reading

fluency profiles. Similar to chapter four, we took inspiration from the ABP to evaluate the quality of the optimal set of profiles, placing a strong emphasis on their practical utility. The results of this study provide evidence that distinct, practically and theoretically relevant reading fluency profiles can be developed.

Throughout chapter six, we evaluated whether Large Language Models (LLM's) can be used to transform SERDA's reading diagnostics, reading profiles, and children's deviations from their profile into instructional feedback for teachers. In addition, we evaluated to what degree various degrees of contextualisation impact feedback quality. Specifically, we compared the readability, coherence, didactical quality, the number of mistakes and the length of human feedback to feedback generated by three LLM's that varied with regard to contextualisation. The results of the study indicate that LLM's generate feedback that is of relatively high quality. Especially the readability and limited length of feedback was deemed promising. At the same time, the coherence and didactical quality of feedback showed lower scores than human feedback, thereby requiring further finetuning before practical implementations are be considered.

Throughout chapter seven, we present a discussion of the work conducted throughout this dissertation. A reflection is provided on whether the project has achieved the goals set out at its inception, and the lessons learned throughout its completion. Finally, we discuss suggestions for future work that might be conducted to build upon the foundation that was build here.

REFERENCES

Aldhanhani, Z. R., & Abu-Ayyash, E. A. (2020). Theories and Research on Oral Reading Fluency: What Is Needed?. *Theory and Practice in Language Studies, 10*(4), 379–388 http://dx.doi.org/10.17507/tpls.1004.05.

Amendum, S.J., Conradi, S.K., & Liebfreund, M.D. (2021). Explaining reading variance by student subgroup: Should we move beyond oral reading fluency? *Journal of Research in Reading, 44*(4), 757–786. https://doi.org/10.1111/1467-9817.12371.

Argyle, S. (1989). Miscue analysis for classroom use. *Reading Horizons, 29*(2), 93–102. https://scholarworks.wmich.edu/reading_horizons/vol29/iss2/2/.

Barthakur, A., Dawson, S., & Kovanovic, V. (2023). *Advancing leaner profiles with learning analytics: A scoping review of current trends and challenges*. LAK23: 13th international learning analytics and knowledge conference, New York, USA. https://doi.org/10.1145/3576050.3576083.

Berge, Z. L. (2011). If you think socialisation in mLearning is difficult, try personalisation. *International Journal of Mobile Learning and Organisation*, 5, 231–238. https://doi.org/10.1504/IJMLO.2011.045314.

Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose (s)?. *Educational Psychology Review, 33*(4), 1675-1715. https://doi.org/10.1007/s10648-021-09615-8.

Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, *3*, 100068. https://doi.org/10.1016/j.caeai.2022.100068.

Bloom, B. S. (1984). The 2 Sigma Problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16. https://doi.org/10.3102/0013189X01300600.

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013). Human and automated assessment of oral reading fluency. *Journal of Educational Psychology, 105*(4), 1142–1151. https://doi.org/10.1037/a0031479.

Bray, B., & McClaskey, K. (2013). A Step-by-Step Guide to Personalize Learning. *Learning & Leading with Technology, 40*(7), 12–19.

Bray, B. & McClaskey, K. (2015). *Make learning personal. The What, Who, Wow, Where and Why.* SAGE Publications Ltd., USA.

Bray, B., & McClaskey, K. (2016). *How to personalize learning: A practical guide for getting started and going deeper.* Corwin Press.

Chen, C. M., & Chung, C. J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education, 51*(2), 624–645. https://doi.org/10.1016/j.compedu.2007.06.011.

Chrysafiadi, K., & Virvou, M. (2015). *Advances in personalized web-based education*. Springer International Publishing. ISO 690.

Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: Implications for policy and practice. *Policy insights from the behavioral and brain sciences, 3*(1), 54–61. https://doi.org/10.1177/2372732215624931.

Cucchiarini, C., Veldkamp, B. P., Strik, H. (2020). *Advanced speech technology and learning analytics for child personalized reading education* [Unpublished Grant Proposal]. Radboud University.

Dera, J. (2023). De leescrisis is overall. *Tijdschrift Lezen.* Retrieved on 10 February 2026 from https://www.lezen.nl/tijdschrift-artikel/de-leescrisis-is-overal/.

Egberink, I.J.L. & Leng, W.E. de. (2024a). *2010, AVI.* Retrieved on 10 February 2026 from https://www.cotandocumentatie.nl/beoordelingen/b/14564/avi-toets/.

Egberink, I.J.L. & Leng, W.E. de. (2024b). 2010, *Drie-Minuten-Toets.* Retrieved on 10 February 2026 from https://www.cotandocumentatie.nl/beoordelingen/b/14566/drie-minuten-toets.

Evans, M. (2012). *A guide to personalized learning: Suggestions for the Race to the Top–District competition.* Innosight Institute: An education white paper.

Falcão, T. P., Peres, F. M. A., Sales de Morais, D. C., & da Silva Oliveira, G. (2018). Participatory methodologies to promote student engagement in the development of educational digital games. *Computers & Education, 116*, 161–175. https://doi.org/10.1016/j.compedu.2017.09.006.

Foorman, B. R., Petscher, Y., Stanley, C., & Truckenmiller, A. (2017). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of Research on Educational Effectiveness, 10*(3), 619–645. https://doi.org/10.1080/19345747.2016.1237597.

Fuchs, L., Fuchs, D., Hosp, M., And Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256. https://doi.org/10.1207/S1532799XSSR0503_3.

Gómez, S., Zervas, P., Sampson, D. G., & Fabregat, R. (2014). Context-aware adaptive and personalized mobile learning delivery supported by UoLmP. *Journal of King Saud University - Computer and Information Sciences, 26*(1), 47–61. https://doi.org/10.1016/j.jksuci.2013.10.008.

Grant, P., & Basye, D. (2014). *Personalized learning: A guide for engaging students with technology.* International Society for Technology in Education.

Groen, M. A., Veenendaal, N. J., & Verhoeven, L. (2018). The role of prosody in reading comprehension: evidence from poor comprehenders. *Journal of Research in Reading, 42*(1), 37–57. https://doi.org/10.1111/1467-9817.12133.

Harmsen, W., Hubers, F., van Hout, R., Cucchiarini, C., Strik, H. (2026). Reading accuracy assessment in first graders: comparing assessments by teachers and machines. *In review.*

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535. https://doi.org/10.1037/0033-2909.112.3.527.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000.

Kim, Y. S. G., Quinn, J. M., & Petscher, Y. (2021). What is text reading fluency and is it a predictor or an outcome of reading comprehension? A longitudinal investigation. *Developmental Psychology, 57*(5), 718–732. https://doi.org/10.1037%2Fdev0001167.`

Knol, J. J. (2024). De gruwel die 'begrijpend lezen' heet. *Cultureel Kapitaal. Debat over cultuurredactie en cultuurparticipatie.* Retrieved on 10 February 2026 from https://www.lkca.nl/opinie/de-gruwel-die-begrijpend-lezen-heet/.

Kompen, R. T., Edirisingha, P., Canaleta, X., Alsina, M., & Monguet, J. M. (2019). Personal learning Environments based on Web 2.0 services in higher education. *Telematics and informatics, 38*, 194-206. https://doi.org/10.1016/j.tele.2018.10.003.

Levis, J., & Suvorov, R. (2012). Automatic speech recognition. In: Chapelle, C. A. (2012). *The encyclopedia of applied linguistics.* Hoboken, NJ : John Wiley & Sons.

Li, K. C., & Wong, B. T. M. (2020). Features and trends of personalised learning: a review of journal publications from 2001 to 2018. *Interactive Learning Environments, 29*(2), 182–195. https://doi.org/10.1080/10494820.2020.1811735.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492.

Loukina, A., Klebanov, B. B., Lange, P. L., Qian, Y., Gyawali, B., Madnani, N., Misra, A., Zechner, K., Wang, Z., & Sabatini, J. (2019). *Automated Estimation of Oral Reading Fluency During Summer Camp e-Book Reading with MyTurnToRead.* INTERSPEECH 2019: Graz, Austria. https://www.iscaarchive.org/interspeech_2019/loukina19_interspeech.pdf.

Kuhn, M. R., Schwanenflugel, P. & Meisinger, E. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 232–253. https://doi.org/10.1598/RRQ.45.2.4.

Maghsudi, S., Lan, A., Xu, J., & van Der Schaar, M. (2021). Personalized education in the artificial intelligence era: what to expect next. *IEEE Signal Processing Magazine, 38*(3), 37-50. https://doi.org/10.1109/MSP.2021.3055032.

Meelissen, M. R. M., Maassen, N. A. M., Gubbels, J., van Langen, A. M. L., Valk, J., Dood, C., Derks, I., In 't Zandt, M., & Wolbers, M. (2023). *Resultaten PISA-2022 in vogelvlucht. Enschede*: Universiteit Twente. https://doi.org/10.3990/1.9789036559461.

Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading research quarterly, 43*(4), 336–354. https://doi.org/10.1598/RRQ.43.4.2.

Morris, D., & Perney, J. (2018). Using a sight word measure to predict reading fluency problems in grades 1 to 3. *Reading & Writing Quarterly, 34*(4), 338–348. https://doi.org/10.1080/10573569.2018.1446857.

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., & Tobin, B. (2003). Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research, 29*, 61–117. https://doi.org/10.2190/06AX-QW99-EQ5G-RDCF.

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). PIRLS 2021 *International Results in Reading.* Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342.

National Institute of Child Health and Human Development. (2000). Report of the National Reading Panel. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* U.S. Government Printing Office: Washington, DC. https://www.nichd.nih.gov/publications/pubs/nrp/smallbook.

Nicolao, M., Sanders, M., & Hain, T. (2018). *Improved acoustic modelling for automatic literacy assessment of children*. Interspeech 2018, Hyderabad, India, 2-6 September. https://doi.org/10.21437/Interspeech.2018-2118.

Patrick, S., Kennedy, K., & Powell, A. (2013). Mean what you say: *Defining and integrating personalized, blended and competency education*. International Association for K-12 Online Learning (iNACOL).

Perfetti, C. (1985). *Reading ability*. New York: Oxford University Press.

Pikulski, J.J., & Chard, D.J. (2005). Fluency: Bridge between decoding comprehension. *The Reading Teacher, 58*(6), 510–519. https://doi.org/10.1598/RT.58.6.2.

Proença, J., Celorico, D., Candeias, S., Lopes, C., & Perdigão, F. (2015). *Children's Reading Aloud Performance: A Database and Automatic Detection of Disfluencies*. Interspeech 2015, Dresden, Germany, 6-10 September. https://doi.org/10.21437/Interspeech.2015-382.

Psyridou, M., Tolvanen, A., de Jong, P. F., Lerkkanen, M. K., Poikkeus, A. M., & Torppa, M. (2021). Developmental profiles of reading fluency and reading comprehension from grades 1 to 9 and their early identification. *Developmental Psychology, 57*(11), 1840–1854. https://doi.org/10.1037/dev0000976.

Rasinski, T. V., Blachowicz, C., & Lems, K. (2006). *Fluency instruction: Research-based best practices*. New York, NY: Guilford. https://doi.org/10.64152/10125/66619.

Reeder, K., Shapiro, J., & Wakefield, J. (2007). *The effectiveness of speech recognition technology in promoting reading proficiency and attitudes for Canadian immigrant children*. Proceedings of the 9th European Conference on Reading.

Risberg, A. K., Widlund, A., Hellstrand, H., Vataja, P., & Salmi, P. (2024). Profiles of reading fluency and spelling skills: Stability and change across the early school years. *Scandinavian Journal of Educational Research, 68*(6), 1231–1246. https://doi.org/10.1080/00313831.2023.2228822.

Sabu, K., & Rao, P. (2018). Automatic assessment of children's oral reading using speech recognition and prosody modeling. *CSI Transactions on ICT, 6,* 221–225. https://doi.org/10.1007/s40012-018-0202-3.

Van Schoors, R., Elen, J., Raes, A., & Depaepe, F. (2021). An overview of 25 years of research on digital personalised learning in primary and secondary education: A systematic review of conceptual and methodological trends. *British Journal of Educational Technology, 52*(5), 1798-1822. https://doi.org/10.1111/bjet.13148.

Scott, E., Soria, A., & Campo, M. (2017). Adaptive 3D virtual learning environments – A review of the literature. *IEEE Transactions on Learning Technologies, 10*(3), 262–276. https://doi.org/10.1109/TLT.2016.2609910.

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: the perils of overreliance on an" outlier" orthography. *Psychological bulletin, 134*(4), 584–615. https://doi.org/10.1037/0033-2909.134.4.584.

Shemshack, A., & Spector, J. M. (2020). A systematic literature review of personalized learning terms. *Smart Learning Environments, 7*(1), 33. https://doi.org/10.1186/s40561-020-00140-9.

Shinn, M. R. (1998). *Advanced Applications of Curriculum-Based Measurement. Guilford,* New York.

Silva, W. A., Carchedi, L. C., Junior, J. G., de Souza, J. V., Barrere, E., & de Souza, J. F. (2021). A framework for large-scale automatic fluency assessment. *International Journal of Distance Education Technologies, 19*(3), 70–88. https://doi.org/10.4018/IJDET.2021070105.

Society for Learning Analytics. (2011). *What is Learning Analytics?*. Retrieved on 10 Febuary 2026 from: https://www.solaresearch.org/about/what-is-learning-analytics/.

Stichting Lezen. (2024). Leesmotivatie in het onderwijs. *Kwestie van Lezen 17.* Retrieved on 10 Febuary 2026 from lezen.nl/sites/default/files/kwestie_van_lezen_17.pdf.

Tapalova, O., & Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIEd for personalised learning pathways. *Electronic Journal of e-Learning, 20*(5), 639-653. ISO 690. https://eric.ed.gov/?id=EJ1373006.

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., Vloedgraven, J. & De Wijs, A. (2018a). *Wetenschappelijke verantwoording DMT*. Cito: Arnhem.

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., & De Wijs, A. (2018b). *Wetenschappelijke verantwoording AVI*. Cito: Arnhem.

University of Oregon (2020). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Administration and Scoring Guide*. Eugene, OR: University of Oregon. https://dibels.uoregon.edu.

Veenendaal, N.J., Groen, M.A. & Verhoeven, L. (2015). What speech text reading fluency can reveal about reading comprehension. *Journal of Research in Reading, 38*(3), 213–225. https://doi.org/10.1111/1467-9817.12024.

Veenendaal, N.J., Groen, M.A., & Verhoeven, L. (2016). Bidirectional Relations between Text Reading Prosody and Reading Comprehension in the Upper Primary School Grades: A Longitudinal Perspective. *Scientific Studies of Reading. 20*(3), 189–202. https://doi.org/10.1080/10888438.2015.1128939.

Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology, 22*(3), 407–423. https://doi.org/10.1002/acp.1414.

Xie, H., Chu, H. C., Hwang, G. J., & Wang, C. C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education, 140*, 103599. https://doi.org/10.1016/j.compedu.2019.103599.

Yilmaz, E., & Pelemans, J., van Hamme, H. (2014). *Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model*. Interspeech 2014, Singapore, 14-18 September. https://doi.org/10.21437/Interspeech.2014-256.

Zheng, L., Zhong, L., Niu, J., Long, M., & Zhao, J. (2021). Effects of Personalized Intervention on Collaborative Knowledge Building, Group Performance, Socially Shared Metacognitive Regulation, and Cognitive Load in Computer-Supported Collaborative Learning. *Educational Technology & Society, 24*(3), 174–193. https://www.jstor.org/stable/27032864.

Zutell, J., & Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory Into Practice, 30*(3), 211–217. https://doi.org/10.1080/00405849109543502.

Chapter 2

What do they say? Assessment of oral reading fluency in early primary school children: A scoping review

van der Velde, M., Molenaar, B., Veldkamp, B. P., Feskens, R.C.W., & Keuning, J. (2024). What do they say? Assessment of oral reading fluency in early primary school children: A scoping review. *International Journal of Educational Research*, *128*, 102444. https://doi.org/10.1016/j.ijer.2024.102444

ABSTRACT

Recent studies have shown concern regarding an increased risk of functional illiteracy. A way to reverse this trend concerns focusing on improving reading fluency assessment. However, reading fluency assessment, which concerns quick, accurate and expressive reading, knows many shortcomings. To learn how to overcome these shortcomings, we conducted a scoping review. This review included primary and grey literature regarding the assessment and definitions of reading fluency and its underlying components within early primary education. Searches resulted in 107 relevant records. The results indicate that definitions of reading fluency are frequently not provided, and show discrepancies with assessment procedures.

2.1 INTRODUCTION

From reading books to filling in your taxes, being able to read is crucial for many aspects of life and is required for participating in our literate society. However, recent large-scale studies (Gubbels et al., 2019; Mullis et al., 2023) warn of reduced reading ability and an increased risk of functional illiteracy among teenagers in many countries around the world. A way to reverse this trend concerns focusing on individualizing reading instruction, which has increasingly become the focus of education interest (Bray & McClaskey, 2015; Davies et al., 2013). Specifically, individualizing instruction based on improved reading fluency assessment can increase comprehension by facilitating the early detection of, and attending to, reading problems and deficits.

Reading fluency is the ability to read with speed, accuracy and proper expression (Kuhn et al., 2010; Pikulski & Chard, 2005), and has been identified as critical for attaining literacy (National Institute of Child Health and Human Development, 2000). In order to identify how reading fluency assessment and instruction can be individualized, it is essential to map existing attempts. Therefore, a scoping review is conducted with the goal of providing an overview of how reading fluency is currently defined, assessed, and validated for children in early primary education.

2.1.1 Understanding: the importance of fluency to literacy

The cruciality of reading fluency to literacy is exemplified through its relationship with reading comprehension (Amendum et al., 2021; Fuchs et al., 2001; Shinn, 1998). To elaborate, the relationship between reading fluency and comprehension is established to such a degree that interventions have been conducted to improve comprehension through fluency (Mastropieri & Scruggs, 1997; Reutzel & Hollingsworth, 1993). Furthermore, reading fluency has been identified as a bridge between early decoding skills and eventual reading comprehension (Pikulski and Chard, 2005), as decoding skills are predictive of later comprehension (Schaars et al., 2019; Verhoeven & van Leeuwe, 2008), and largely dependent on reading speed (Verhoeven & van Leeuwe, 2009). Prosody, which concerns proper expression, is also related to comprehension (Veenendaal et al., 2016), as it facilitates attaining and maintaining meaning while reading (Cowie et al., 2002; Miller & Schwanenflugel, 2006). In short, the relationship between reading fluency and comprehension has been illustrated both practically and theoretically within the literature.

2.1.2 Criticisms of current measurement tools

Despite the importance of reading fluency, its assessment is subject to criticism. Though measures for speed and accuracy are readily available, such as the Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DORF; University of Oregon, 2020), and the Test of Word Reading Efficiency Sight Word Efficiency (TOWRE-SWE; Torgesen

et al., 1997), these tools limit their informativeness to the test level, forgoing item level diagnostics and complicating individualized instruction. Furthermore, the assessment of expressiveness is currently independent of the assessment of speed and accuracy, and tends to be subjective (Kuhn et al., 2010).

When looking at reading fluency assessment for a more consistent orthography, such as Dutch, the same pattern emerges. Although validated Dutch tools are available (i.e., Drie-minuten-test (DMT), Dutch Decoding Test and AVI; van Til et al., 2018a, 2018b; Verhoeven & Keuning, 2018; Verhoeven et al., 2022), these are also primarily informative at the task level. The time-consuming nature of these tasks further complicates extracting item-level diagnostics, forcing teachers to abandon them in practice.

Taken together, the lack of diagnostics and objectivity, the costs, as well as the recent interest in personalized reading education (Bray & McClaskey, 2015; Davies et al., 2013), call for the creation of a tool that enables the objective, valid and affordable in-class assessment of reading fluency, while providing detailed diagnostic insights that can guide personal learning-to-read trajectories.

2.1.3 The present study

To realize this tool, it is necessary to understand how reading fluency is currently being assessed, and how attempts at overcoming current criticisms have been made. Here, we chose to conduct a scoping review. A scoping review aims to provide an overview of currently available research, to clarify definitions, and to illustrate how research is conducted within the field of interest (Munn et al., 2018; Peters et al., 2015; Peters et al., 2020; Peters et al., 2021; Xiao & Watson, 2019).

Here, the focus on definitions is especially relevant. Though reading fluency's working definition mentions the speed, accuracy and expressiveness of reading (Kuhn, et al., 2010; Pikulski & Chard, 2005), as shown in Figure 1, the tools discussed earlier have all disregarded prosody. This limits construct validity, which concerns the degree to which a tool measures the construct it aims to measure (American Educational Research Association [AERA] et al., 2014; Drenth & Sijtsma, 2005; Reynolds & Livingston, 2021), as assessment practices do not match constructions or definitions of fluency. Therefore, instead of solely investigating fluency assessment, it is also of interest to investigate how reading fluency is current defined within the literature, and whether definitions match up with assessment procedures. Indeed, if the literature shows discrepancies between reading fluency assessment and construction, a scoping review would allow for the specification of an alternative working definition or assessment approach.

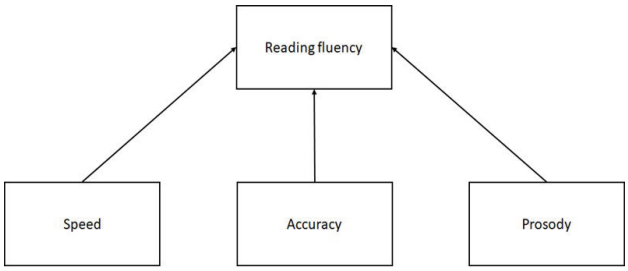


Figure 1. *Traditional Model for Reading Fluency*

Thus, the present study aims to clarify and increase the scientific knowledge on reading fluency assessment and construction through a scoping review. In addition, discrepancies between definitions and assessment descriptions of reading fluency were investigated to evaluate the appropriateness of the current working definition. Specifically, the review aspires to answer the following research questions:

2.1.4 Main question

How are reading fluency, and the components underlying reading fluency, measured in early primary education?

2.1.5 Sub questions

- *How are reading fluency, and the components underlying reading fluency, defined within the current literature?*
- *What tools and techniques are used to assess reading fluency, and the components underlying reading fluency, in early primary education?*
- *How valid and reliable are current measurement tools?*

2.2 METHODS

The scoping review was conducted based on the methodology described by the Joanna Briggs Institute (JBI; Peters et al., 2020), and reported based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) standards (Tricco et al., 2018). A protocol was pre-registered for this review (van der Velde et al., 2022).

2.2.1 Inclusion criteria

This review included studies that discussed the measurement of reading fluency and the components underlying reading fluency (e.g. speed, accuracy, expressiveness), as well as studies that focused on their definitions. Studies that discussed alternative components

were also eligible, as long as they were described or treated as a component of reading fluency. In addition, this review focused on studies that discussed the tools and techniques used to measure reading fluency and its components, as well as studies that focused on the provision of validity or reliability evidence for tools. Here, tools concerned measurement instruments such as the DORF (University of Oregon, 2020), whereas techniques referred to the specific manner in which assessment was conducted, such as calculating the number of words read correctly (WCPM) for a passage. The current review was aimed at early primary education. Therefore, studies that specifically focused on children outside of Grades 1 to 4 of primary education, or mostly discussed children attending special education, second language learners, deaf children, children with dyslexia or studies that refer to "poor readers", have been excluded.

In order to make sure both consistent and inconsistent orthographies are represented, this review included published and unpublished literature written in Dutch and English. This included manuals, justification reports, dissertations, and unpublished papers. However, studies that focused on languages that do not use a Latin-based writing system, like logo-syllabic languages (e.g. Japanese), were excluded to limit the impact of alphabet on the construction of fluency.

2.2.2 Search strategy

The search strategy primarily focused on unearthing published literature. Additionally, gray literature published in Dutch was investigated, as Dutch literature was scarce. The search was created with the help of an information specialist, and consisted of three steps. First, a limited initial search was conducted on the Scopus and Web of Science databases to identify relevant articles. Followingly, the titles, index terms and abstracts of the obtained papers were used to specify keywords for a second, full search. This second search included the Scopus, Web of Science Core Collection, PsycInfo, Psychology and Behavioral Sciences Collection (EBSCO) and Teacher Reference Center (EBSCO) databases. The keywords used are presented in Appendix A. Finally, forward and backward searches were conducted for all literature that passes literature selection.

Dutch gray literature was searched using the Commission Of Test Affairs Netherlands (COTAN) and Meetinstrumenten in de zorg [Measurement-instruments in healthcare] databases.

A preliminary review search was conducted on the fourth of March 2022, using two prominent systematic review databases (PROSPERO and JBI evidence synthesis). Additionally, Scopus, Web of Science and PsycInfo were investigated. Though relevant reviews were found, these focused on aspects of specific tools (Ardoin et al., 2013), test accuracy (Kilgus et al., 2014), or the use of reading fluency as a screening tool (Newell, 2018; Newell et al., 2020). Thus, no current or in-progress reviews on this topic were identified.

2.2.3 Literature selection

Identified literature was stored into EndNote20 (version 20.2.1), after which duplicates were removed. Followingly, two screening procedures were conducted by two independent reviewers.

2.2.3.1 *Screening procedure 1: titles and abstracts.*

The reviewers first screened the literature based on titles and abstracts, which was conducted using ASReview (version 1.1). An explanation of ASReview is provided in the original paper by van de Schoot et al. (2021), while its implementation is presented in the protocol for this review (van der Velde et al., 2022). The screening of titles and abstracts was evaluated using the inter-rater agreement, Cohen's Kappa (Cohen, 1960), and the efficiency of ASReview.

As presented in Appendix B, an overall agreement of 0.84 was found for the first pilot, with a Cohen's Kappa of 0.64, indicating substantial agreement (Cohen, 1960). For the second pilot, an overall agreement of 0.71 was found, with a Cohen's Kappa of 0.34, indicating fair agreement. For the final stage of screening an overall agreement of 0.86 was found, with a Cohen's Kappa of 0.33, indicating fair agreement. ASReview's efficiency was evaluated by calculating the percentage of papers that were deemed irrelevant by one rater, given that it was not evaluated by the other. As the first pilot concerned a predetermined set of papers, ASReview's efficiency was not calculated. During the second pilot 29% (n = 128) of the papers were presented to a single reviewer, of which 88% (n = 112) were deemed irrelevant. For the finalizing stage, 33% (n = 64) of the papers were presented to a single reviewer, of which 94% (n = 60) were deemed irrelevant.

2.2.3.2 *Screening procedure 2: full-text review.*

The second screening procedure concerned entire papers and included one pilot. Within this pilot, both reviewers evaluated the relevance of 10 pieces of literature. Inclusion was evaluated through the use of an exclusion criteria form. After completing the pilot, discrepancies between reviewers were discussed, leading to the finalization of the exclusion criteria form, presented in Appendix C.

As presented in Appendix B, an overall agreement of 0.70 was found for the first pilot, with a Cohen's Kappa of 0.40, indicating fair to moderate agreement (Cohen, 1960). For the final stage, an overall agreement of 0.83 was found, with a Cohen's Kappa of 0.64, indicating substantial agreement.

2.2.4 Data extraction

Data extraction was systematically conducted by two independent reviewers using the extraction tool found in Appendix D. The extracted data concerns general information (e.g. year of publication/authors/Grade), the prevalence and definitions of reading fluency

and its components, as well as the tools and techniques discussed and the validity and reliability evidence for tools.

2.2.5 Analysis and presentation of the evidence

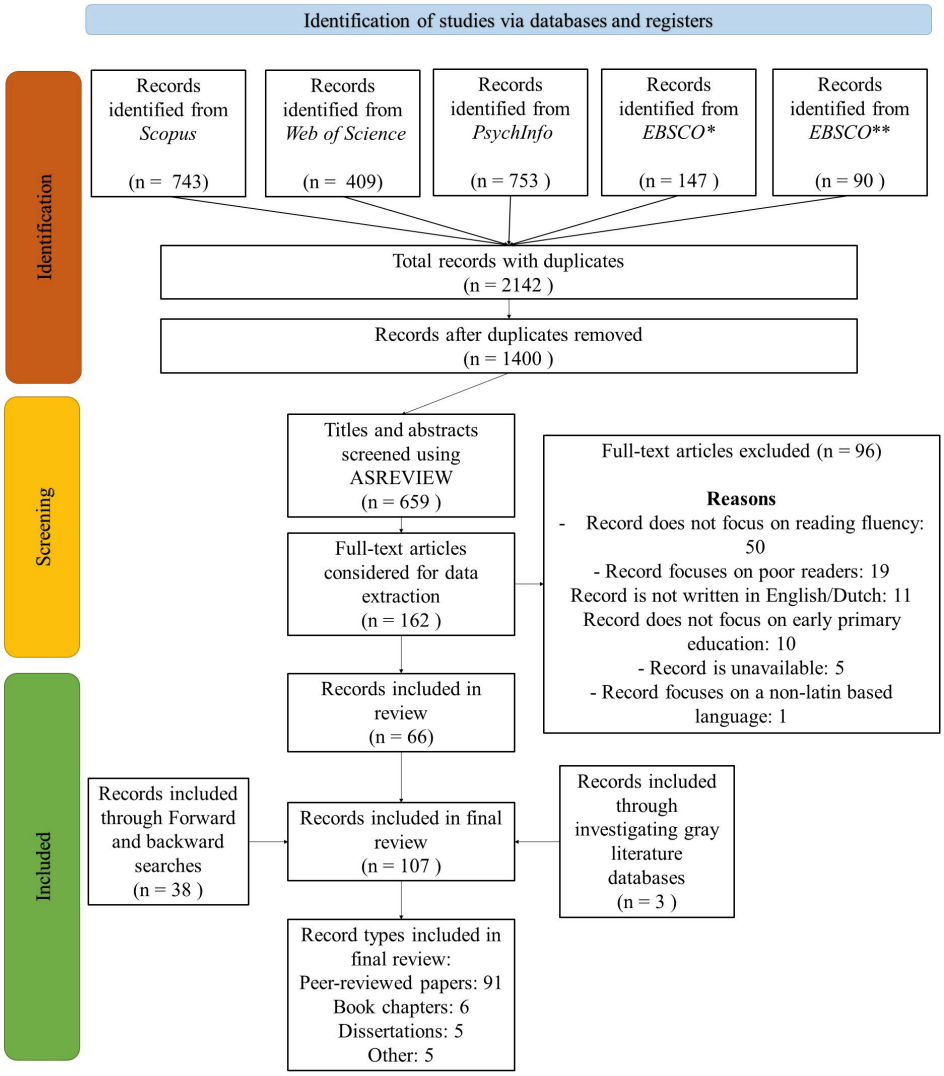
A narrative summary of the findings related the results of the scoping review to the research questions. First, we discussed the results of the search, and some general characteristics of the included records. Then, we answered the research questions by discussing the components that featured in at least 10% (n = 11) of all records, as well as their definitions, tools and techniques. To ensure that the results did not solely reflect the country in which the studies were conducted, and specifically the consistency of the corresponding language, we also compared the definitions and technique descriptions for studies conducted in English speaking countries to those conducted in a country that uses a more consistent language. The selection of this more consistent language was based on the frequency of occurrence. Results will primarily be presented through the use of frequency-tables and word-clouds. Additionally, validity and reliability evidence were averaged per tool, per type, within and over records, after which they were tabulated, and compared between tools. For automatic tools, which shall be elaborated upon within the results, we averaged over all automatic tools instead.

2.3 RESULTS

Searches of online databases resulted in the identification of 2,142 titles and abstracts. After removing duplications and screening titles and abstracts, 162 records were retained for the full-text review. 96 records were excluded based on the criteria: 50 did not focus on reading fluency or its components, 19 focused on “poor” readers, 11 were not written in English or Dutch, 10 did not focus on early primary education, five were unavailable and one was based on a non-Latin based language. The forward and backward search resulted in 38 additional records, while the investigation of gray literature databases provided 3. Hence, data was extracted for 107 records (91 peer-reviewed papers, 6 book chapters, 5 dissertations, 5 other documents). Specifics regarding the identification, screening and inclusion procedures are presented in Figure 2, which is based on the PRISMA methodology (Moher et al., 2009).

Included records were published between 1922 and 2022, and were primarily published between 2005 and 2022. Studies were mostly conducted in the United States of America (n = 49), Portugal (n = 8), and the Netherlands (n = 5). As a result, the definitions and assessment technique descriptions of Portuguese records were compared to those presented in English records. Studies most frequently focused on children attending Grades 2 (n = 55) and 3 (n = 55), followed by Grade 1 (n = 39), and Grade 4 (n = 37). Figure 3 shows

the relative frequency of the constructs discussed in the records, with frequency being represented by area size. Here, reading fluency (n = 82), accuracy (n = 46), speed (n = 43), prosody (n = 36), automaticity (n = 29), reading aloud (n = 14) and word recognition (n = 12) were sufficiently featured, making them the focus of the rest of the review.



Note. * = Psychology and Behavioral Sciences Collection, ** = Teacher Reference Center

Figure 2. *PRISMA Flow Diagram of Inclusion of the Systematic Literature Review*

Note. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses,* = Psychology and Behavioral Sciences Collection, ** = Teacher Reference Center

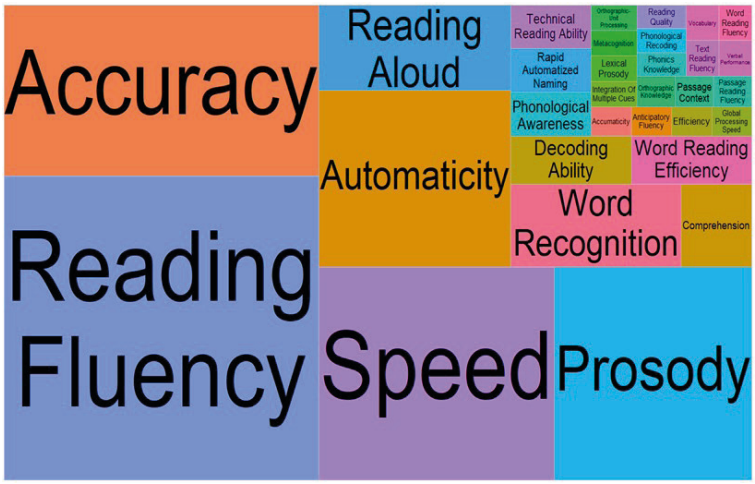


Figure 3. *Relative Frequencies for the Mentioned Constructs, Represented by Area Size*

2.3.1 Definitions of reading fluency

To provide an overview of the definitions found in the literature we describe the most frequently provided definitions of the most frequently mentioned constructs, as well as the words most frequently featured in their definitions.

Table 1 shows the frequency with which the most relevant constructs were discussed and defined. Reading fluency itself was only provided with a definition in 61% of the records in which it was mentioned. Fewer definitions were provided for word recognition (54%), accuracy (22%), speed (17%), and reading aloud (0%), while Prosody (81%), and automaticity (90%) were defined more frequently.

Table 1. Frequency With Which the Most Prevalent Constructs Were Mentioned and Defined

Construct	Discussed	Defined	Proportion defined
Reading fluency	82	50	0.61
Accuracy	46	10	0.22
Speed	43	7	0.16
Prosody	36	29	0.81
Automaticity	29	26	0.90
Reading aloud	14	0	0
Word recognition	12	7	0.58

The definitions provided tend to describe reading fluency, as a whole, as the ability to read text -orally or aloud- quickly, accurately, and with proper expression[1]. Alternative definitions stressed more complex constructs, such as automaticity[2], effort[3] or comprehension[4]. The words that most frequently feature in definitions of reading fluency are presented in Figure 4A. Among them, "read(ing)" (*n* = 50), "accuracy" (*n* = 38), "speed" (*n* = 31), "expression/prosody" (*n* = 28), and "text(s)" (*n* = 25), are the most prevalent. Based on word-frequency distribution comparisons, no differences were found between the definitions of reading fluency for English and Portuguese records.

Speed is most frequently defined in terms of the pacing of reading[5], and in relation to automaticity[6]. Specifically, speed is described as an indication or operationalization of automaticity. The words that most frequently feature in definitions of speed are presented in Figure 4B. Among them, "automaticity" (*n* = 4), "read(ing)" (*n* = 4), "word(s)" (*n* =3), "speed" (*n* = 2), and "text(s)" (*n* = 2) are the most prevalent. Definitions for accuracy tend to describe accuracy as the ability to read, identify, or decode words correctly[7]. The words that most frequently feature in definitions of accuracy are presented in Figure 4C. Among them, "word(s)" (*n* = 8), "decode(ing)" (*n* = 5), "ability" (*n* = 4), "correctly" (*n* = 3), and "recognize/identify" (*n* = 3) are the most prevalent. Definitions for prosody tended to described prosody as the ability to properly use a combination of phrasing, expression, intonation, stress, pitch, and pauses while reading text[8]. In addition, definitions stressed prosody's relation to comprehension through affecting, preserving, expressing or otherwise facilitating the conveying of meaning[9]. The words that most frequently feature in definitions of prosody are presented in Figure 4D. Among them, "read(ing)" (*n* = 18), "phrasing" (*n* = 17), "intonation" (*n* = 15), "expression" (*n* = 14), "stress" (*n* = 13), "text" (*n* = 13), "pitch" (*n* = 11), and "pause(s/ing)" (*n* = 10) are the most prevalent. Definitions for automaticity focused on two interpretations. On the one hand, automaticity was described as a combination of speed and accuracy[10]. On the other hand, definitions focused on more complex cognitive concepts such as effort, attention, or the facilitation of freeing up cognitive resources for other tasks, such as comprehension[11]. The words that most frequently feature in definitions of automaticity are presented in Figure 4E. Among them, "speed" (*n* = 15), "word(s) (*n* = 12), "effort(less)" (*n* = 11), "accuracy" (n = 8), "read(ing)" (*n* = 8), and "attention" (*n* = 7) are the most prevalent. Definitions used for word recognition tended to describe it as the ability to read, based on some type of internal information or representation[12], stressing the reading, or decoding, of words in isolation, without context[13]. The words that most frequently feature in definitions of word recognition are presented in Figure 4F. Among them, "information" (*n* = 4), and "word(s)" (*n* = 2) are the most prevalent.

[1-48] The indices presented throughout the result section link results to specific sets of records. The linkage is presented in Supplementary Appendix A. The references of these records are found Supplementary Appendix B.

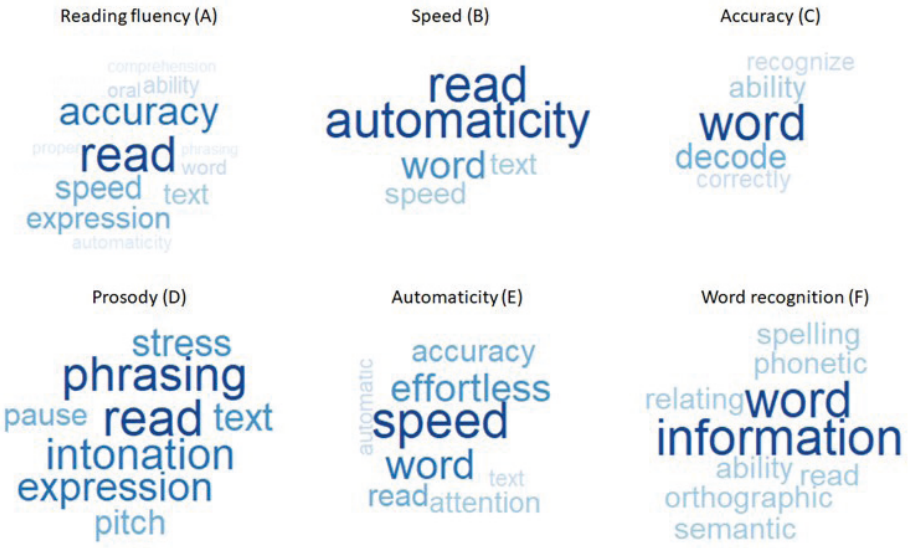


Figure 4. *Word Clouds Showing the Words That Most Frequently Feature in Definitions of Reading Fluency (A), Speed (B), Accuracy (C), Prosody (D), Automaticity (E), and Word Recognition (F)*

2.3.2 Tools

We investigated the most frequently mentioned tools, as well as the frequency with which validity and reliability evidence was provided. In total, 121 mentions of tools were made, most of which assessed reading fluency (n = 72), prosody (n = 14), or accuracy (n = 12).

Including subscales, 55 unique tools were used. Among these, the most frequently mentioned tools are standardized tests: the DORF[14] the TOWRE-SWE[15] and the Test of Word Reading[16] (TLP; Viana et al., 2014). The second most frequently mentioned tools are more generally applicable rating scales like the Multidimensional Fluency Scale[17] (MDFS; Rasinski et al., 2006; Zutell & Rasinski, 1991) and the National Assessment of Educational Progress Oral Reading Fluency Scale[18] (NAEP; Daane et al., 2005). Here, it should be noted that the Curriculum Based Measurement of Oral Reading Fluency (CBM-R) was relatively frequently mentioned. However, as CMB-R primarily functions as assessment methodology, it has been excluded from further analyses. An overview of the most popular tools, and the number of records that discussed them, is presented in Table 2.

Table 2. Number of Records That Discussed and Validated the Most Prevalent Tools

Tool	Discussed	Validity	Reliability
DORF	21	19 (90%)	6 (29%)
TOWRE SWE	9	6 (67%)	5 (56%)
MDFS	8	2 (25%)	1 (13%)
NAEP	5	2 (40%)	3 (60%)
TLP	4	3 (75%)	1 (25%)

Note. DORF = Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency, TOWRE SWE = Test Of Word Reading Efficiency Sight Word Efficiency, MDFS = Multidimensional Fluency Scale, NAEP = National Assessment of Educational Progress Oral Reading Fluency Scale, TLP = Test of Word Reading.

Additionally, 11 mentions were made regarding "automatic assessment tools". Due to the low number of tools mentioned for the other constructs, we only focused on reading fluency, prosody, accuracy, and automatic tools throughout the remainder of this section. Reading fluency, as a whole, was measured by 31 unique tools. The most popular being the DORF, TOWRE-SWE, and the MDFS.

Prosody was measured by seven unique tools. The most popular were the MDFS and the NAEP.

Accuracy was measured by 11 unique tools, out of which only the TLP was used more than once.

2.3.3 Automatic assessment tools

Starting from 2010, authors started mentioning automatic assessment tools. Automatic tools primarily used some type of advanced language processer, such as automatic speech recognition (ASR), to automatically transform speech data into fluency or prosody measures. These tools tended to take the form of online applications, favoring mobiles and computer devices over traditional pen-and-paper methods. Moreover, automatic tools tended to focus on the provision of feedback as opposed to a traditional test-score.

In total, 9 unique tools were identified[19], among which the Fluent Oral Reading Assessment (FLORA; Bolaños et al., 2013) was most frequently mentioned.

2.3.4 Validity and reliability evidence

To determine the validity and reliability of reading fluency assessment tools, original evidence was extracted and evaluated. Included evidence only concerned original results reported in research, and did not include reported results that were not the result of the record itself, such as results from the manual of the instrument.

A complete overview of all collected validity and reliability evidence is presented in Supplementary Appendix C. In total, 171 pieces of validity evidence were extracted from 48 records. Based on prevalence, we investigated the concurrent validity, predictive validity, and classification accuracy for the most relevant tools. With regard to reliability,

62 pieces of evidence were extracted from 36 records. Here, based on prevalence, we evaluated the inter-rater, test-retest and alternate-form reliability for the most relevant tools. The frequency with which the most relevant tools were mentioned, and provided with any validity and reliability evidence, is presented in Table 2.

Table 3 shows the validity evidence, averaged over records (after averaging within records), as well as the standard-deviations between records, and the number of records that provided validity evidence for the most relevant assessment tools. For automatic tools, we averaged the relevant evidence over the records of all instruments, after averaging within records. Here, concurrent validity was calculated using correlations, the predictive validity was determined for both correlations and explained variances, and the classification accuracy concerned accuracy, sensitivity, specificity, positive predictive values, and negative predictive values.

Table 3. Counts, Averages, and Standard Deviations (Between Records) of the Validity Evidence for the Most Frequently Used Measurement Tools

	Concurrent Validity	**Predictive Validity**		**Classification accuracy**				
Tool	**COR**	**COR**	**R2**	**ACC**	**SENS**	**SPEC**	**PPV**	**NPV**
DORF[20]								
Mean	0.72	0.66	49%	0.69	0.72	0.78	0.44	0.93
N	11	5	6	4	5	5	4	4
SD	0.11	0.08	16%	0.16	0.18	0.12	0.14	0.05
TOWRE SWE[21]								
Mean	0.64	0.68		0.78	0.76	0.78	0.38	0.94
N	8	1	-	1	1	1	1	1
SD	0.11							
MDFS[22]								
Mean	0.91	0.81						
N	2	2	-	-	-	-	-	-
SD	0.10	0.12						
NAEP[23]								
Mean	0.80	0.90						
N	2	1	-	-	-	-	-	-
SD	0.26							

	Concurrent Validity	Predictive Validity		Classification accuracy				
Tool	**COR**	**COR**	**R2**	**ACC**	**SENS**	**SPEC**	**PPV**	**NPV**
TLP[24]								
Mean	0.45	0.51		0.75	0.90	0.73		
N	2	1	-	1	1	1	-	-
SD	0							
AUTO[25]								
Mean	0.61	30%		0.84				
N	2	-	1	1	-	-	-	-
SD	0.12							

Note. DORF = Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency, TOWRE SWE = Test Of Word Reading Efficiency Sight Word Efficiency, MDFS = Multidimensional Fluency Scale, NAEP = National Assessment of Educational Progress Oral Reading Fluency Scale, TLP = Test of Word Reading, AUTO = Combined Scores of Automatic Tools, COR = Correlation, R2 = Explained Variance, ACC = Accuracy, SENS = Sensitivity, SPEC = Specificity, PPV = Positive Predictive Value, NPV = Negative Predictive Value. Averages and standard-deviations are calculated between records that discuss evidence, after averaging evidence within records.

With the exception of the MDFS and NAEP, all discussed tools are validated in at least half of the records in which they are mentioned. Here, most validity evidence was found for the DORF, which was almost always validated. Tools showed moderate to strong concurrent and predictive validity. However, the classification accuracy evidence indicated that the performance of tools depended on their intended usage. Namely, low positive predictive values (PPV) were found, while tools showed high negative predictive values (NPV). PPVs provide the probability that a person has a specified condition or disorder, given that the test used to diagnose it provides a positive result (Iverson, 2011b). Contrastingly, NPVs represent the probability that a person does not have a specified condition or disorder, given that the test used to diagnose it provides a negative result (Iverson, 2011a). Thus, the tools mentioned in Table 3 seem useful to assess the absence of a reading deficit or disorder, while diagnosing deficits might be ill-advised. Finally, automatic tools performed comparably to their pen-and-paper counterparts with regard to concurrent validity, while showing high classification accuracy.

Table 4 shows the reliability evidence, averaged over records (after averaging within records), as well as the standard-deviations between records, and the number of records that provided reliability evidence, for the most prevalent assessment tools. Again, the scores for automatic tools were averaged over relevant evidence found in the records of all automatic instruments, after averaging within records.

Table 4. Counts, Averages and Standard Deviations (Between Records) of the Reliability Evidence for the Most Frequently Used Measurement Tools

	Inter-Rater Reliability	Test-Retest Reliability	Alternate-form Reliability
Tool	**COR**	**COR**	**COR**
DORF[26]			
Mean	0.96	0.90	0.95
N	3	1	1
SD	0.05		
TOWRE SWE[27]			
Mean	0.99	0.86	
N	3	1	-
SD	0.01		
MDFS[28]			
Mean	0.90	0.90	
N	1	1	-
SD			
NAEP[29]			
Mean	0.85	0.88	
N	3	1	-
SD	0.03		
TLP[30]			
Mean			
N	-	-	-
SD			
AUTO[31]			
Mean	0.93		
N	3	-	-
SD	0.06		

Note. DORF = Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency, TOWRE SWE = Test Of Word Reading Efficiency Sight Word Efficiency, MDFS = Multidimensional Fluency Scale, NAEP = National Assessment of Educational Progress Oral Reading Fluency Scale, TLP = Test of Word Reading. AUTO = Combined Scores of Automatic Tools, COR = Correlation. Averages and standard- deviations are calculated between records that discuss evidence, after averaging evidence within records.

Compared to validity evidence, reliability evidence was less frequently provided. Specifically, alternate form reliability and test-retest reliability were scarcely presented. However, the reliability estimates that were found, indicated high reliability. The evidence regarding the inter-rater reliability was especially impressive, showing near-perfect scores for the DORF, TOWRE-SWE and automatic tools. These results indicate that although,

ironically, reliability evidence for assessment tools is less prevalent and varied, current assessment tools are highly reliable.

2.3.5 Techniques

We investigated how often techniques were mentioned, what constructs they were related to, and the words used in their descriptions. In total, 185 mentions of techniques were made, most of which were used to assess reading fluency (n = 89), prosody (n = 25), accuracy (n = 24), speed (n = 18), automaticity (n = 5), and word recognition (n = 5).

Reading fluency techniques describe reading fluency assessment using the number of words read -correctly- within 45 seconds, one minute, or three minutes, for a passage of text[32]. Frequently, a median or average score is calculated over multiple passages[33]. The words that most frequently feature in descriptions of reading fluency assessment techniques are presented in Figure 5A. Among them, "word(s)" (n = 102), "read(ing)" (n = 94), "correct" (n = 61), "minute" (n = 59), and "passages" (n = 40), were the most prevalent. Based on word-frequency distribution comparisons, no relevant differences were found between the technique descriptions of reading fluency for English and Portuguese records.

The technical descriptions in Figure 5A show partial overlap with their definitional equivalents in Figure 4A. First, both definitions and techniques stress the reading of passages of text, over reading lists of isolated words. Secondly, both definitions and techniques consistently include the concepts of speed and accuracy, using different terms. However, the definitions of reading fluency frequently include the concept of expression/prosody, whereas techniques solely reflect speed and accuracy.

To quantify the severity of this mismatch, we investigated how often prosody was included in reading fluency technique descriptions, or assessed separately, given that prosody was included in the definition of reading fluency, and given that reading fluency was assessed. Of the 50 records that defined reading fluency, 32 made some mention of prosody, or components of prosody[34]. Out of these, 21 records measured reading fluency[35]. Among them, only 9 (43%) measure prosody in some manner[36]. This discrepancy indicates that reading fluency is not measured in accordance with the working definition, nor with provided definitions.

Speed techniques describe the assessment of speed as the words read per minute[37] for a passage or story[38]. The words that most frequently feature in descriptions of speed assessment techniques are presented in Figure 5B. Among them, "read(ing)" (n = 18) "word(s)" (n = 13), "minute" (n = 12), "time" (n = 8), and "passage" (n = 6), were the most prevalent.

Comparing the definitions presented in Figure 4B to the techniques in Figure 5B, results in few differences. Specifically, both focus on the speed of reading words, and on reading passages of text. However, the techniques do not specifically mention automaticity, while definitions do.

Accuracy techniques mostly used the number, or percentage, of words read correctly[39] for a passage[40], though focus was sometimes placed on errors instead[41]. The words that most frequently feature in descriptions of accuracy assessment techniques are presented in Figure 5C. Among them, "read(ing)" (n = 20) "word(s)" (n = 20), "correctly" (n = 15), "percentage" (n = 8), and "errors" (n = 7), were the most prevalent. The definitions and technique descriptions of accuracy are mostly equivalent. While both focus on the correctness of word reading, the definitions stress the decoding process more strongly. Both also make few statements with regard to the type of task conducted, though techniques did slightly focus on the reading of passages.

Prosody was primarily assessed using rating scales[42]. These scales mostly ran from 1 to 4, tended to be filled in by experts, or were automatically calibrated through an algorithm. Scores for multiple aspect of prosody were combined into a finalized score by averaging or calculating a median score. The aspects most frequently assessed were expression and volume, phrasing, smoothness, and pacing[43]. The words that most frequently feature in descriptions of prosody assessment techniques are presented in Figure 5D. Among them, "read(ing)" (n = 18) "rating" (n = 15), "expression" (n = 14), "pausing" (n = 13) "score" (n = 10), "word" (n = 10), "phrasing" (n = 8), and "volume" (n = 8) were the most prevalent. Prosody definitions had much in common with their technique counterparts. Firstly, both word clouds contain a relatively large number of aspects. Secondly, though their frequencies vary, both word clouds include expression, intonation, phrasing, and pausing. In contrast, stress, pitch, and rhythm only occur frequently within the definitions, while smoothness is only found in technique descriptions. Finally, though both definitions and techniques described prosody as a concept related to reading texts or passages, the definitions tended to stress this more strongly.

Automaticity assessment was always described using the number of words read correctly[44]. In most cases, this concerned the number of words read correctly per minute[45]. The words that most frequently feature in descriptions of automaticity assessment techniques are presented in Figure 5E. Among them, "correctly" (n = 5) "read" (n = 5), "words" (n = 5), and "minute" (n = 4) were the most prevalent.

Though few technique descriptions were provided, the technique-based word cloud does clarify that the measurement of automaticity includes the speed and accuracy aspects found in its definitions. However, the more cognitive aspects of automaticity, such as effort and attention, are left unrepresented. Interestingly, the technique-based word cloud for automaticity is almost entirely equivalent to the technique-based word cloud of reading fluency, and resembles the synthesis of the technique-related speed and accuracy word-clouds.

Word recognition assessment was described as the number of words correctly read, or pronounced[46]. In contrast to the other constructs, here the focus is specifically on reading isolated words from a list[47]. Additionally, stress was placed on the difference between

regular and irregular word reading[48]. The words that most frequently feature in descriptions of word recognition assessment techniques are presented in Figure 5F. Among them, "word(s)" (n = 11) "correct" (n = 5), and "read" (n = 5) were the most prevalent.

Although few technique descriptions were provided, word recognition techniques closely matched their definitions. Namely, both the definitions and techniques stress the reading of isolated words, while using information that is intrinsic to the word being read. Furthermore, technique descriptions closely matched the accuracy counterparts, indicating how similarly these constructs are used.

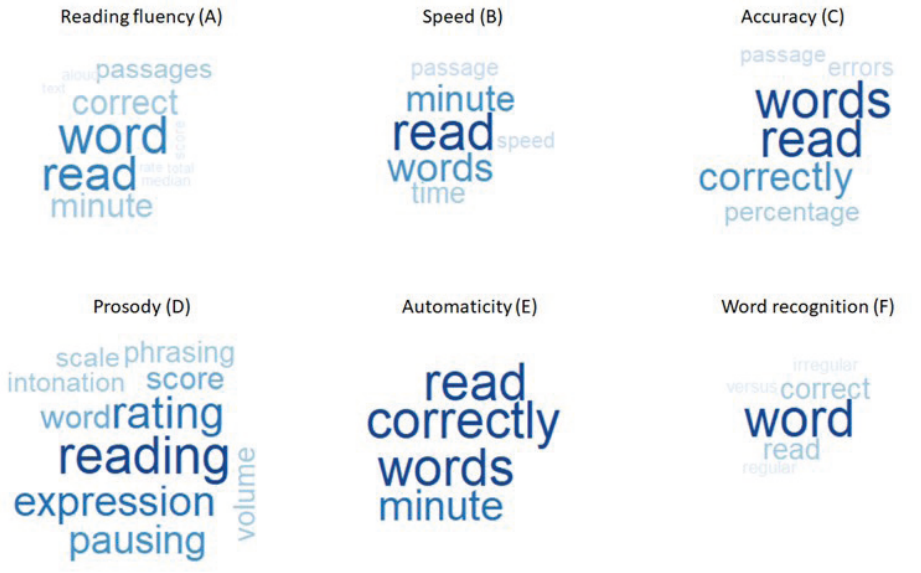


Figure 5. *Word Clouds Showing the Words That Most Frequently Occur in Techniques of Reading Fluency (A), Speed (B), Accuracy (C), Prosody (D), Automaticity (E), and Word Recognition (F)*

2.4 DISCUSSION

This scoping review aimed to improve the scientific knowledge on how reading fluency and its components are defined, assessed, and validated. In addition, discrepancies between definitions and assessment technique descriptions of reading fluency were investigated to evaluate the appropriateness of the current working definition.

2.4.1 Definitions and measures

Reading fluency was defined as the ability to read text -orally or aloud- quickly, accurately, and with proper expression. However, it is important to note that the provision of definitions was scarce, especially given the complex nature of reading fluency. This finding is striking, given that earlier work stressed the importance of defining reading errors (Leu Jr, 1982). And indeed, when we combine the provided definitions with the evidence gathered on assessment, the problematic nature of infrequently defining fluency becomes apparently clear.

To elaborate, the primary issues illustrated throughout this review relate to the mismatches between definitions and assessment techniques. Here, the most frequently discussed construct, reading fluency, also contained the biggest discrepancy. While reading fluency definitions often included prosody, or components of prosody, less than half of the records that mentioned prosody within reading fluency definitions actually included its assessment when assessing reading fluency. Although this finding is remarkable, the incorporation of information on reading fluency assessment tools provides context.

Namely, the DORF and TOWRE SWE are currently fluency's most prominent assessment tools. Fittingly, both have been thoroughly validated, showing promising validity and reliability evidence. However, both tools assess reading fluency through the number of words that a student reads correctly per minute, or 45 seconds. In other words, both tools neglect prosody, reducing construct validity. Thus, though the speed and accuracy components of fluency can currently be validly and reliably assessed, incorporating prosody would improve the validity of reading fluency assessment.

After reading fluency, accuracy was most prominently discussed. Surprisingly, accuracy was almost never defined, providing little information on what this construct represents. The definitions that were provided strongly resembled the word recognition definitions. Namely, the literature describe accuracy as the ability to correctly read, identify or decode words. Correspondingly, word recognition was defined as the ability to read or decode words, in isolation, through some type of internal representation. Within the literature, accuracy tended to be described in more practical terms, whereas word recognition was placed in a theoretical context. Likewise, the technique descriptions for accuracy and word recognition strongly overlapped. Both focussed on the number of words read correctly, with a stronger focus on theoretical components, such as regularity, for word recognition. Taken together, the results of this review imply that accuracy and word recognition are, respectively, used as measurement and theoretical equivalents within the context of reading isolated words. The relationship between speed and automaticity showed a similar, yet more complicated, pattern. Namely, though speed was only incidentally defined, its definitions explicitly described it as an operationalization of automaticity. Meanwhile, automaticity was commonly defined in terms of speed and/or accuracy. However, automaticity definitions often included complex cognitive constructs such as attention, effort, or the

freeing up of resources for other skills, such as comprehension. Similarly, speed techniques only partially overlapped with their automaticity counterparts. While automaticity was described as the number of words read correctly per minute, speed was described as the number of words read per timeframe, excluding correctness. Indeed, Figure 5 illustrates that the discrepancy between speed and automaticity assessment dissipates if accuracy is incorporated. To summarize, though the results portray speed as an operationalization of automaticity, automaticity concerns a more complex construct that, at the very least, requires the inclusion of accuracy.

Furthermore, prosody is defined as the ability to properly use a combination of phrasing, expression, intonation, stress, pitch, and pauses, while reading text. However, definitions and measurement techniques tended to vary with regard to the exact components included. Fittingly, prosodic assessment tools such as the MDFS and the NAEP focus on the evaluation of separate prosody components through the use of rating scales. Though the scales are rated by humans, making them relatively subjective, the evidence presented for these tools is promising, outperforming most reading fluency tools with regard to validity. In short, though only a few tools were frequently mentioned, and despite the fact that they currently lack objectivity, the MDFS and NAEP showed promising validity and reliability evidence and could be useful for building towards future tools.

Finally, though frequently mentioned, reading aloud was never defined or measured. It was only mentioned as a component of fluency in more theoretical discussions.

2.4.2 A working definition and componential model for reading fluency

Based on the results of the review we followingly discuss suggestions to adapt the traditional reading fluency model, which are presented in Figure 6.

Speed and accuracy are now seen as operationalizations of word recognition and automaticity. Specifically, word recognition is the conceptual equivalent of the accuracy of isolated word reading, while automaticity is the conceptual equivalent of the speed and accuracy of word and passage reading. Correspondingly, word recognition is assumed to unidirectionally influence automaticity through mastery of isolated word reading. In addition, prosody is assumed to be affected by automaticity through automaticity's role in freeing up cognitive resources for other tasks. In contrast, the lack of contextual information during word recognition is represented through the absence of a direct relationship with prosody. In essence, the proposed model assumes that speed and accuracy measures tell us something about how well we are able to make the process of reading aloud automatic, while automaticity itself indicates how much cognitive resources are available to generate proper expression.

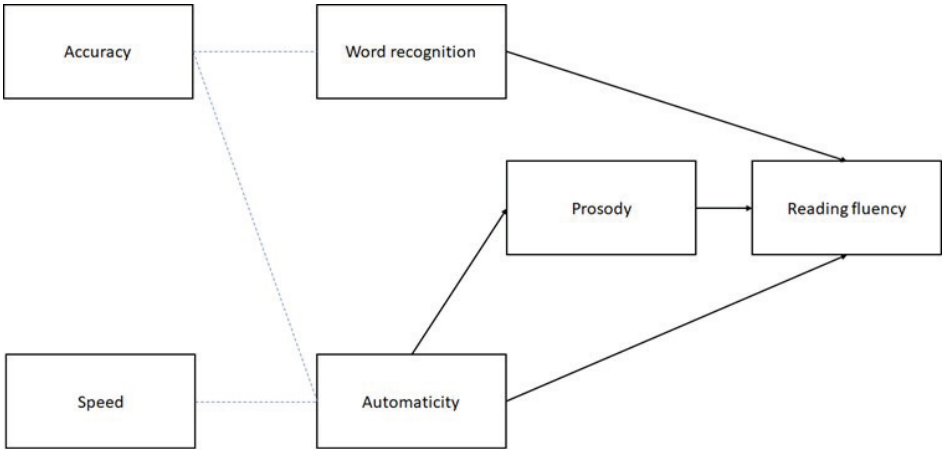


Figure 6. *The Proposed Model for Reading Fluency*

Note. The striped blue lines indicate that constructs are an alternative form of the construct they are related to. Here, speed and accuracy are the measurement equivalents of the theoretical constructs "Word recognition" and "Automaticity".

To conclude, based on the proposed model our advice with regard to a working definition is as follows. Reading fluency should be seen as the ability to fluently read aloud individual words and passages. This ability primarily results from proficient word recognition, automaticity and prosody. Reading fluency is reflected in measures of speed and accuracy, and through the proper use of intonation, phrasing, pitch, pauses, and the usage of stress. Finally, being a fluent reader allows one to focus on more complex tasks or processes, such as comprehension.

2.4.3 Limitations and suggestions

A first limitation relates to the software used to screen titles and abstracts, ASReview, as ASReview might only present a subset of identified titles and abstracts to the reviewer. To minimize this potential loss of information we conducted the screening of titles and abstracts using two reviewers. In addition, we investigated the efficiency of ASReview. The results, presented in Appendix B, suggest a limited loss of information when using two reviewers.

A second limitation concerns the prevalence of records from the United States of America (USA). Given that practice can differ between countries, this could harm the generalizability of the review. Further issues emerge when taking into account the argumentation provided by Share (2008), which illustrates how English, as a language, functions as an outlier with regard to reading. However, though Share (2008) convincingly describes English as an outlier orthography, this does not necessarily imply that assessing English fluency should be viewed as such. Evidence against this view is found in work by Arnesen et al. (2017), who adapted the most popular reading fluency assessment tool

in America, the DORF, for Norwegian. Though Norwegian has a structurally different orthography compared to English, the validation of this adaptation identified it as a reliable and valid measure of reading fluency. In addition, the current study showed no differences between records from English speaking countries and records from a more consistent language, Portuguese, with regard to reading fluency definitions and technique descriptions. Therefore, although most studies included within this review stem from English speaking countries, this does not directly indicate that the results can not be generalized.

Finally, future researchers are advised to implement the suggested adaptations in practice. We suggest researchers to work towards creating an automatic assessment tool, as automatic tools should allow for the simultaneous assessment of all reading fluency components, could facilitate item-level diagnostics by using speech data, and might reduce the costs and time-consuming nature of traditional assessment tools. However, traditional tools should not be abandoned. Instead we suggest adopting the thoroughly validated speed and accuracy assessment methodology of the DORF, and the prosody assessment methodology of the established MDFS and NAEP instruments, as its foundation.

2.5 CONCLUSION

Reading fluency is a relatively complex construct. It is scarcely defined, has varying components, and shows discrepancies between its definitions and assessment techniques. Most worryingly, reading fluency assessment is mostly concerned with measures of speed and accuracy, disregarding prosody. To oppose this trend, we suggest an adapted working definition and measurement model for reading fluency. The definition and model stress the importance of speed, accuracy and prosody, while incorporating word recognition and automaticity. Researchers are advised to validate this model by implementing it in an automatic assessment tool that builds upon the best currently available tools, allowing for the simultaneous assessment of speed, accuracy, and prosody. This tool has the potential to improve assessment validity, reduce costs, lessen testing burdens, and allow for the integration of formative- and summative assessment, opening the door for personalized reading instruction to fend of the increasingly imminent risk of functional illiteracy among future generations.

REFERENCES

Amendum, S. J., Conradi, S. K., & Liebfreund, M. D. (2021). Explaining reading variance by student subgroup: should we move beyond oral reading fluency? J*ournal of Research in Reading, 44*(4), 757¬¬¬–786. https://doi.org/10.1111/1467-9817.12371.

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for Educational & Psychological Testing.* American Educational Research Association. Available from: standards_2014edition.pdf (testingstandards.net)

Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding curriculum-based measurement of oral reading fluency (CBM-R) decision rules. *Journal of school psychology, 51*(1), 1–18. https://doi.org/10.1016/j.jsp.2012.09.004

Arnesen, A., Braeken, J., Baker, S., Meek-Hansen, W., Ogden, T., & Melby-Lervåg, M. (2017). Growth in oral reading fluency in a semitransparent orthography: concurrent and predictive relations with reading proficiency in Norwegian, Grades 2–5. *Reading Research Quarterly, 52*(2), 177–201. https://doi.org/10.1002/rrq.159

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013). Human and automated assessment of oral reading fluency. *Journal of educational psychology, 105*(4), 1142–1151. https://psycnet.apa.org/doi/10.1037/a0031479

Bray, B. & McClaskey, K. (2015). *Make learning personal. The What, Who, Wow, Where and Why.* SAGE Publications Ltd., USA.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cowie, R., Douglas-Cowie, E., & Wichmann, A. (2002). Prosodic characteristics of skilled reading: Fluency and expressiveness in 8–10-year-old readers. *Language and Speech, 45*(1), 47–82. https://doi.org/ 10.1177/00238309020450010301

Daane, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2005). *Fourth-Grade students reading aloud: NAEP 2002 special study of oral reading* (NCES 2006–469). Washington, DC: U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

Davies, D., Jindal-Snape, D., Collier, C., Digby, R., Hay, P., & Howe, A. (2013). Creative Learning Environments in Education – A Systematic Literature Review. *Thinking Skills and Creativity, 8*, 80–91. https://doi.org/10.1016/j.tsc.2012.07.004

Drenth, P. J. D., & Sijtsma, K. (2005). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen*. Bohn Stafleu Van Loghum.

Fuchs, L., Fuchs, D., Hosp, M., And Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scient. Stud. Read. 5*, 239–256. https://doi.org/10.1207/S1532799XSSR0503_3

Gubbels, J., van Langen, A., Maassen, N., & Meelissen, M. (2019). *Resultaten PISA-2018 in vogelvlucht*. Universiteit Twente. https://doi.org/10.3990/1.9789036549226

Iverson, G.L. (2011a). Negative Predictive Power. In: Kreutzer, J.S., DeLuca, J., Caplan, B. (eds) *Encyclopedia of Clinical Neuropsychology*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-79948-3_1219

Iverson, G.L. (2011b). Positive Predictive Power. In: Kreutzer, J.S., DeLuca, J., Caplan, B. (eds) *Encyclopedia of Clinical Neuropsychology*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-79948-3_1234

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of school psychology, 52*(4), 377–405. https://doi.org/10.1016/j.jsp.2014.06.002

Kuhn, M., Schwanenflugel, P. & Meisinger, E. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 232–253. https://doi.org/10.1598/RRQ.45.2.4

Leu Jr, D. J. (1982). Oral reading error analysis: A critical review of research and application. *Reading Research Quarterly, 17*(3), 420–437. https://doi.org/10.2307/747528

Mastropieri, M. A., & Scruggs, T. E. (1997). Best practices in promoting reading comprehension in students with learning disabilities. *Remedial and Special Education, 18*, 197–213. https://doi.org/10.1177/074193259701800402

Miller, J., & Schwanenflugel, P.J. (2006). Prosody of syntactically complex sentences in the oral reading of young children. *Journal of Educational Psychology, 98*(4), 839–853. https://doi.org/10.1037/0022-0663.98.4.839.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The, P. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine, 6*(7), e1000097. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). PIRLS 2021 International Results in Reading. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342

Munn, Z., Stern, C., Aromataris, E., Lockwood, C., & Jordan, Z. (2018). What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC medical research methodology, 18*(1), 1–9. https://doi.org/10.1186/s12874-017-0468-4.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. U.S. Government Printing Office: Washington, DC.

Newell, K. (2018). *An evaluation of the use of oral reading fluency as a screening tool with emerging biliterates.University of Minnesota Digital Conservancy*. Retrieved on 6 February 2026, from https://hdl.handle.net/11299/206639

Newell, K. W., Codding, R. S., & Fortune, T. W. (2020). Oral reading fluency as a screening tool with English learners: A systematic review. *Psychology in the Schools, 57*(8), 1208–1239. https://doi.org/10.1002/pits.22406

Peters, M. D., Godfrey, C. M., Khalil, H., McInerney, P., Parker, D., & Soares, C. B. (2015). Guidance for conducting systematic scoping reviews. *JBI Evidence Implementation, 13*(3), 141–146. https://doi.org/10.1097/XEB.0000000000000050.

Peters, M. D., Godfrey, C. M., McInerney, P., Munn, Z., Tricco, A.C., & Khalil, H. (2020). Chapter 11: Scoping Reviews. In: Aromataris E, Munn Z (Editors). *JBI Manual for Evidence Synthesis*, JBI, 2020. https://doi.org/10.46658/JBIMES-20-12.

Peters, M. D., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., & Khalil, H. (2021). Updated methodological guidance for the conduct of scoping reviews. *JBI evidence implementation, 19*(1), 3–10. doi/10.11124/JBIES-20-00167.

Pikulski, J.J., & Chard, D.J. (2005). Fluency: Bridge between decoding comprehension. *The Reading Teacher, 58*(6), 510–519. https://doi.org/10.1598/RT.58.6.2.

Rasinski, T. V., Blachowicz, C., & Lems, K. (2006). *Fluency instruction: Research-based best practices*. New York, NY: Guilford.

Reutzel, D. R., & Hollingsworth, P M. (1993). Effects of fluency training on second graders reading comprehension. *Journal of Educational Research, 86*, 325–331. https://doi.org/10.1080/00220671.1993.9941225.

Reynolds, C. R., & Livingston, R. A. (2021). *Mastering modern psychological testing*. Springer International Publishing.

Schaars, M. M. H., Segers, E., & Verhoeven, L. (2019). Cognitive and linguistic precursors of early first and second language reading development. *Learning and Individual Differences, 72*, 1–14. https://doi.org/10.1016/j.lindif.2019.03.008.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence, 3*(2), 125–133 (2021). https://doi.org/10.1038/s42256-020-00287-7.

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: the perils of overreliance on an" outlier" orthography. *Psychological bulletin, 134*(4), 584–615. doi/10.1037/0033-2909.134.4.584.

Shinn, M. R. (1998). *Advanced Applications of Curriculum-Based Measurement*. Guilford, New York.

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., & De Wijs, A. (2018a). *Wetenschappelijke verantwoording AVI*. Cito: Arnhem.

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., Vloedgraven, J. & De Wijs, A. (2018b). *Wetenschappelijke verantwoording DMT.* Cito: Arnhem.

Torgesen, J. K., Wagner, R., & Rashotte, C. (1997). *Test of word reading efficiency*. Austin, TX: PRO-ED.

Tricco A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of internal medicine, 169*(7), 467–473. https://doi.org/10.7326/M18-0850

University of Oregon (2020). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Administration and Scoring Guide.* Eugene, OR: University of Oregon. Available from: https://dibels.uoregon.edu

Veenendaal, N.J., Groen, M.A., & Verhoeven, L. (2016). Bidirectional Relations between Text Reading Prosody and Reading Comprehension in the Upper Primary School Grades: A Longitudinal Perspective. *Scientific Studies of Reading. 20*(3), 189–202. https://doi.org/10.1080/10888438.2015.1128939

van der Velde, M., Molenaar, B., Veldkamp, B. P., Feskens, R.C.W., & Keuning, J. (2022). The Holistic and fragmented assessment of oral reading fluency in children attending early primary education: A scoping review protocol. OSF. https://doi.org/10.17605/OSF.IO/7B6WT.

Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology, 22*(3), 407–423. https://doi.org/10.1002/acp.1414

Verhoeven, L. & Van Leeuwe, J. (2009). Modeling the Growth of Word-Decoding Skills: Evidence From Dutch. *Scientific Studies of Reading, 13*(3), 205–223. https://doi.org/10.1080/10888430902851356.

Verhoeven, L., & Keuning, J. (2018). The nature of developmental dyslexia in a transparent orthography. *Scientific Studies of Reading, 22*(1), 7–23. https://doi.org/10.1080/10888438.2017.1317780

Verhoeven, L. T. W., Voeten, M. J. M., & Keuning, J. (2022). *Modeling developmental changes in print tuning in a transparent alphabetic orthography*. https://doi.org/10.3389/fnins.2022.934590

Viana, F. L., Ribeiro, I., Vale, A. P., Chaves-Sousa, S., Santos, S. C. S., & Cadime, I. M. D. (2013). *TLP-Teste de Leitura de Palavras: manual técnico.* Liboa: CEGOC-TEA https://hdl.handle.net/1822/51181

Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research, 39*(1), 93–112. https://journals.sagepub.com/doi/pdf/10.1177/0739456X17723971

Zutell, J., & Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory Into Practice, 30*(3), 211–217. https://doi.org/10.1080/00405849109543502

Chapter 3

The framework and development of SERDA: Speech enabled reading fluency assessment for dutch

Van der Velde, M., Veldkamp, B. P., Keuning, J., Feskens, R. C. W., Swart, N. M., Harmsen, W. N. (2024). The framework and development of SERDA: Speech enabled reading fluency assessment for Dutch. In Ranđelović B., Karalić E., Aleksić K., Đukić D. (Eds.), *E-testing and computer-based assessment. CIDREE Yearbook 2024* (pp. 99– 123). CIDREE. https://www.cidree.org/wp-content/uploads/2024/11/cidree_yearbook- 2024.pdf

ABSTRACT

The importance of reading for educational, vocational and societal life cannot be understated. Nonetheless, recent large-scale studies reveal that the reading comprehension of students has declined globally, and specifically in the Netherlands. Developing fluent reading skills allows children to read quickly, accurately and with proper expression, which is fundamental to become a good reader. To monitor this development, teachers need to assess fluency on a regular basis. However, fluency assessment is currently time-consuming for teachers, provides limited information, and neglects prosody assessment. This chapter presents a framework for, and the development of, a digital automatic fluency assessment tool for early primary education that overcomes current issues through incorporating Automatic Speech Recognition (ASR): the Speech Enabled Reading Diagnostics App (SERDA). Three word- and passage reading tasks were developed based on popular pen-and-paper instruments, and administered to 653 primary school children. The results provide usability, validity and reliability evidence for SERDA's speed and accuracy measures. Furthermore, SERDA reduces the testing burden placed on teachers, increases the information gained, and facilitates prosody assessment.

3.1 INTRODUCTION

Being a proficient reader is essential to succeed throughout educational, vocational, and societal life (Horning, 2007). Nonetheless, recent large-scale studies reveal that the reading ability of fourth grade students has been on the decline globally (Mullis et al., 2023), and especially in the Netherlands (Swart et al., 2023). In addition, while less than a fourth of Dutch fifteen year-olds were found to be at risk of functional illiteracy in 2018 (Gubbels et al., 2019), recent research shows this now concerns every third (Meelissen et al., 2023). A means to counteract this trend is found through improving the development of fluent reading skills, a widely acknowledged and critical component for the development of proficient reading (National Institute of Child Health and Human Development, 2000). Given that the monitoring of this development requires teachers to assess the fluency of children's reading at a regular basis, and given that current assessment practices know many shortcomings, improving the assessment of reading fluency could help impede the imminent increase in functional illiteracy.

Reading fluency is defined as the ability to read quickly, accurately and with proper expression (Kuhn et al., 2010; Pikulski & Chard, 2005). The speed and accuracy of reading are often referred to as the automaticity of reading (e.g. Kim et al., 2021). This conceptualization dates back to the rationale discussed by Logan (1988), who argues that once a person has had sufficient practice, allowing them to read both quickly and accurately, reading becomes automatic. Here, automaticity indicates that reading requires little effort, which frees up cognitive resources, and allows the reader to focus on more complex aspects of reading, such as comprehending (Aldhanhani & Abu-Ayyash, 2020; Morris & Perney, 2018). The remaining component, expressiveness or prosody, is described by the literature as the ability to properly use a combination of phrasing, expression, intonation, stress, pitch, and pauses (van der Velde et al., 2024). The ability to read expressively has previously been linked to both earlier and later reading comprehension, the directionality of the relationship being dependent on children's primary school Grade (Veenendaal et al., 2016). To summarize, reading fluency can be seen as the degree of automaticity, or speed and accuracy, and prosody of reading.

While the construct of reading fluency is generally agreed upon, its assessment has proven problematic. Fluency assessment currently focusses on the number of words read correctly per minute (WCPM), which is an operationalization of automaticity rather than fluency (Benjamin et al., 2013; van der Velde et al., 2024). This focus on WCPM is found for popular international instruments such as the Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DORF; University of Oregon, 2020), as well as for popular instruments in the Netherlands for both word reading (i.e., the Three Minute Task [Drie-minuten-toets; DMT] (van Til et al. 2018a)) and passage reading (i.e., AVI [Analyse van Individualiseringsvormen; AVI] (van Til et al. 2018b)). Though this underrepresentation

of prosody assessment is nothing new (Paige et al., 2017), it persistently influences the validity and viability of using fluency scores in practice.

That is not to say that the overrepresentation of automaticity assessment is incomprehensible from a practical point of view. Automaticity assessment can generally be conducted both swiftly and easily (e.g. University of Oregon, 2020). Meanwhile, prosody assessment tends to be more complicated, as it requires further training, demands the administration of a separate instrument, and provides relatively subjective information (Kuhn et al., 2010). Given that test administration and scoring is work that is mostly carried out by teachers, placing a heavy testing burden on them, it is not difficult to understand why incorporating prosody assessment is often deemed too time-consuming. Moreover, even when only taking into account automaticity assessment, the extraction of detailed diagnostics tends to be limited in practice, as these require a more thorough and time-consuming investigation of the reading performance.

In short, the assessment of reading fluency is overrepresented by its speed and accuracy components. In addition, fluency assessment currently places a large testing burden on teachers and does not yield detailed diagnostics when conducted in a practically feasible manner. Therefore, creating an assessment tool that could limit teacher burden while providing detailed and objective fluency diagnostics on all fluency components could considerably help teachers, children and society at large. In this chapter, we describe the proposed framework to overcome these assessment shortcomings, which will subsequently be implemented within the development of a reading fluency assessment instrument referred to as the Speech Enabled Reading Diagnostics App (SERDA).

3.1.1 SERDA's framework

SERDA's framework describes how to improve reading education at the primary school level by means of assessing reading fluency through the analysis of speech from reading aloud tasks, and by means of modelling the resulting data to provide individualized feedback on how to improve reading. Specifically, the final goal is for SERDA to visualize the reading ability of children for teachers at both the class and individual level. Information should be presented on children's general ability to read fluently, as well as more specific information on the speed, accuracy and expressiveness of reading. In addition, SERDA should be able to differentiate between children's performance on the reading of word lists and passages, providing a comparison of proficiency in context-free and context-specific reading.

In order to manifest these ambitions, SERDA's framework combines automatic speech recognition (ASR), speech diagnostics and learning analytics to create an innovative, integrated approach to reading diagnostics and automated feedback, as illustrated in Figure 1. Throughout this framework, ASR concerns the "independent, machine-based process of decoding and transcribing oral speech" (Levis & Suvorov, 2012, p. 1). Speech diagnostics refer to the relevant speed, accuracy and expressiveness measures extracted from speech

data by the ASR-algorithm. Learning analytics is generally defined as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” (Society for Learning Analytics, 2011). Within the context of this framework, learning analytics specifically relate to the analyses conducted to transform speech diagnostics into recommendations that can be used to improve personal learning-to-read trajectories in primary education. In essence, when compared to its pen-and-paper contemporaries, SERDA’s most fundamental difference is its usage of speech data and the transformation of speech into relevant diagnostics through ASR.

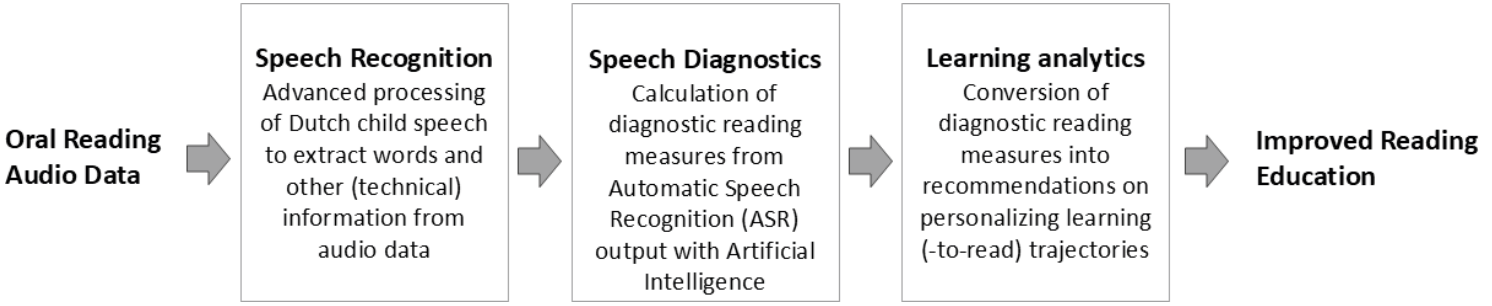


Figure 1. *SERDA's Reading Fluency Assessment Framework*

3.1.2 Automatic reading fluency assessment

Research on implementing ASR with the goal of improving reading ability dates back more than two decades (Mostow et al., 2003; Reeder et al., 2007). This incorporation was fruitful, as evidenced by the wealth of successful research on ASR-based reading tutors for reading practice and automatic assessment in English (Bolaños et al., 2013; Loukina et al., 2019; Sabu & Rao, 2018) and other languages (Godde et al., 2017; Proença et al, 2015; Silva et al., 2021).

Within the context of the Dutch language, Bai et al (2020) have recently shown that ASR can be successfully used to assess and provide feedback on the reading accuracy and speed of first graders. In addition, Wei et al (2022) showed the potential of ASR in assessing reading errors for non-native speech. Indeed, up to now, much research on the use of ASR for the speech of Dutch children has focussed on extracting reading errors (Nicolao et al., 2018; Yilmaz et al., 2014), while much less work has focussed on prosody (e.g. Cucchiarini et al., 2000). However, previous research has shown that it is possible to extract automatic measures from speech that are related to subjective fluency and prosody ratings (Benjamin et al., 2013; Cheng, 2011; Chung & Bidelman, 2022; Dimzon & Pascual, 2023; Truong et al., 2018).

3.1.3 The present study

Reading fluency assessment in the Netherlands currently places too large a testing burden on teachers and does not yield detailed diagnostics when conducted in a practically feasible manner. The current study presents a framework to overcome these shortcomings. Based on this framework a reading fluency assessment instrument was developed. Throughout the remainder of this chapter, we will discuss the development of this instrument and the collected speech data. In addition, we provide usability, reliability and validity evidence to substantiate the use of SERDA's speed and accuracy measures in practice. The extraction and evaluation of SERDA's prosody measures is discussed in another paper, as these require different algorithms and methodology, which reaches beyond the scope of this chapter.

3.2 METHODS

The development of SERDA followed the following steps: First, we constructed reading tasks in collaboration with subject-area experts based on currently popular fluency instruments. Then, we administered the reading tasks to 653 children attending Grade 2 and 3 of primary schools in the Netherlands. In addition, to evaluate the validity of SERDA's tasks we obtained the most recent results of the DMT and AVI, which are the standard Dutch fluency instruments used throughout primary education.

3.2.1 Task-development

3.2.1.1 *SERDA: Word reading.*

In order to assess the ability of children to read context-free words, three Dutch word lists were developed based on the DMT (Van Til, 2018a). Each list contained 50 words. The first set consisted of one-syllable words with varying consonant-vowel (cv) combinations, cv/vc/cvc/ccv/ccvc/vcc/cvcc/ccvcc, and included various reading difficulties (i.e., *sch-*, *-ng*/*nk*, open syllable). The second set consisted of one-, two- and three-syllable words, including various advanced reading difficulties (i.e., *be-*/*ge-*/*ver-*, *-lijk*). The third set consisted of two-, three-, and four-syllable words, including various complex reading difficulties (i.e., loanwords, *-isch, -x-*, *-y-*). Words were chosen by experts in the field, based on a Dutch reading fluency test (Keuning & Verhoeven, 2005).

In order to obtain accurate word reading speed estimates, a progressive demasking design was used (Grainger & Sugui, 1990). During the progressive demasking task, a word was individually presented in the middle of the screen with a mask placed over the word, resulting in a seemingly empty screen. Then, the mask was removed for 17 milliseconds (ms). This left children with 17 ms to read the word, after which the mask returned and the first cycle was completed. The removal time of the mask gradually increased to 340 ms, in steps of 17ms per cycle. Children were instructed to tap the screen as soon as they recognized the word, after which they read the word out loud. When children were not able to read the word with a presentation time of 2,200 ms, corresponding to the 20th cycle, the child moved on towards the next word. An example of the masking-design is visualized in Figure 2.

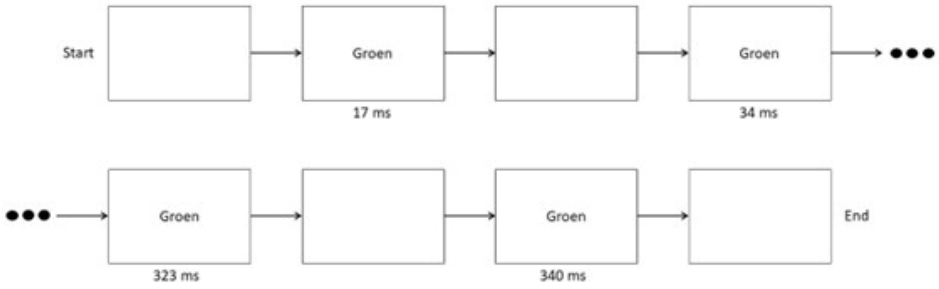


Figure 2. *Example of the Start and end of a Single Progressive Demasking Trial Using the Word "Green" [Groen].*

3.2.1.2 *SERDA: Passage reading.*

In order to assess the ability of children to read context-specific passages, three short passages of increasing difficulty were created based on guidelines that were also used in the construction of the passages of the AVI (Van Til, 2018b). Passages contained around 175

words, were written by professional child literature authors, and were selected by experts in the field. Passage-content was selected with the aim to match the interests of young children. The passages were created to, respectively, match characteristics representative of the expected reading level at the end of Grade 2, the middle of Grade 3 and the end of Grade 3. The texts contained monosyllabic, bisyllabic and polysyllabic words with incidental orthographic inconsistencies and complexities (see Van Til et al., 2018b).

3.2.2 Participants

Data was collected from October 2022 to February 2023, and from October 2023 to February 2024. To establish a generalizable sample with respect to important background variables, the distribution of school-weight (Inspectie van het Onderwijs, 2024), which is an indication of children's socio-economic background, and the distribution over dialectical regions in the Netherlands (Cucchiarini et al., 2008), were used to draw two samples over all primary schools in the Netherlands. Schools for special education and special needs education were excluded from the sample. The first sample consisted of 20 school, of which 7 (35%) participated. The second sample consisted of 30 school, of which 12 (40%) participated. For participating schools, all available Grade 2 and 3 classes were included in the sample. Ethical approval was obtained from the local ethics committee, after which the approval for the participation of individual children was acquired from the children's parents or caregivers (from here: parents). The participating children were all asked to complete all reading tasks.

In total, 19 schools participated, providing a sample of 653 children. Out of all participating students 48% (n = 311) attended Grade 2, 47% (n = 310) attended Grade 3 and 5% (n = 32) attended a combined Grade 2/3 class. Boys made up 48% (n = 312) of the sample. On average, children were seven years of age when attending Grade 2 and eight years and two months of age when attending Grade 3.

The sample showed a representative distribution of schools in the Netherlands. However, it has to be remarked that the sample overrepresented schools with a low school-weight and underrepresented schools from the west of the Netherlands to some degree. With regard to school-weight, this means that relatively well-performing schools with less complex or diverse populations were more willing to participate within the current study. This finding is unsurprising, given that data collection was started a few months after the final COVID-19 lockdown was concluded. As a result, most schools in the Netherlands, and especially those with a complex or diverse population of children, were exceptionally busy and less receptive to participate in research. As for dialect region, although schools from the west of the Netherlands were underrepresented compared to the population, this resulted in a sample that provided a more equally distributed representation of schools from all dialect regions. Given the intended use of SERDA throughout all of the Netherlands, speech from children with all types of accents and dialects should be equally eligible. As a result, this underrepresentation is deemed unproblematic.

3.2.3 Materials

3.2.3.1 *SERDA: Automatic measures of word and passage reading.*

For each reading task, SERDA yields an audio recording and a log file. The audio recordings contain the recorded speech of the child during the task. The log file contains metadata, information about the children's interactions with the application, and the total duration of the tasks. Using the audio and log data, SERDA generated item-, task- and person-level accuracy and speed measures for the word- and passage reading tasks, as well as task- and person-level WCPM-scores. An overview of all extracted metrics is presented in Table 1.

Table 1. Item, Task and Passage Level Measures Extracted by SERDA.

Measure	Word-reading	Passage-reading
Item level		
Accuracy	0 or 1	0 or 1
Speed	Flashing time (seconds)	Speaking duration (seconds)
WCPM	-	-
Task level		
Accuracy	Number of words read correctly	Number of words read correctly
Speed	Words read divided by total flashing time	Words read divided by task duration
WCPM	Accuracy divided by total flashing time	Accuracy divided by task duration
Person level		
Accuracy	Average task-level accuracy	Average task-level accuracy
Speed	Average task-level speed	Average task-level speed
WCPM	Average task-level WCPM	Average task-level WCPM

SERDA automatically computed children's word- and passage reading accuracy based on the audio recordings. The transformation from audio to accuracy followed three steps, which are visualized in Figure 3. Firstly, each audio recording was automatically transcribed using the Dutch Large-v2 ASR model Whisper Timestamped (Louradour, 2023). This model is based on OpenAI's Whisper ASR models (Radford et al., 2023) and uses Dynamic Time Warping (DTW; Giorgino, 2009) to predict word segment timestamps. Secondly, prompts (i.e., the text of the original word list or passage) were aligned with the ASR transcription of the audio using the reversed-ADAPT algorithm for grapheme alignment (Bai et al., 2021; Elffers et al., 2013). Thirdly, each prompt word that is aligned with exactly that word from the ASR transcription is labeled as read correctly (1), while all others were labeled as read incorrectly (0), resulting in item-level accuracy scores. Followingly, the

number of words read correctly were calculated for each word- and passage reading task to obtain task-level accuracy scores. Finnaly, averaging over task-level measures resulted in person-level accuracy scores for the word- and passage reading tasks.

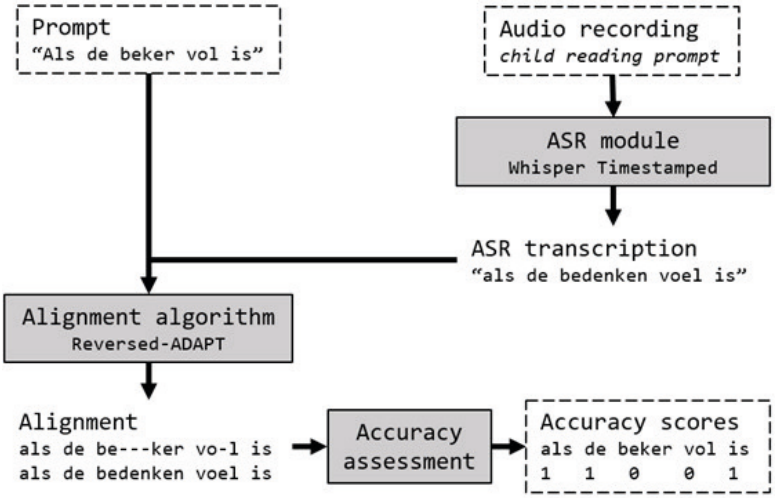


Figure 3. *Visualization for the Transformation of Audio Into Accuracy Scores*

The calculation of reading speed differed between the word- and passage reading task. For the word-reading task logged timestamps were used. Specifically, for each word in each word-reading subtask, SERDA stored three timestamps. The first was the onset of word presentation (T1), while the second timestamp concerned the first tap on the screen, indicating that the child has recognized the word (T2), and the final timestamp was the second tap on the screen, which signaled that the child has read the word out loud and wants to continue to the next word (T3). Then, the duration between T1 and T2, referred to as the flashing time, was calculated and functioned as the item-level speed measure. Next, the number of words read per minute (WPM) was calculated for each word-reading task by dividing the total number of words read by the total flashing time, resulting in task-level speed measures. Calculating the average WPM over all word-reading tasks resulted in person-level speed measures for the word reading task.

For the passage reading task, SERDA used the ASR output to obtain item-level speed measures. This output contains the word segments automatically recognized in the audio, together with their start and end timestamp. From the alignment of the recognized word segments with the prompt, we extract the begin and end timestamps, and thus duration, of the correctly read words. Subsequently, the task-level speed measures (WPM) were obtained by dividing the total number of words read by the total task duration. The person-level speed measures were defined as the average of the task-level speed measures.

Lastly, WCPM-scores were calculated for each word- and passage reading task. For the word-reading task, WCPM was defined as the number of words read correctly, divided by the total flashing time. For the passage reading task, WCPM concerned the number of words read correctly, divided by the time it took to complete the passage. Then, average WCPM-scores were calculated over all administered tasks, which were used as person-level measures for the word- and passage reading task.

3.2.3.2 *Word and passage decoding: the three minute task [Drie-minuten-toets; DMT] and AVI [Analyse van Individualiseringsvormen; AVI].*

The DMT and AVI are paper based tests, intended to assess the development of the word- and passage reading ability of children attending primary education in the Netherlands (van Til et al., 2018a, 2018b). Their usability, reliability and validity have been thoroughly investigated, and positively evaluated by the Dutch Committee on Tests and Testing (COTAN; Egberink & Leng, 2024a, 2024b). During the DMT and AVI, children respectively read up to three word lists or passages of increasing difficulty. Children were explicitly instructed to read the words and passages out loud as quickly and accurately as possible.

For the DMT, children had one minute to read each individual word list. The first list only contained single-syllable words with one consonant at a time. The second list also contained words with two syllables, and included words with multiple consonants. Finally, the third list allowed for the inclusion of words of more than two syllables. For the AVI, the child kept on reading increasingly difficult Grade-level passages, until the child made too many mistakes, started reading too slow, or showed a combination of both.

Then, the children's DMT and AVI classifications were obtained, which are based on their performance compared to Grade-specific Dutch norms. For the DMT, classifications concerned placing children into one of five categories, ranging from the 20% lowest- to the 20% best performing children. For the AVI, children were classified into categories that correspond to, and represent, the Grades of primary education in the Netherlands.

3.2.4 Procedure

SERDA's tasks were individually administered by a test-leader in a quiet room at the children's schools, without any additional personnel. Participants conducted the tasks on a Samsung Galaxy Tab A6 tablet, while their speech was captured using a headset with an in-build microphone. To mimic current assessment practices, administration was conducted during school-hours. With the exception of school- specific break timings and opening hours, no variation existed with regard to the timing of assessment between schools. In order to evaluate the children's experience with SERDA, as well as practical and technical issues, test-leaders recorded noteworthy situations and comments from children.

During the administration of SERDA's reading tasks, children were first introduced to the word reading task. Due to the novel nature of the task, children were first taught

how to correctly conduct the exercise by completing three examples with the test-leader. Once children felt comfortable with the word-reading task they completed the first word list, which contained the easiest words. After completing the first word-list the second and third word-reading tasks were conducted.

The passage-reading task was very recognizable to the children, as most have had experience with the AVI. Therefore, instructions were limited. For each passage, students were instructed to read the passage out loud, including the title, as quickly and accurately as possible. A passage was completed once the child read the entire passage, or after three minutes had passed. After the instructions, the child was allowed to start on the first passage, followed by the second and third passage.

The AVI and DMT measures were provided by the schools of the children, all of whom were familiar with conducting and scoring the AVI and DMT.

3.2.5 Data analysis

Data analysis was primarily aimed at obtaining an indication of the usability, reliability and validity of SERDA's reading tasks. The usability of SERDA's reading tasks was investigated by comparing their average administration duration to those of the DMT and AVI. For SERDA's word-and passage reading tasks, this concerns the duration of administering all subtasks. For the DMT and AVI, we took the sum of the administration and scoring duration, as reported by COTAN (Egberink & Leng, 2024a, 2024b). Scoring-time was excluded for SERDA's reading tasks, as SERDA performs this task automatically.

To evaluate the reliability of SERDA's reading tasks we evaluated the internal consistency and split-half reliability. The internal consistency was evaluated by calculating Cronbach's Alpha (Cronbach, 1951) for the accuracy-scores of all items in the word- and passage reading tasks, as well as for their separate subtasks. Split-half reliability was estimated for the item-level accuracy, speed and WCPM measures of the word- and passage reading tasks by correlating 10.000 randomly generated 50/50 splits.

Then, to evaluate the validity of SERDA's reading tasks, we investigated their construct validity. Construct validity, which concerns the degree to which a tool measures the construct it aims to measure (American Educational Research Association [AERA] et al., 2014; Reynolds & Livingston, 2021), was determined by correlating the WCPM-scores of SERDA's word- and passage reading tasks with one-another. In addition, we used Spearman's Rho to compare SERDA's WCPM scores to the DMT and AVI classifications. All analyses were conducted in RStudio (version 4.3.1; Posit Team, 2023).

Finally, the feedback of test leaders was evaluated. We investigated whether comments were made with regard to experiences of children with SERDA. In addition, we evaluated and tried to remedy issues that emerged.

3.3 RESULTS

3.3.1 Descriptives

SERDA was administered to 653 children, resulting in 176.6 hours of speech data. In addition, 569 and 622 classifications were obtained for the DMT and AVI respectively.

The usability of SERDA's tasks was evaluated by comparing their administration duration to the administration duration of the DMT and AVI. The administration of SERDA's word reading task took about 10 minutes on average, while the average duration of the passage reading tasks was 6 minutes. Administering both reading tasks generally took between 10 to 25 minutes, with an average duration of about 16 minutes. For the DMT and AVI, the COTAN reported that the combined duration of administration, scoring and interpretation are, respectively, 20 and 25 minutes. Thus, the total duration of administering SERDA's word and passage reading tasks is feasibly lower compared to the DMT and AVI, especially when both types of reading need to be administered, scored and interpreted.

During the word reading task children, on average, read 31 (*SD* = 8.6) words correctly, 24 (*SD* = 7.2*)* words per minute and 16 (*SD* = 7.4) words correctly per minute. For the passage reading task, children averaged 84 (*SD* = 19.6) words read correctly, 98 (*SD* = 34.7*)* words per minute, and 48 (*SD* = 20.1) words read correctly per minute. Figure 4 presents the distribution of administration duration, as well as the average accuracy, speed and WCPM measures for the word reading task. Likewise, Figure 5 presents the distribution of administration duration, as well as the average accuracy, speed and WCPM measures for the passage reading task.

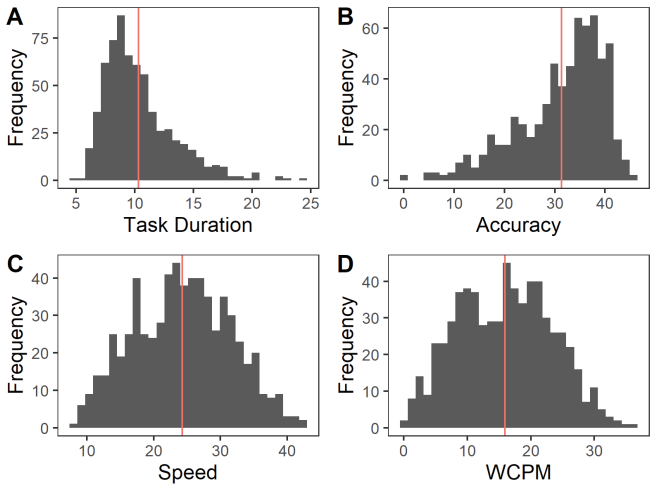


Figure 4. *Average Task Duration (A), Accuracy (B), Speed (C) and WCPM (D) Metrics for the Word-reading Task, With Sample Averages Indicated by the Vertical Lines*

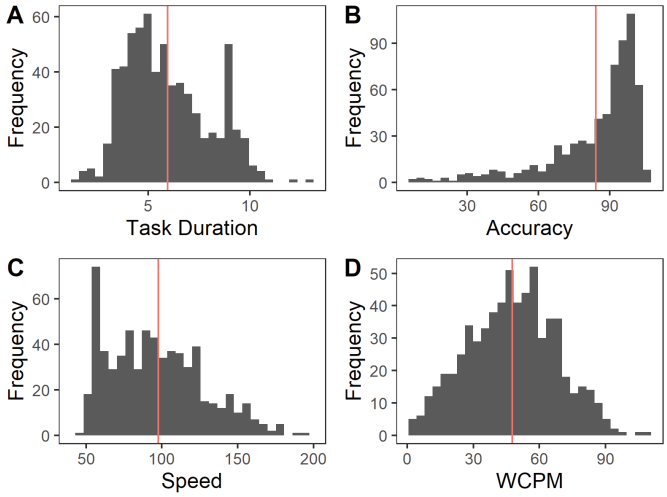


Figure 5. *Average Task Duration (A), Accuracy (B), Speed (C) and WCPM (D) Metrics for the Passage-reading Task, With Sample Averages Indicated by the Vertical Lines*

3.3.2 Internal consistency

Table 2 shows Cronbach's Alpha for the accuracy scores of the word- and passage reading tasks, as well as their subtasks. Cronbach Alpha ranged from 0.89 to 0.97 for the subtasks, and between 0.96 to 0.98 for the complete tasks. This indicates that SERDA's reading tasks, when administered in their entirety, provide good internal consistency when used to make important individual decisions according to the COTAN guidelines (Evers et al., 2009).

Table 2. Cronbach's alpha for the accuracy scores of the word- and passage reading tasks, as well as their subtasks

Task	Word-reading	Passage-reading
Complete	0.96 [0.96, 0.97]	0.98 [0.98, 0.99]
Sub-task 1	0.94 [0.93, 0.95]	0.97 [0.96, 0.97]
Sub-task 2	0.89 [0.87, 0.91]	0.96 [0.96, 0.97]
Sub-task 3	0.94 [0.93, 0.95]	0.96 [0.96, 0.97]

Note. Bracketed numbers are 95% confidence intervals.

3.3.3 Split-half reliability

Table 3 shows the split-half reliability estimates of SERDA's word- and passage reading tasks, averaged over 10.000 random 50/50 splits. Average split-half reliability estimates ranged from 0.92 to 0.99, showing good reliability for the word- and passage reading task, when used for important individual decisions.

Table 3. Split-half reliability for the speed, accuracy, and wcpm measures of the complete word- and passage reading tasks, averaged over 10000 randomly selected splits

Task-type	Words	Passages
Speed	0.99 (SD < 0.01)	0.93 (SD = 0.01)
Accuracy	0.93 (SD = 0.01)	0.97 (SD = 0.01)
WCPM	0.97 (SD < 0.01)	0.92 (SD = 0.01)

Note. SD = standard deviation over 10000 split-half reliability estimates.

3.3.4 Construct validity

To evaluate the construct validity of SERDA's reading tasks, we conducted a Pearson correlation between the WCPM-scores of the word- and passage reading tasks. In addition, we calculated spearman's rho between the WCPM-scores of the word reading task and the DMT classifications, as well as between the WCPM-scores of the passage reading task and the AVI classifications.

As shown in table 4, correlations varied between 0.54 to 0.81, showing moderate to strong positive relationships (Schober et al., 2018). The Pearson correlation analysis showed a significant positive relationship between the WCPM scores of SERDA's word and passage reading task, $r(644) = .79$, $p < .001$. The spearman correlation between the WCPM scores of the word reading task and the classification of the DMT showed a significant positive relationship, $rho(502) = 0.54$, $p < .001$. The spearman correlation between the WCPM scores of the passage reading task and the classification of the AVI showed a significant positive relationship, $rho(591) = 0.81$, $p < .001$.

Table 4. Correlations between the WCPM-scores of SERDA's word- and passage reading tasks, and the performance categories of the DMT and AVI

Task	WCPM Words	WCPM Passages	CLASS DMT	CLASS AVI
WCPM Words	1	-	-	-
WCPM Passages	0.79	1	-	-
Class DMT	0.54	0.68	1	-
Class AVI	0.68	0.81	0.65	1

Note. All correlations were significant at a = 0.001

3.3.5 Experiences with SERDA

Test leaders noted that children tended to enjoy SERDA's reading tasks, as exemplified through comments such as "I wasn't sure whether I should let her do the third set of words, because she hardly said anything right, but she enjoyed doing it 0so much that I thought it was fine", and " [I] experienced the task as very enjoyable". At the same time, some deemed

the number of tasks to numerous: “She thought 3 word lists was a lot.”, while others got tired: “had too little energy + concentration to finish the last story”.

Regarding testing-issues, we observed some technical, practical and task-related problems. The test-leaders noted that children incidentally skipped tasks, while the application failed to store the recordings. Though these issues were mostly resolved by the end of the first round of data collection, this has led to some data loss. In addition, comments were made regarding disturbances on-site. For example, one test leader remarked: “There was a parent arguing in the hallway, which was quite distracting”, while another noted that “the class next door was singing”. Though these disturbances are problematic, as they reduce the quality of the audio recordings, they are deemed characteristic of primary schools. Finally, two issues were noted regarding children’s performance on the word reading task. Namely, children “clicked too early”, leaving them unable to recognise and read out the word, or read words before tapping the screen, leading to larger flashing times.

3.4 DISCUSSION

The current study aimed to improve reading fluency assessment by developing a novel digital reading fluency assessment instrument that utilizes ASR. Specifically, the goal of SERDA is to reduce the testing burden placed on teachers, while increasing the amount of available fluency diagnostics. Throughout this chapter we have investigated the usability, reliability and validity of SERDA’s speed, accuracy and automaticity measures.

The results of the current study illustrate some advantages of SERDA compared to the DMT and AVI, based on its mode of administration. First of all, SERDA’s administration time is relatively short. On average, administering both reading tasks takes about 16 minutes, whereas the combined administration time for the DMT and AVI comes down to around 20 minutes (COTAN; Egberink & Leng, 2024a, 2024b). In addition, SERDA only requires teachers to provide instructions and a microphone, while the DMT and AVI require test administration, scoring and interpretation to be done by hand. In practice, this can lead to total administration durations of up to 45 minutes. All the while, SERDA’s usage of speech data allows for a more elaborate investigation of children’s strengths and difficulties, as speed and accuracy information can easily and quickly be obtained at the item, task, and person level. Indeed, as long as tasks are conducted correctly and the speech of the child is properly captured, SERDA can be administered more quickly, reducing teacher’s testing burden while providing detailed information on the speed and accuracy of children’s reading.

Furthermore, the results of the current study indicate that both the word- and passage reading task provide reliable scores that resemble their pen-and-paper contemporaries moderately well to good. It is important to note, however, that the validity of the word-reading task was lower than the validity of the passage-reading task. This is not surprising,

given that SERDA's word-reading task used a progressive demasking design. This design creates an administrative discrepancy with the DMT that reaches beyond the difference between the passage-reading task and the AVI, which primarily reflect their administrative modes. At the same time, the correlation between SERDA's word-reading task and the AVI was almost identical to the correlation between the DMT and AVI, indicating that SERDA's word-reading scores do not resemble the AVI's conceptually worse than the DMT's.

Though these results are promising, SERDA still requires work before it can solve the issues that currently plague fluency assessment. First of all, future research should be conducted to transform SERDA's provision of individual performance information into relevant diagnostics, as well as their translation into feedback towards teachers. An important step towards this goal can be made by incorporating the types of mistakes that children make at the item, task and person level. These mistakes and successes could then be used to discover reading profiles, based on state of the art learning-to-read models such as the DRC (Coltheart et al., 2001), Triangle (Harm & Seidenberg, 2004) and Connectionist Dual Process model (Perry et al., 2010). In turn, these reading profiles could be used to provide teachers with insightful individualized suggestions with regard to children's learning-to-read trajectories, an approach that has successfully been applied in the Netherlands within the context of reading comprehension (Keuning et al., 2019).

When more elaborate fluency diagnostics are incorporated, it is advised to undertake a more thorough validation of SERDA's reading tasks. Though the present study provided some evidence for the usability, reliability and validity of SERDA's reading tasks, more evidence is required if statements are to be made about the use of SERDA's speed, accuracy and WCPM scores in practice. Within this validation, extensive efforts should be made to evaluate the validity of the ASR-algorithm used throughout the current study. In addition, a more thorough validation could provide insights with regard to the reliability of score-estimation for children of different reading abilities, as well as a more elaborate evaluation of the quality and informativeness of individual items.

When a more elaborate validation of SERDA's improved measures has been established, researchers are advised to focus on the generation, evaluation and validation of prosody measures. Although some work has been conducted on extracting prosody measures using ASR (e,g. Truong et al., 2018), little research has been done to evaluate their usefulness and informativeness within a Dutch context, let alone for Dutch primary school children. Therefore, an elaborate investigation of the usefulness and validity of extracting established prosody measures from Dutch primary school children's speech should be conducted.

Finally, the comments made by the test-leaders require consideration. While children enjoyed the reading tasks, some found them too lengthy, reducing their ability to concentrate. However, the complete set of tasks was primarily administered for the purpose of the task evaluation. For applications of SERDA in practice, a subset of tasks could suffice, as substantiated by the split-half reliability results. In addition, comments

were made regarding the tapping behavior of children, leading to errors at the item-level. These errors could be attributed to the novelty of the task. Therefore, we suggest a more thorough practice exercise to reduce these procedural errors.

To sum up, the current study was conducted with the goal of creating a reading fluency assessment tool that overcomes current assessment shortcoming. To achieve this goal, digital word- and passage reading tasks were developed based on expert opinion and currently popular reading fluency assessment instruments in the Netherlands. The results of this study suggest that SERDA's reading tasks provide reliable and valid indications of children's reading speed and accuracy, while reducing teacher's testing burden. Future researchers are advised to build upon the work conducted here by increasing the diagnostics SERDA can provide, through conducting a more comprehensive validation study on SERDA's reading tasks, and through the extraction of prosody features from the available speech data. If SERDA's development is successfully completed, its tasks could help individualize reading instruction, further reducing teacher testing burden, and improve reading education, providing the means to reduce the emerging emaciation of children's literacy the world over.

REFERENCES

Aldhanhani, Z. R., & Abu-Ayyash, E. A. (2020). Theories and Research on Oral Reading Fluency: What Is Needed?. *Theory and Practice in Language Studies, 10*(4), 379–388. http://dx.doi.org/10.17507/tpls.1004.05

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for Educational & Psychological Testing.* Washington, DC: American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Bai, Y., Hubers, F., Cucchiarini, C., & Strik, H. (2020). *ASR-Based Evaluation and Feedback for Individualized Reading Practice.* Paper presented at Interspeech 2020, Shanghai, China, October 25–29 [online]. Available: https://www.researchgate.net [March 21, 2024].

Bai, Y., Hubers, F. C. W., Cucchiarini, C., & Strik, H. (2021). *An ASR-based Reading Tutor for Practicing Reading Skills in the First Grade: Improving Performance through Threshold Adjustment*. Paper presented at Iberspeech 2021, Valladolid, Spain, March 24-25 [online]. Available: https://repository.ubn.ru.nl/bitstream/handle/2066/245151/245151.pdf [March 21, 2024]

Benjamin, R. G., Schwanenflugel, P. J., Meisinger, E. B., Groff, C., Kuhn, M. R., & Steiner, L. (2013). A spectrographically grounded scale for evaluating reading expressiveness. *Reading Research Quarterly, 48*(2), 105–133. https://doi.org/10.1002/rrq.43

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013). Human and automated assessment of oral reading fluency. *Journal of Educational Psychology, 105*(4), 1142–1151. https://doi.org/10.1037/a0031479 \

Cheng, J. (2011). *Automatic assessment of prosody in high-stakes English tests*. Paper presented at Interspeech 2011, Florence, Italy, August 27-31 [online]. Available: https://caimber-cdn.s3.us-west-2.amazonaws.com/papers-and-presentations/cheng11c_interspeech.pdf [Januari 21, 2024].

Chung, W. L., & Bidelman, G. M. (2022). Acoustic Features of Oral Reading Prosody and the Relation With Reading Fluency and Reading Comprehension in Taiwanese Children. J*ournal of Speech, Language, and Hearing Research, 65*(1), 334–343. https://doi.org/10.1044/2021_JSLHR-21-00252.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–258.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cucchiarini, C., van Hamme, H., Driesen, J., Sanders, E. (2008). *THE JASMIN-CGN: CORPUS Design, recording, transcription and structure of the corpus.*

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America, 107*(2), 989–999. https://doi.org/10.1121/1.428279.

Dimzon, F. D., & Pascual, R. M. (2023). Prosodic characterisation of children's Filipino read speech for oral reading fluency assessment. *International Journal of Technology Enhanced Learning, 15*(1), 74–94. https://doi.org/10.1504/IJTEL.2023.127939.

Egberink, I.J.L. & Leng, W.E. de. (2024a). *2010, AVI [2010, AVI]* [online]. Available: www.cotandocumentatie.nl [Januari 21, 2024].

Egberink, I.J.L. & Leng, W.E. de. (2024b). *2010, Drie-Minuten-Toets* [2010, Three Minute Test] [online]. Available: www.cotandocumentatie.nl [Januari 21, 2024].

Elffers, B., van Bael, C., Strik, H. (2005). *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcription*s [online]. Available: https://hstrik.ruhosting.nl/wordpress/wp-content/uploads/2013/03/a121-adapt.pdf

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2009). *COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie)* [COTAN evaluation system for the quality of tests (completely revised version)]. NIP. Available: http://www.psynip.nl/website/wat-doet-het nip/tests/beoordelingsprocedure/beoordelingsprocedure [Januari 21, 2024].

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software, 31,* 1-24. https://doi.org/10.18637/jss.v031.i07.

Godde, E., Bailly, G., Escudero, D., Bosse, M. L., & Gillet-Perret, E. (2017). *Evaluation of reading performance of primary school children: Objective measurements vs. subjective ratings*. Presented at the 6th Workshop on Child Computer Interaction, Glasgow, Scotland, November 13 [online]. Available: https://doi.org/10.21437/WOCCI.2017-4.

Grainger, J., & Segui, J. (1990). Neighborhood frequency effects in visual word recognition: A comparison of lexical decision and masked identification latencies. *Perception & psychophysics, 47*, 191-198. https://doi.org/10.3758/BF03205983

Gubbels, J., van Langen, A., Maassen, N., & Meelissen, M. (2019). *Resultaten PISA-2018 in vogelvlucht* [Results of PISA-2018 from a bird's eye view] [online]. Enschede: Universiteit Twente. Available: https://doi.org/10.3990/1.9789036549226 [Januari 21, 2024].

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662–720. https://doi.org/10.1037/0033-295X.111.3.662.

Horning, A. S. (2007). Reading across the curriculum as the key to student success. *Across the disciplines, 4*(1), 1–17.

Inspectie van het Onderwijs. (2024). *Schoolweging primair onderwijs* [Schoolweightprimary education] [online]. Available: https://www.onderwijsinspectie.nl/trends-en-ontwikkelingen/onderwijsdata/schoolweging-po [March 20, 2024].

Keuning, J., Swart, N., Scheltinga, F., Gruhn, C.S., Segers, E. & Verhoeven, L. (2019). *Evaluatie en planning van leesleertrajecten: Een dynamisch perspectief (eindrapport NRO project 405-15-548)*[Evaluation and planning of learning-to-read trajectories, a dynamic perspective (final report NRO project 405-15-548)] [online]. Arnhem: Cito. Available: https://www.nro.nl/sites/nro/files/migrate/eindrapport-405-15-548.pdf [Januari 21, 2024].

Keuning, J. & Verhoeven, L. (2005). *Signaleren van lees- en spellingproblemen in groep 4-8* [Detection of reading- and spelling problems in Grades 2-6]. Nijmegen: Expertisecentrum Nederlands

Kim, Y. S. G., Quinn, J. M., & Petscher, Y. (2021). What is text reading fluency and is it a predictor or an outcome of reading comprehension? A longitudinal investigation. *Developmental Psychology, 57*(5), 718–732. https://doi.org/10.1037%2Fdev0001167

Kuhn, M., Schwanenflugel, P. & Meisinger, E. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 232–253. https://doi.org/10.1598/RRQ.45.2.4

Levis, J., & Suvorov, R. (2012). Automatic speech recognition. In: Chapelle, C. A. (2012). *The encyclopedia of applied linguistics.* Hoboken, NJ : John Wiley & Sons.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Loukina, A., Klebanov, B. B., Lange, P. L., Qian, Y., Gyawali, B., Madnani, N., Misra, A., Zechner, K., Wang, Z., & Sabatini, J. (2019). *Automated Estimation of Oral Reading Fluency During Summer Camp e-Book Reading with MyTurnToRead.* Paper presented at Interspeech 2019, Graz, Austria, 15-19 September [online]. Available: https://www.iscaarchive.org/interspeech_2019/loukina19_interspeech.pdf [Januari 21, 2024].

Louradour, J. (2023). *Whisper-timestamped* [online]. GitHub Repository. Available: https://github.com/linto-ai/whisper-timestamped [March 21, 2024].

Meelissen, M. R. M., Maassen, N. A. M., Gubbels, J., van Langen, A. M. L., Valk, J., Dood, C., Derks, I., In 't Zandt, M., & Wolbers, M. (2023). *Resultaten PISA-2022 in vogelvlucht* [Results of PISA-2022 from a bird's eye view] [online]. Enschede: Universiteit Twente. Available: https://doi. org/10.3990/1.9789036559461 [March 18, 2024].

Morris, D., & Perney, J. (2018). Using a sight word measure to predict reading fluency problems in grades 1 to 3. *Reading & Writing Quarterly, 34*(4), 338–348. https://doi.org/10.1080/10573569.2018.1446857

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., & Tobin, B. (2003). Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research, 29*, 61–117. https://doi.org/10.2190/06AX-QW99-EQ5G-RDCF

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 International Results in Reading* [online]. Boston College, TIMSS & PIRLS International Study Center. Available: https://doi.org/10.6017/lse.tpisc.tr2103.kb5342 [Januari 21, 2024].

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washington, DC: U.S. Government Printing Office.

Nicolao, M., Sanders, M., & Hain, T. (2018). I*mproved acoustic modelling for automatic literacy assessment of children.* Paper presented at Interspeech 2018, Hyderabad, India, 2-6 September [online]. Available: https://doi.org/10.21437/Interspeech.2018-2118 [Januari 21, 2024].

Paige, D. D., Rupley, W. H., Smith, G. S., Rasinski, T. V., Nichols, W., & Magpuri-Lavell, T. (2017). Is prosodic reading a strategy for comprehension?. *Journal for educational research online, 9(*2), 245–275. https://doi.org/10.25656/01:14951.

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology, 61*, 106–151. https://doi.org/10.1016/j.cogpsych.2010.04.001

Pikulski, J.J., & Chard, D.J. (2005). Fluency: Bridge between decoding comprehension. *The Reading Teacher, 58*(6), 510–519. https://doi.org/10.1598/RT.58.6.2

Posit team (2023). RStudio: Integrated Development Environment for R, version 4.3.1. Posit Software, PBC, Boston, MA. http://www.posit.co/.

Proença, J., Celorico, D., Candeias, S., Lopes, C., & Perdigão, F. (2015). *Children's Reading Aloud Performance: A Database and Automatic Detection of Disfluencies.* Presented at Interspeech 2015, Dresden, Germany, 6-10 September [online].

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision.* Paper presented at the 40th International Conference on Machine Learning, Hawaii, USA, 23-29 July [online]. Available: https://proceedings.mlr.press/v202/radford23a.html [March 21, 2024].

Reeder, K., Shapiro, J., & Wakefield, J. (2007). *The effectiveness of speech recognition technology in promoting reading proficiency and attitudes for Canadian immigrant children.* Proceedings of the 9th European Conference on Reading [online].

Reynolds, C. R., Livingston, R. A., & Allen, D. N. (2021). *Mastering modern psychological testing: Theory & methods (Second Edition).* Cham: Springer Nature Switzerland.

Sabu, K., & Rao, P. (2018). Automatic assessment of children's oral reading using speech recognition and prosody modeling. *CSI Transactions on ICT, 6*, 221–225. https://doi.org/10.1007/s40012-018-0202-3

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia, 126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864.

Silva, W. A., Carchedi, L. C., Junior, J. G., de Souza, J. V., Barrere, E., & de Souza, J. F. (2021). A framework for large-scale automatic fluency assessment. International *Journal of Distance Education Technologies (IJDET), 19*(3), 70–88. https://doi.org/10.4018/IJDET.2021070105.

Society for Learning Analytics. (2011). *What is Learning Analytics?* [online]. Available: https://www.solaresearch.org/about/what-is-learning-analytics/ [Januari 21, 2024].

Swart, N. M., Gubbels, J., in 't Zandt, M., Wolbers, M. H. J., & Segers, E. (2023). *PIRLS-2021: Trends in leesprestaties, leesattitude en leesgedrag van tienjarigen uit Nederland* [online] [PIRLS-2021: Trends in the reading performance, reading attitude and reading behaviour of ten year olds from the Netherlands]. Expertisecentrum Nederlands. Available: PIRLS-2021_Rapportage.pdf(expertisecentrumnederlands.nl) [Januari 21, 2024].

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., Vloedgraven, J. & De Wijs, A. (2018a). *Wetenschappelijke verantwoording DMT* [Scientific justification DMT] [online]. Cito: Arnhem. Available: https://cito.nl/media/2juinmdl/106-cito_lvs-dmt_gr-3-tm-halverwege-gr-8_wet-verantwoording.pdf [Januari 21, 2024].

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., & De Wijs, A. (2018b). *Wetenschappelijke verantwoording AVI* [Scientific justification AVI] [online]. Cito: Arnhem. Available: https://cito.nl/media/hdqkk1if/109-cito_lvs-avi_gr-3-tm-halverwege-gr-8-wet-verantwoording.pdf [Januari 21, 2024].

Truong, Q. T., Kato, T., & Yamamoto, S. (2018). *Automatic Assessment of L2 English Word Prosody Using Weighted Distances of F0 and Intensity Contours.* Paper presented atInterspeech 2018, Hyderabad, India, 2-6 September [online].

University of Oregon (2020). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Administration and Scoring Guide*. Eugene, OR: University of Oregon. https://dibels.uoregon.edu

Veenendaal, N.J., Groen, M.A., & Verhoeven, L. (2016). Bidirectional Relations between Text Reading Prosody and Reading Comprehension in the Upper Primary School Grades: A Longitudinal Perspective. *Scientific Studies of Reading. 20*(3), 189–202. https://doi.org/10.1080/10888438.2015.1128939

van der Velde, M., Molenaar, B., Veldkamp, B. P., Feskens, R.C.W., & Keuning, J. (2024). What do they say? Assessment of oral reading fluency in early primary school children: A scoping review. *International Journal of Educational Research, 128*, 102444. https://doi.org/10.1016/j.ijer.2024.102444

Wei, X., Cucchiarini, C., van Hout, R. W. N. M., & Strik, H. (2022). Automatic Speech Recognition and Pronunciation Error Detection of Dutch Non-native Speech: cumulating speech resources in a pluricentric language. *Speech Communication, 144*, 1–9. https://doi.org/10.1016/j.specom.2022.08.004

Yilmaz, E., & Pelemans, J., van Hamme, H. (2014). *Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model.* Paper presented at Interspeech 2014, Singapore, 14-18 September [online]. https://doi.org/10.21437/Interspeech.2014-256.

Chapter 4

Speech enabled reading fluency assessment: A validation study

van der Velde, M., Harmsen, W., Veldkamp, B. P., Feskens, R.C.W., Keuning, J., & Swart, N. (2025). Speech enabled reading fluency assessment: A validation study. *International Journal of Artificial Intelligence in Education*, 1-27. https://doi.org/10.1007/s40593-025-00480-y.

ABSTRACT

Although the ability to comprehend what one is reading is one of the most fundamental necessities to function within society, the reading comprehension skills of students have recently been on the decline in many countries. An essential prerequisite to reading comprehension is the ability to read fluently, which is defined as the ability to read (aloud) with accuracy, speed, automaticity and prosody. Current oral reading fluency assessment instruments seldom provide detailed diagnostics however, and bestow a heavy testing burden on practitioners. Recent developments in Artificial Intelligence-based assessment methodology might provide a solution to current assessment issues, but thorough validations of such procedures have proven scarce. This study evaluates whether valid word decoding and passage reading measures (accuracy, speed and automaticity) can be generated for a semi-transparent language, using an automatic speech recognition (ASR) based oral reading fluency assessment instrument. A validation study was conducted, using the Argument Based Approach to Validation. Data concerned 176 h of speech data, and the results of 569 and 622 oral word- and passage reading tests that are currently administered in primary schools, from 653 children attending the second- or third grade of Dutch primary education. The results of the validation indicate that it is possible to generate fluency metrics for a semi-transparent language, using an ASR-based oral reading fluency assessment instrument. Future researchers are advised to further optimize the ASR, evaluate its errors, and realize a prosody component, completing the envisioned reading fluency assessment instrument, thereby improving reading fluency assessment throughout primary education.

4.1 INTRODUCTION

Being able to comprehend what you are reading is one of the most fundamental necessities to function within society. Nevertheless, the reading comprehension skills of students have recently been on the decline in many countries (Meelissen et al., 2023; Mullis et al., 2023), increasing the risk of functional illiteracy among future generations. One of the most important prerequisites to reading comprehension is the ability to read fluently (Fuchs et al., 2001; Hoover & Gough, 1990), which is often defined as the ability to read aloud with accuracy, speed and proper expression (Kuhn et al., 2010; Pikulski & Chard, 2005). Indeed, the importance of fluency to comprehension is so well-established that some interventions aimed at improving comprehension have focused on improving fluency instead (Mastropieri & Scruggs, 1997; Reutzel & Hollingsworth, 1993). While it is not suggested that reading fluency interventions should substitute for specialized reading comprehension training, such implementations do demonstrate the relevance of fluency to comprehension.

To elaborate, research has increasingly linked comprehension to the automaticity and prosody of reading (Groen et al., 2018; Kuhn et al., 2010). Here, automaticity reflects the ability to decode written text with sufficient accuracy and speed (e.g. Kim et al., 2021), which a reader attains through instruction and practice (Logan, 1988). Automatic reading reduces the mental effort required to read, allowing the reader to focus on more cognitively demanding tasks, like comprehension (Aldhanhani & Abu-Ayyash, 2020; Morris & Perney, 2018). To elaborate, Perfetti's (1985) verbal efficiency theory states that lower complexity tasks must be mastered to some degree, before more complex tasks can occur during reading. Correspondingly, automaticity can be related to comprehension through proficiency in decoding, word identification and the retention of limited cognitive resources (Perfetti & Stafura, 2014)

Meanwhile, prosody reflects expressive components of reading, such as phrasing, expression, intonation, stress and pitch (Miller & Schwanenflugel, 2008; Share, 2008), which facilitate or enhance the retention of meaning (Kuhn et al., 2010; Miller & Schwanenflugel, 2008; Silva et al., 2021). Prosody and comprehension have been shown to affect one another throughout most of primary education (Veenendaal et al., 2016), even when controlling for automaticity (Groen et al., 2018; Veenendaal et al., 2015). In short, the relationship between fluency and comprehension has been well-documented and involves automaticity and prosody.

In contrast, standardized oral reading fluency assessment tools, such as Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DORF; University of Oregon, 2020), and the Test of Word Reading Efficiency Sight Word Efficiency (TOWRE-SWE; Torgesen et al., 1997), assess fluency using the number of words read correctly per minute (WCPM). This metric only integrates measures of accuracy (words read correctly) and speed (per minute), thereby solely reflecting automaticity. Meanwhile, prosody is

separately measured through subjective rating scales (Kuhn et al., 2010; Morrison & Wilcox, 2020), that require multiple trained raters to obtain reliable results.

In practice, the time-consuming nature of oral reading fluency assessment, the requirement for training, and the resulting testing burden, has led teachers and other practitioners to relinquish the assessment of prosody. In addition, practitioners stray from obtaining detailed automaticity diagnostics, as their extraction further increases assessment duration. This lack of complete and detailed diagnostics is problematic, as it complicates the implementation of individualized reading instruction, which has increasingly been identified as crucial for the optimal development of individual readers (Bray & McClaskey, 2015; Connor & Morrison, 2016).

To summarize, although fluency is a crucial prerequisite for developing reading comprehension skills, its assessment seldom supplies practitioners with detailed diagnostics, and bestows a heavy testing burden upon them. Therefore, an assessment instrument that provides detailed individualized diagnostics, and that reduces the testing burden placed on practitioners, could improve the development of reading fluency and comprehension alike. To fulfill this ambition, a recent review on oral reading fluency assessment has suggested the use of artificial intelligence-based speech technology (van der Velde et al., 2024a), an approach that has served a multitude of scientific fields.

Over the last two decades, major advances have been made regarding the availability and complexity of artificial intelligence-based technology, increasing their applicability and relevance for assessment (Clarke-Midura & Dede, 2010). For example, developments in text processing have allowed for the automatic evaluation of theoretical papers (Rokade, 2018). Likewise, the development of automatic speech recognition (ASR), which concerns the "independent, machine-based process of decoding and transcribing oral speech" (Levis & Suvorov, 2012, p. 1), has made audio data a valuable source of information.

With regard to reading fluency, the effectiveness of ASR has been demonstrated in early work by Mostow et al. (2003) and Reeder et al. (2007), while the current relevance of ASR is reflected through its central role within recent reading fluency assessment frameworks (Silva et al., 2021). Correspondingly, much work has been conducted to automate fluency assessment through ASR, especially for English readers (Cheng & Shen, 2010; Loukina et al., 2019; Sabu & Rao, 2018). This, in turn, has led to the creation of automatic reading fluency tools such as the Fluent Oral Reading Assessment (FLORA; Bolaños et al., 2013), and Moby.Read (Cheng, 2018).

Even though the body of literature that implements ASR within the English language is substantial, this does not necessarily indicate that these results are generalizable to all languages. Namely, although most literary research focusses on English, English has long been established as an outlier orthography (Share, 2008). Specifically, English has relatively high irregularity, or low transparency, with regard to grapheme-phoneme correspondences when compared to other languages. As previous research has illustrated, the transparency

of a language impacts the way in which children process, learn, and attempt to express a language (Smith et al., 2021; Wimmer & Goswami, 1994). For example, children learning less transparent orthographies favor the direct recognition of words or letter-strings over converting graphemes into phonemes.

Throughout the last decade, attempts have been made to implement ASR within non-English languages. For example, Proença et al (2015; 2017) automatically evaluated disfluencies in the speech of Portuguese children. Another example concerns the Dutch language, where crucial steps have been made regarding the automation of assessment of first graders (Bai et al., 2020; Bai et al., 2021), and secondary language learners (Cucchiarini et al., 2009; Wei et al., 2022). However, less attention has been placed on children attending Grades 2 and 3, even though these prominently featured in the reading fluency assessment literature (van der Velde et al., 2024a). In addition, current ASR implementations primarily focus on accuracy, ignoring speed and prosody.

To overcome current assessment shortcomings, and to investigate the full potential of ASR based fluency assessment, an assessment instrument that utilizes automatic speech recognition to provide detailed diagnostic information on all fluency components has been developed (van der Velde et al., 2024b).

4.1.1 SERDA: Automatic oral reading fluency assessment for dutch

The Speech Enabled Reading Diagnostics App (SERDA) is a Dutch oral reading fluency assessment instrument, developed to improve reading education at the primary school level (van der Velde et al., 2024b). Through the incorporation of speech recognition, speech-based diagnostics, and their conversion into didactic suggestions for practitioners, SERDA allows for the provision of individualized feedback on children's oral word and passage reading performance, as well as detailed information on all fluency components. All the while, SERDA's short administration duration and automatic scoring should reduce the testing burden placed on teachers and other practitioners.

Although these findings are promising, the ASR-based accuracy, speed, automaticity and prosody metrics of any automatic fluency instrument should be thoroughly validated before statements can be made about their usability in practice. Given that reading fluency is currently assessed through the speed, accuracy and automaticity of reading in practice, we argue that it should first be proven that ASR-based metrics can validly substitute for their pen-and-paper contemporaries, before prosody is considered. Therefore, the current study will focus on validating SERDA's word decoding and passage reading tasks, excluding prosodic metrics, to determine whether an ASR-based reading fluency assessment instrument can provide valid word decoding and passage reading metrics.

4.1.2 Validating ASR-based decoding scores

To substitute for current instruments, an ASR-based decoding instrument should provide observable and reliable scores based on children's oral reading performance. Moreover, the ASR-based scores, obtained over a limited sample of reading items and primary school children, should be generalizable to all potential samples of reading items and primary school children. Furthermore, the scores should be proven to reflect oral reading skills, allowing for claims to be made regarding children's oral reading performance. Finally, the reading tasks should provide scores that allow practitioners to differentiate between good, average and less proficient oral readers, such that decisions with regard to development and proficiency can be made.

In order to evaluate whether these requirements are met, we will apply the Argument-Based Approach to validation (ABP; Kane, 1992, 2006, 2013). Within the ABP, an Interpretation and Use Argument (IUA) is specified, which describes the inferences and assumptions underlying the proposed interpretation of assessment results. Then, a validity argument is defined, describing the process of evaluating the components of the IUA through gathered evidence. Lastly, the validation as a whole is evaluated. Specifically, it is evaluated whether the correct assumptions and inferences are addressed, whether the inferences can be justified, and whether the validity argument, as a whole, is plausible.

4.1.3 The present study

The present study aims to evaluate whether it is possible to generate valid word decoding and passage reading measures for a semi-transparent language, using an ASR-based oral reading fluency assessment instrument.

Based on the ABP framework, the main question answered with this study is:

- Can an oral reading fluency assessment instrument that utilizes automatic speech recognition provide valid word decoding and passage reading scores?

In order to answer this question, we will answer the following sub-questions:

1. Can performances on the reading tasks be translated into observable and reliable ASR-based scores?
2. Is the sample of reading tasks and primary schools representative of the population of reading tasks and primary schools?
3. Do the ASR-based scores reflect oral reading skills, allowing for claims to be made regarding children's oral reading performance?
4. Can the ASR-based oral reading scores differentiate between good, average and less proficient oral readers, such that they can be used to make decisions regarding their proficiency and development?

4.2 METHODS

To determine whether an ASR-based oral reading fluency assessment instrument could provide valid word decoding and passage reading scores, we administered SERDA's word- and passage reading tasks (van der Velde et al., 2024b), as well as the most popular instruments to monitor oral word and passage reading skills in the Netherlands: the Three Minute Task [*Drie-minuten-toets*; DMT] (van Til et al. 2018a) and AVI [*Analyse van Individualiseringsvormen*; AVI] (van Til et al. 2018b).

4.2.1 Participants

176 hours of speech data were obtained, as well as the results of 569 DMT and 622 AVI administrations, from 653 (52% girls) children attending the second and third grade of Dutch primary education. Children attended 19 different primary schools, selected to represent dialect regions (Cucchiarini et al., 2008) and school-weights, an indicator of expected school performance and social economic status of children's parents (Inspectorate of Education, 2024). The average age of the children was seven and a half (*SD = 0.74)*, with children attending Grade 2 being one year younger, on average, than children attending Grade 3.

4.2.2 Materials

4.2.2.1 *SERDA: Word and passage decoding.*

SERDA's word and passage reading tasks were individually administered on a tablet, during which children's speech was recorded through a microphone. The word decoding task contained 150 words, chosen based on the DMT (van Til et al., 2018a) and expert opinion. Words were divided over three 50-word subtasks, which varied with regard to the number of syllables per word and the complexity of reading difficulties. The presentation of words followed a progressive demasking design (Grainger & Segui, 1990) to allow for accurate reading speed estimation. During the progressive demasking task, a mask was placed over the words at an increasing interval, such that the to be read word became visible for longer over time, until the participant indicated that they were able to recognize the word. Before administration, children were instructed to tap the screen as quickly as possible once the presented word was recognized, after which they read the word out loud as accurately as possible.

The passage reading task contained three passages of about 175 words, which were constructed using the guidelines of the AVI (van Til et al., 2018b). The passages were written by children's authors, discussed topics of interest to children, and contained multi-syllable words with reading complexities that corresponded to those, respectively, expected at the end of second grade, the middle of third grade and the end of the third grade. Children

were instructed to read the passages as quickly and accurately as possible, including the title. The task was finalized once the entire passage was read, or after 3 minutes had passed.

Each word and passage reading subtask yielded audio- and log files, based on which item-, subtask- and person-level measures were extracted. An overview of the extracted measures can be found in Table 1. To extract these measures, the same methodology was adopted as described in van der Velde et al. (2024b), using an updated version of the ASR model. In addition, we utilized Item Response Theory (Hambleton & Swaminathan, 1985) to extract item- and person parameters. Specifically, we applied the Hierarchical Bayesian joint modeling approach (van der Linden, 2007), which allows for the joint modelling of accuracy and speed scores, integrating both accuracy and speed information during the estimation of children's oral reading skills. Here, we used the "LNIRT" (Fox et al., 2021) R-package to obtain item difficulty and discrimination parameters for all items. Finally, we calculated LNIRT word decoding and passage reading ability and speed estimates for each person, using their item-level accuracy and speed scores.

Before model specification, words or items with little to no variability and persons with extremely unlikely scores or mostly missing data were removed. For the word decoding task, we retained observations of 633 children for 149 items. For the passage reading task, we retained observations of 631 children for 526 items.

Table 1. Item, subtask and person level measures extracted by SERDA.

Measure	Word Decoding	Passage Reading
Item level		
Accuracy	0 or 1	0 or 1
Speed	Flashing time (seconds)	Speaking duration (seconds)
WCPM	-	-
Subtask level		
Accuracy	Number of words read correctly	Number of words read correctly
Speed	Words read divided by total flashing time	Words read divided by task duration
WCPM	Accuracy divided by total flashing time	Accuracy divided by task duration
Person level		
Accuracy	Average subtask-level accuracy	Average subtask-level accuracy
Speed	Average subtask-level speed	Average subtask-level speed
WCPM	Average subtask-level WCPM	Average subtask-level WCPM

Note. Adapted from van der Velde et al. 2024b.

4.2.2.2 *Word decoding: Three minute task [Drie-minuten-toets; DMT].*

The DMT is an on-paper examination, aimed at monitoring the development of word decoding skills of children during Grades 1 to 6 of primary education (van Til et al., 2018a). DMT administrations were individually conducted and scored by teachers of the schools the children attended. Children read up to three word lists of increasing difficulty, as quickly and accurately as possible, for a duration of one minute per list. Then, based on their performance compared to grade-specific norms for the population of primary schoolers, children were classified into one of five categories, ranging from the 20% best to least developed readers. As these categories were provided by school after specification, no reliability or validity information was collected. However, the reliability and validity of the DMT has previously been thoroughly investigated (Van Til., 2018a).

Based on the DMT classifications, we specified three proficiency classes. To elaborate, children in the 20% least developed DMT group were classified as "Less Proficient", while children classified into the top 20% were classified as "Highly Proficient". Children classified into the middle 60% of readers were classified as "Averagely Proficient".

4.2.2.3 *Passage reading: AVI [Analyse van Individualiseringsvormen; AVI].*

The AVI is an on-paper examination, aimed at monitoring the development of passage reading skills of children during Grades 1 to 6 of primary education (van Til et al., 2018b). AVI administrations were individually conducted and scored by teachers at the schools the children attended. During the AVI, children read Grade-level passages of increasing difficulty. The reading of passages continued until the child was unable to meet national norms, either by making too many mistakes, by not reading quickly enough, or by a combination of these factors. Then, based on the highest Grade-level passage read successfully, an AVI classification was provided. Specifically, the AVI categorizes children into one of twelve levels, ranging from the start to the end of primary education, providing an indication of the child's progress throughout primary education. As these categories were provided by the school after specification, no reliability or validity information was collected. However, the reliability and validity of the AVI has previously been thoroughly investigated (Van Til., 2018b)

Based on the AVI classification, we specified three proficiency classes. Specifically, children who obtained an AVI categorization below the level of second grade were classified as "Less proficient", while children with an AVI class at the level of the second or third grade were classified as "Averagely Proficient". Children with an AVI classification above the third grade were classified as "Highly Proficient".

4.2.3 Argument-based validation

The validation was implemented through the extended ABP (Kane, 1992, 2006, 2013; Wools et al., 2010). First, we specified the explicit inferences made about the decoding

scores, by means of an IUA. Then, we described the validity arguments for each step of the IUA. Correspondingly, we specified the analyses conducted and evidence gathered for each validity argument. Finally, we evaluated the validity in its entirety, including the validation procedure.

Making the proposed interpretations and uses of test scores explicit was done through the specification of claims. For example, it could be claimed that the performance on a test provides a score that is observable. To evaluate this claim, warrants, rebuttals and backings were specified, which respectively concern statements that allow for the acceptance of the claim, evidence that refutes the claim or warrant, and evidence that supports the claim or warrant (Toulmin, 2003). It follows that the presentation of sufficient and qualitatively sound backings and warrants, alongside the justified rejection of rebuttals, leads to the acceptance of a claim.

The IUA for the current study is presented in Figure 1. Additionally, a detailed overview of the proposed inferences, assumptions, and sources of evidence is provided in Table 2. Correspondingly, the exact claims, warrants, rebuttals and backings for each inference are shown in Appendix A.

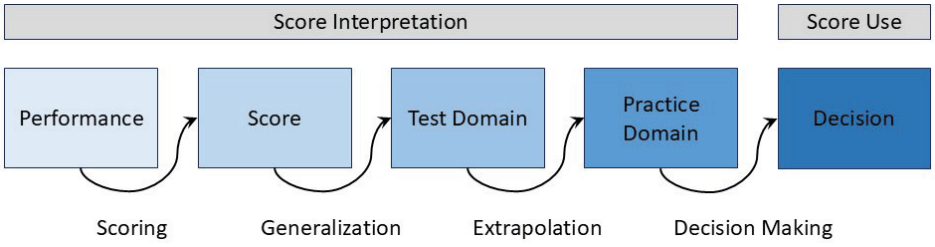


Figure 1. *Interpretation and Use Argument (IUA) for the Validation*

Table 2. Inferences, assumptions, and sources of evidence used to validate the reading tasks

Inferences	Assumptions	Sources of evidence
Scoring: Performances on the reading tasks can be translated into observable and reliable ASR-based scores.	- The scoring algorithm generates meaningful scores	- Evaluation of the ASR-based scoring algorithm through a comparison with human raters.
	- The ASR-based word decoding and passage reading scores are reliable.	- Psychometric evaluation of the reliability of the word decoding and passage reading scores for all levels of ability and speed.
Generalization: the oral reading scores, obtained over a limited sample of items and primary schools, can be used to make inferences about all possible samples of items and primary schools.	- The tasks reflect relevant aspects of the learning goals and methods for oral reading in early primary education.	- Compare the reading tasks to relevant learning goals and methods for early primary education in the Netherlands.
	- The reading tasks reflect the difficulty level expected of oral reading tasks in Grades 2 and 3.	- Item difficulty parameters (mostly) cover the range of observed ability estimates.
	- The sample of primary schools is representative of the population.	- Compare the dialect region and school-weight distributions in the sample and population of primary schools.
Extrapolation: The ASR-based oral reading scores can be used to make claims about children's oral reading performance.	- The reading tasks measure the same underlying construct as the DMT and AVI.	- Correlate the LNIRT ability and speed estimates, and the person-level WCPM measures, with the DMT and AVI classifications.
	- The reading tasks are authentic, representing all relevant aspects of oral reading.	- Evaluate whether the reading tasks provide information on oral reading accuracy, speed and automaticity.
Decision Making: The ASR-based oral reading scores can be used to make decisions about proficiency and development.	- The oral reading scores differentiate between good and less proficient oral readers.	- Evaluate the discrimination-parameters of all items.
	- Misclassifications into highly, averagely and less proficient oral readers are minimized.	- Predict DMT/AVI proficiency classes, using the oral reading scores.

4.2.4 Data analysis

The analyses conducted throughout the present study were used to evaluate the validity arguments underlying the proposed inferences. All statistical analyses were conducted using RStudio (version 4.3.1; Posit Team, 2023).

4.2.4.1 *Scoring inference.*

To evaluate whether children's performances on the reading tasks could be translated into observable and interpretable scores, we validated the ASR scoring-algorithm by comparing its item-level accuracy and speed scores to human annotations. In addition, we evaluated the reliability of the accuracy and speed scores.

To evaluate the validity of the item-level ASR-based accuracy and speed scores, we compared them with manual annotations. For the word decoding task, we only validated the accuracy scores, as the speed scores concerned logged data. Specifically, we obtained human annotations for 333 word decoding subtasks. These annotations were made by test leaders during task administration, using SERDA's build-in test-leader app. Test leaders could label words as "read incorrectly", while leaving them unlabelled marked them as read correctly.

For the passage reading task, test leaders reported that children read too fast to accurately annotate. Therefore, we obtained orthographic transcriptions of 18 subtasks, made by two Linguistics graduates. Transcribers respectively transcribed 12 and 9 subtasks, three of which were transcribed by both, in two tiers, using PRAAT (Boersma & Weenink, 2024). The first tier contains the prompts presented to the speaker. Each prompt is a word from the passage reading task, transcribed in an interval that contains all attempts a speaker made to read the prompt. The second tier contains the orthographic transcription of the audio, where each attempt to read a word was transcribed in a separate interval. If the last attempt in tier 2 was equal to the prompt word in tier 1, the prompt was labelled as read correctly. Meanwhile, the item-level speed scores were defined as the duration of the final reading attempt for a prompt.

Consistent with earlier studies on automatic accuracy assessment (Kheir et al., 2023), accuracy measures were encoded in terms of reading errors. Thus, correctly read words were labeled as False (i.e., word reading does not contain an error) and incorrectly read words as True (i.e., word reading does contain an error). Subsequently, we compared the ASR-based accuracy scores of both tasks to human annotations, using Matthew's Correlation Coefficient (MCC). MCC yields a score between -1 and 1, where a score of 0 indicates that the correspondence is no better than chance. The MCC was chosen since our dataset is unbalanced, containing more correctly- than incorrectly read words, making the MCC a trustworthy and complete performance indication (Chicco et al., 2021). In addition, to enable a more thorough interpretation of these results, we computed the sensitivity (i.e., the proportion of decoding errors that are predicted to be incorrect), specificity (i.e. the proportion of actually correct readings that are predicted to be correct), and precision (i.e., the proportion of predicted incorrect readings that are actually incorrect).

Finally, to evaluate the passage reading speed scores, we computed correlations between human and ASR-based item-level speed scores. These are the correlations between the speed scores of transcriber 1 and the ASR, the speed scores of transcriber 2 and the ASR, and the speed scores of both raters and the ASR. However, the ASR is currently only able to produce speed scores for correctly read words. To elaborate, due to the large variation in possible reading errors that a child can make (e.g. repetitions at sub-word, word, phrase level or insertions of words that are not in the prompt), defining the desired output is very difficult. As a result, we only included speed scores for words that were read correctly.

To estimate the reliability of the ASR-based accuracy and speed scores we calculated Cronbach's Alpha, Goodman's Lambda 2 and the Greatest Lower Bound for the accuracy and speed measures of the word and passage reading tasks. In addition, we investigated the posterior standard deviations for children's LNIRT ability and speed estimates. However, as the LNIRT model's assumption of log-normal residuals was violated for most items, we replicated the generation of person- and item parameter estimates using a different IRT model, which only incorporates the item-level accuracy scores. The results, which can be found in Supplementary Appendix A, provided no indication that the violation of the log-normal residuals assumption substantially affected the results.

4.2.4.2 *Generalization inference.*

We assume that the sample of oral reading items can be used to make inferences about all samples of items if they reflect Dutch learning goals and methods for early primary education, if they match the difficulty level expected of oral reading tasks in Grades 2 and 3, and if the sample of primary schools represents the population.

To determine whether the reading tasks reflect Dutch learning goals and methods, the main argumentation of current guidelines was compared to the content of the reading tasks. The difficulty of the oral reading items was investigated by evaluating the distribution of the LNIRT item difficulty parameters and ability estimates. Finally, we evaluated the representativeness of the sample of primary schools by comparing the distribution of dialect region and school-weight to those found in the population.

4.2.4.3 *Extrapolation inference.*

The oral reading scores are assumed to provide information about oral reading performance if they measure the same underlying construct as their pen-and-paper predecessors, and if they provide information on oral reading accuracy, speed and automaticity.

To investigate whether the reading tasks measure the same construct as the DMT and AVI, we calculated correlations between the LNIRT ability and speed metrics, the person-level WCPM-scores, and the classifications of the DMT and AVI. The reading tasks' authenticity was determined by discussing whether they reflect all relevant aspects of oral reading.

4.3.4.4 *Decision making inference.*

It is assumed that children's oral reading scores can be used to guide decisions regarding their performance if the reading tasks contain items that can discriminate between good and less proficient oral readers, and if misclassifications into highly, averagely and less proficient readers are minimized.

The discriminative ability of the reading tasks was determined by evaluating the LNIRT item discrimination parameters. To investigate the minimization of misclassification, we predicted the DMT and AVI proficiency classes through ordinal regression. Models included the LNIRT ability and speed estimates and the person-level WCPM scores of the word decoding and passage reading task. We also included children's Grade, as the DMT and AVI are normed based on grade. Model performance was evaluated by calculating the weighted kappa with quadratic weights, and the average F1-score over all proficiency classes.

4.3 RESULTS

4.3.1 Scoring inference

To validate the ASR-based item level accuracy scores, we compared them to accuracy scores from human raters, using the MCC, sensitivity, specificity and precision. The results are presented in Table 3.

Table 3. Evaluation metrics for the ASR-Based Item-level Accuracy Scores.

Task	Subtasks	Items (inc, cor)	MCC	Sensitivity	Specificity	Precision
Word	333	16650 (2117, 14533)	0.43	0.93	0.69	0.31
Passage	18	3156 (542, 2614)	0.55	0.76	0.86	0.54

We found moderate agreement between human and automatic accuracy measures for the word decoding (MCC = 0.43) and passage (MCC = 0.55) reading task, and for the inter-rater agreement of the passage reading accuracy scores (MCC = 0.59). Additionally, the word (sensitivity = 0.93, specificity = 0.69) and passage (sensitivity = 0.76, specificity = 0.86) reading tasks showed moderate to high sensitivity and specificity. However, the word decoding task showed low precision (0.31), while the passage reading task showed moderate precision (0.54).

Then, we compared the ASR-based item-level speed scores of the passage reading task to their human equivalents. We found moderate to strong correlations between the ASR and transcriber 1 (r = 0.61), transcriber 2 (r = 0.57) and both transcribers (r = 0.59).

The reliability of the accuracy and speed scores was evaluated using Cronbach's Alpha, Goodman's Lambda 2 and the Greatest Lower Bound, and by evaluating the posterior

standard deviation estimates from the LNIRT model for all LNIRT ability and speed estimates. Table 4 presents Cronbach's Alpha, Goodman's Lambda 2 and the Greatest Lower Bound for the word decoding and passage reading task. Reliability estimates ranged from 0.96 to 1, indicating that both tasks show excellent reliability.

Table 4. Cronbach's alpha, goodman's lambda 2 and the greatest lower bound for the accuracy and speed measures of the word decoding and passage reading tasks

	Word Decoding		Passage Reading	
Measure	**Accuracy**	**Speed**	**Accuracy**	**Speed**
Cronbach Alpha	0.96	0.99	0.99	0.99
Goodman's Lambda 2	0.96	0.99	0.99	0.99
Greatest Lower Bound	0.98	1.0	1.0	1.0

Figure 2 shows the posterior standard deviation estimates for the LNIRT ability and speed estimates. Posterior standard deviations were relatively low, especially for the LNIRT speed estimates. Higher uncertainty was observed for relatively low and high ability estimates, and for a handful of negative passage speed estimates.

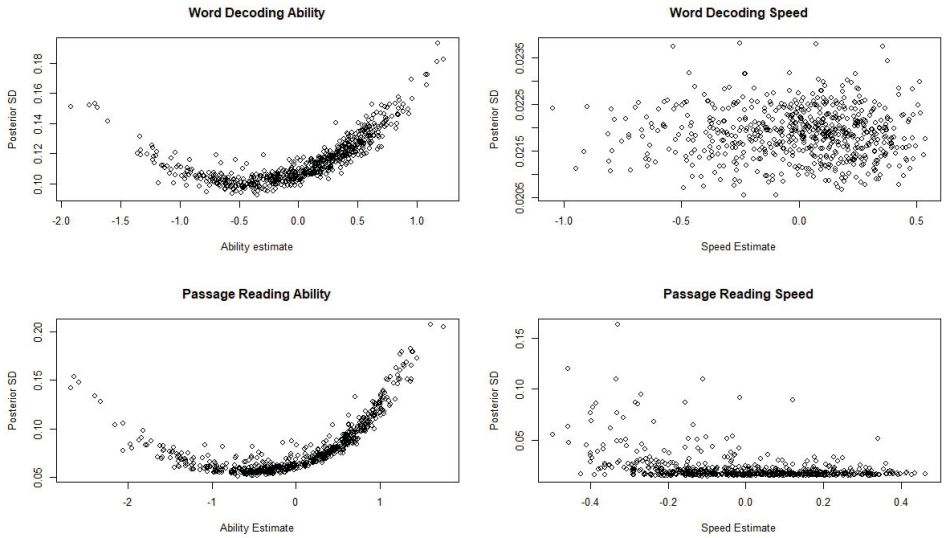


Figure 2. *Posterior Standard Deviations for the LINIRT Word Decoding and Passage Reading Ability and Speed Estimates*

4.3.2 Generalization inference

To evaluate the generalizability of the oral reading scores, we compared the content of the reading tasks to current Dutch learning goals and methods for early primary education, and compared the distributions of the LNIRT ability- and item-difficulty estimates. In addition, we compared the sample of primary schools to their population with regard to dialect region and school weight.

Dutch can be considered a semi-transparent language because of its relatively consistent grapheme-phoneme correspondences (Borgwaldt et al., 2004; Seymour et al., 2003). Correspondingly, learning to read is primarily based on grapheme-phoneme rules. Accordingly, most schools in the Netherlands use a reading instructional method (mostly either Veilig Leren Lezen [learning to read safely; Zwijsen Educatieve Uitgeverij, 2023] or Lijn 3 [Track 3; Malmberg, n.d.]) that focusses on these rules to instruct decoding. These methods focus on teaching children grapheme-phoneme correspondences (including digraphs such as *ei, ui* and *ou*) and on learning to read simple structured words (both mono- and bi-syllabic) throughout the first half of Grade 1. Then, focus gradually shifts towards automatizing the reading process and reading more complex structured words (e.g. consonant clusters and bi-/polysyllabic words). After Grade 1, schools either use a method for advanced decoding instruction, incorporate it in instruction for other language-related instruction (e.g., reading comprehension of language arts), focus on furthering the automatization of the reading process, and/or focus instruction on advanced reading difficulties (e.g., loanwords and the use of *c, x* and *y* in words).

Likewise, SERDA's reading tasks build up in difficulty over its subtasks. Specifically, the first word decoding subtask primarily contains one-syllable words with various consonant-vowel combinations and relatively basic reading difficulties (e.g. open syllable, sch-). Meanwhile, the second subtask focusses on one-, two- and three syllable words with more advanced reading difficulties (e.g. ge-, -lijk), while the third subtask contains two-to-four syllable words with more complex reading difficulties (e.g. -isch, loanwords). The same can be said for the passage reading task, which focusses on reading mono-, bi- and polysyllabic words with an increase in orthographic inconsistencies and complexities over subtasks. Finally, conform the instructional methods discussed, the reading tasks place specific focus on automatizing the reading process through a focus on both accuracy and speed in instruction and performance metrics alike. Altogether, the reading tasks closely match the way in which children are taught how to read in the Netherlands.

Figure 3 shows the distribution of the LNIRT ability- and item-difficulty estimates of the word decoding and passage reading task. For both tasks, the item difficulty parameters cover the range of ability estimates, with the exception of some extremely low passage reading ability estimates. However, the difficulty of the items is generally low, showing relatively few items with positive difficulty estimates.

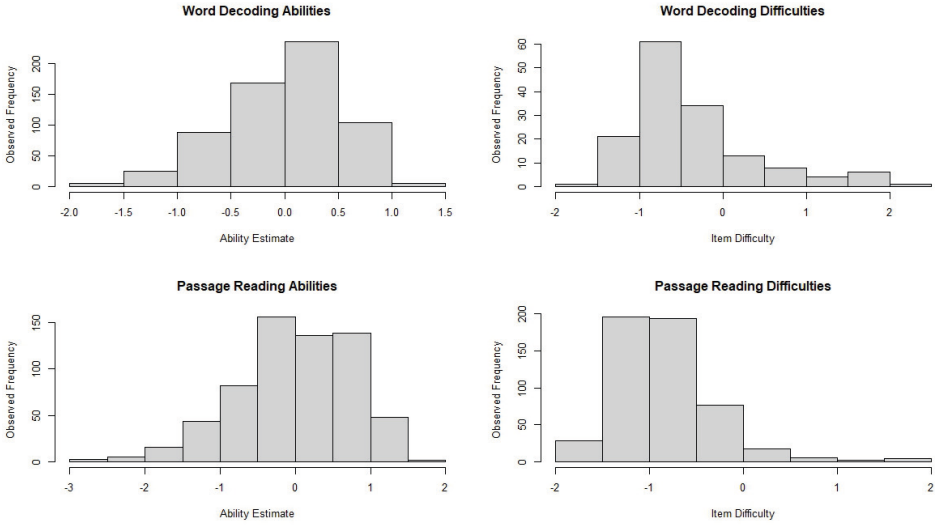


Figure 3. *LNIRT Ability and Difficulty Estimates for the Word Decoding and Passage Reading Task*

Figure 4 shows the dialect region and school-weight distributions for the population and sample of primary schools in the Netherlands. Generally, the sample resembles the population. However, an underrepresentation of schools in the Western dialect region was observed, while the sample overrepresents schools with low school-weights.

Figure 4. *Sample and Population Distributions for Dialect Region and School Weight*

4.3.3 Extrapolation inference

To evaluate whether the oral reading scores allow for claims about oral reading performance, we calculated correlations between the LNIRT ability and speed estimates, the person level WCPM scores, and the categories of the DMT and AVI. In addition, we evaluated whether the word decoding and passage reading tasks provide information on all aspects of oral reading.

Table 5 presents the correlations between the ability estimates of the word decoding and passage reading task, the person level WCPM measures, and the categories of the DMT and AVI. Correlations varied between 0.32 and 0.88, showing weak to very strong correspondences between the oral reading metrics.

Table 5: Pearson and spearman correlations between the LNIRT ability and speed estimates, the person level WCPM Metrics, and the categories of the DMT and AVI

Metric	Ability Word	Speed Word	Ability Passage	Speed Passage	WCPM Word	WCPM Passage	DMT	AVI
Ability Words	1	-	-	-	-	-	-	-
Speed Word	0.65	1	-	-	-	-	-	-
Ability Passage	0.64	0.50	1	-	-	-	-	-
Speed Passage	0.61	0.75	0.43	1	-	-	-	-
WCPM Word	0.87	0.88	0.56	0.74	1	-	-	-
WCPM Passage	0.68	0.77	0.70	0.86	0.79	1	-	-
DMT	0.40	0.57	0.32	0.60	0.54	0.67	1	-
AVI	0.55	0.69	0.42	0.78	0.68	0.80	0.65	1

Note. Correlations with the DMT and AVI were Spearman correlations, while the others concerned Pearson correlations. All correlations were significant at a = 0.001.

To reflect oral reading skills, the reading tasks should provide information regarding children's oral reading accuracy, speed and automaticity. Table 1 shows that the word and passage reading tasks provide information on oral reading accuracy, speed and automaticity (WCPM). Meanwhile, the validation has shown that item-level information is reliable and that most of the resulting child-specific metrics moderately to strongly resemble current oral reading metrics. Therefore, we argue that child-specific information is provided on children's oral reading accuracy, speed and automaticity through, respectively, the LNIRT ability and speed estimates, and the WCPM metrics.

4.3.4 Decision making inference

To evaluate whether the oral reading items differentiate between good and less proficient oral readers, we evaluated the LNIRT item discrimination parameters. In addition, we investigated the classification accuracy of ordinal regression models that predicted oral reading proficiency, using the LNIRT ability and speed estimates, the person level WCPM metrics, and children's Grade.

Figure 5 shows the item discrimination parameters of the LNIRT model for the word decoding and passage reading task. Discrimination parameters primarily ranged between 0.8 and 1.4 for both the word and passage reading task, indicating that items generally discriminate moderately well or better (Baker, 2001; Bichi & Talib, 2018). Lower discrimination parameters were primarily observed for single-syllable words and articles.

The ordinal regression model used to predict the DMT proficiency classes showed a weighted kappa of 0.65 and an average F1 score of 0.70, indicating moderate to good classification accuracy. The ordinal regression model used to predict the AVI proficiency classes showed a weighted kappa of 0.73 and an average F1 score of 0.77, indicating moderate to good classification accuracy.

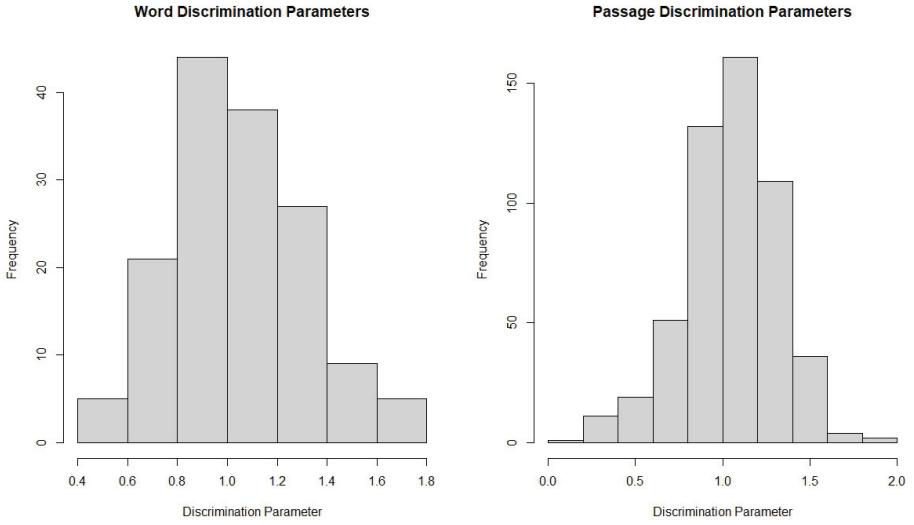


Figure 5. *LNIRT Item Discrimination Parameters for the Word Decoding and Passage Reading Task*

4.3.5 Validation evaluation

Following the presentation of validity evidence, we applied the ABP (Kane, 1992, 2006, 2013) to evaluate whether the proposed interpretations of SERDA's reading tasks can be substantiated. For these purposes, we used the three criteria specified by Wools et al. (2010) to validate the validation process in its entirety.

The first criterion focusses on the complexity of the interpretive argument, as evidenced through the number of inferences and their level of detail. To meet the first criterion, at least four inferences should be specified, each including at least one backing, warrant and rebuttal. As shown in Appendix A, the current study specified four inferences, which were all supplied with at least one backing, warrant and rebuttal. Therefore, the interpretive argument is deemed sufficiently complex and detailed, allowing for the acceptance of the first criterion.

The second criterion focusses on evaluating the presented evidence with regard to plausibility and coherence. For each inference, the validity argument is evaluated in full, resulting in the acceptance or rejection of the inference.

For the scoring inference, we found that the ASR-based oral reading scores resemble human ratings moderately well. In addition, the reliability of the word and passage reading tasks was deemed excellent, showing high internal consistency and low posterior standard deviations for most ability levels. Therefore, the validity evidence substantiates the scoring inference sufficiently to warrant its acceptance.

For the generalization inference, the gathered validity evidence indicates that the reading tasks reflect the way in which children learn to read in the Netherlands, while the difficulty of the tasks matches the expected difficulty in Grade 2 and Grade 3 of primary education. In addition, although some differences with the population were observed, the sample of schools generally represented the population well with regard to the distribution of dialect region and school-weight. Therefore, the validity evidence sufficiently substantiates the acceptance of the generalization inference.

For the extrapolation inference, we found that most LNIRT ability estimates, all LNIRT speed estimates, and the person level WCPM metrics, resembled the DMT and AVI classifications moderately well or better. In addition, we argued that the reading tasks provide item and person level information that reflect oral reading accuracy, speed and automaticity. Based on these results, we conclude that the reading tasks provide scores that reflect oral reading skills, thereby justifying the acceptance of the extrapolation inference.

For the decision making inference, the discrimination parameters from the LNIRT model suggest that most items are able to discriminate between good and less proficient oral readers moderately well or better. In addition, we found moderate to good classification accuracy when predicting the DMT and AVI proficiency classes with the decoding metrics. Therefore, the evidence indicates that the decision making inference is validly assumed, leading to its acceptance.

The third criterion emphasizes the plausibility of the validity argument as a whole, thereby taking into account all validity evidence of all inferences. Accordingly, the third criterion can only be justified if the first two criteria have been met. Based on the acceptance of the first criterion, and each step of the second criterion, the third criterion is also deemed justified. Thus, given that sufficiently numerous and detailed validity evidence has been presented to deem each of the specified inferences plausible, the validation as a whole is also deemed plausible.

4.4 DISCUSSION

The aim of the current study was to investigate whether valid word decoding and passage reading metrics could be generated for a semi-transparent language, using an ASR-based oral reading fluency assessment instrument. The validation was conducted using the extended ABP (Kane, 1992, 2006, 2013; Wools et al., 2010), which includes a validation

evaluation. Subsequently, we present the interpretation of the results, suggestions for future research, and the conclusion.

Based on the results of the validation, the scoring inference was deemed plausible, indicating that oral reading performances can be reliably translated into observable ASR-based oral reading scores. However, while the reliability of the ASR-based scores was excellent, the ASR-based scores showed only moderate resemblance to human raters. To be more specific, both tasks showed high sensitivity and specificity, but only moderate to low precision was observed, indicating that the ASR somewhat overestimates the number of errors a reader makes.

The lower precision could be explained through the unbalanced state of the data, as most items were read correctly. Given that reading errors were the clear minority group, and given that errors were coded as positives, false positive classifications occurred most frequently, leading to lower precision. As this unbalance was more prominently the case for the word decoding task, its precision was especially affected. More generally, a possible explanation for these moderate results is that state-of-the-art ASR models (as used in this study) are trained on adult speech, and therefore perform worse on child speech, which shows more variability than adult speech (Feng et al., 2024). Child speech recognition is a challenging problem, mainly because of the lack of annotated data to capture speech variability. In future research, we would like to extent the SERDA speech corpus with annotations, so that they can be used to improve child speech recognition. Based on the moderate agreement and low to moderate precision, even though the current study has shown evidence that observable and reliable scores can be generated using an ASR-based approach, we suggest against immediate usage in high-stakes settings.

While these limitations also argue against the acceptance of the scoring inference, the purpose of this validation was not to evaluate whether ASR-based oral reading metrics can be used in high stakes settings, nor whether each individual ASR prediction is correct. Instead, we evaluated whether observable ASR-based oral reading scores could be generated such that practitioners can be reliably informed about children's oral reading skills. In short, while future researcher should focus on optimizing ASR performance for this speech corpus, and ASR based on children's speech in general, the observed results were deemed sufficient to satisfy the scoring inference within more formative test settings, matching SERDA's developmental purposes (van der Velde et al., 2024b).

The investigation of the generalization inference resulted in its acceptance, suggesting that the sample of oral reading items and primary schools represent the population. However, a potential issue concerned task difficulty, as many items were shown to be relatively easy. An explanation could be found in the relatively large number of primary schools with a low school-weight. To elaborate, a low school-weight indicates that the children attending a school, on average, are expected to perform relatively well compared to the population. In other words, the prevalence of well-performing schools could have

resulted in a sample with relatively many well-performing children, for whom the items are relatively easy. Thus, although the results provide evidence that generalisations towards the general population of second and third graders is possible, further investigations are required into both the performance of the ASR, and the resulting scores, in samples of predominantly highly and less skilled oral readers.

The evaluation of the validity evidence resulted in the acceptance of the extrapolation inference, indicating that the reading tasks reflect oral reading skills. Specifically, we found that most LNIRT ability and speed estimates were strongly related to the person level WCPM metrics. With regard to the DMT and AVI, the speed estimates and WCPM metrics outperformed the ability estimates. This finding is unsurprising, given that earlier research has stressed the importance of variability in reading speed, and therefore automaticity, throughout reading development (e.g. Verhoeven & van Leeuwe, 2009), while accuracy tends to show lower variability between children as they progress through primary education. To conclude, the results support the assumption that the reading tasks measure oral reading skills, while simultaneously highlighting the importance of including the assessment of speed when using ASR to measure oral reading fluency.

The decision making inference was also justified, showing that the reading tasks allow users to make decisions regarding children's oral reading proficiency and development. However, some of the passage reading items did not have much, or any, discriminative ability. Investigations unearthed that these items primarily concerned articles (e.g. "a", "an" [*een, de, het]*), and other single-syllable words, which contain little to no orthographic complexity. Although an investigation into LNIRT model performance without these items could be of interest, this would also reduce the amount of items in the passage reading task, potentially making the passage reading task more difficult. Therefore, this consideration should be more thoroughly investigated before it is implemented.

Based on these findings, the following recommendations for further research are specified. Firstly, while the present study has demonstrated promise regarding the use of ASR to assess oral reading fluency skills for young children, researchers are advised to investigate the optimalization of ASR performance for the current speech corpus, including an evaluation of the Word Error Rate (WER). Especially interesting would be an assessment of ASR performance for samples containing primarily highly or less proficient orally fluent readers. When compared to the results of the current study, such investigations can provide a more thorough understanding of the behaviour of the ASR, and the resulting LNIRT item parameters and person estimates, potentially leading to higher ASR performances on children's speech in general. Secondly, the performance of the LNIRT model should be compared for different subsets of items, allowing for the specification of an optimal, or minimally required, set of items or subtasks. Researchers are advised to focus on the evaluation of models that exclude items with low to no discrimination ability, and on a more specific evaluation of the word and passage decoding subtasks. Finally, now that it

has been shown that oral reading metrics can be extracted from SERDA, work should focus on including and validating a prosody component.

4.5 CONCLUSION

In conclusion, the gathered evidence suggests that valid word decoding and passage reading measures can be generated for a semi-transparent language, using an ASR-based oral reading fluency assessment instrument. The results provide evidence that the reading tasks can be used to obtain observable and reliable oral reading metrics, while the samples of reading tasks and Dutch primary schools were deemed plentiful and representative enough to warrant generalizations towards their general populations. Evidence also substantiated that the reading tasks measure oral reading skills, and that the tasks allow users to make some claims and decisions regarding children's oral reading proficiency and development. However, the ASR requires optimalization and its errors require further exploration through the analysis of, for example, Word Error Rates. Furthermore, generalizations towards high and low proficiency populations should be more thoroughly investigated, allowing for comparisons of ASR performance, and LNIRT item and person characteristics behavior, such that ASR performance for children's speech can be improved. Finally, future researchers are advised to realize and validate a prosody component. If implemented correctly, these changes would complete the envisioned oral reading fluency assessment instrument, thereby improving the provision of detailed diagnostics, reducing teacher's testing burden, and improving the assessment of oral reading fluency throughout all of primary education.

REFERENCES

Aldhanhani, Z. R., & Abu-Ayyash, E. A. (2020). Theories and Research on Oral Reading Fluency: What Is Needed?. *Theory and Practice in Language Studies, 10(*4), 379–388. http://dx.doi.org/10.17507/tpls.1004.05.

Bai, Y., Hubers, F. C. W., Cucchiarini, C., & Strik, H. (2020). *ASR-Based Evaluation and Feedback for Individualized Reading Practice.* INTERSPEECH 2020: Shanghai, China. https://www.isca-archive.org/interspeech_2020/bai20b_interspeech.pdf

Bai, Y., Hubers, F. C. W., Cucchiarini, C., & Strik, H. (2021). *An ASR-based Reading Tutor for Practicing Reading Skills in the First Grade: Improving Performance through Threshold Adjustment.* IberSPEECH 2021: Valladolid, Spain. https://repository.ubn.ru.nl/bitstream/handle/2066/245151/245151.pdf.

Baker, F. B. (2001). *The basics of item response theory.* http://ericae. net/irt/baker..

Bichi, A. A., & Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. *International Journal of Evaluation and Research in Education, 7*(2), 142–151. http://doi.org/10.11591/ijere.v7i2.12900.

Boersma, P., & Weenink, D. (2024). Praat: doing phonetics by computer [Computer program]. Version 6.4.13. Retrieved 10 June 2024 from http://www.praat.org/

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013). Human and automated assessment of oral reading fluency. *Journal of Educational Psychology, 105*(4), 1142–1151. https://doi.org/10.1037/a0031479

Borgwaldt, S. R., Hellwig, F. M., & de Groot, A. M. (2004). Word-initial entropy in five languages: Letter to sound, and sound to letter. *Written Language & Literacy, 7*(2), 165–184. https://doi.org/10.1075/wll.7.2.03bor

Bray, B. & McClaskey, K. (2015). *Make learning personal. The What, Who, Wow, Where and Why.* SAGE Publications Ltd., USA.

Cheng, J. (2018). *Real-Time Scoring of an Oral Reading Assessment on Mobile Devices*. INTERSPEECH 2018: Hyderabad, India. https://doi.org/10.21437/Interspeech.2018-34.

Cheng, J., & Shen, J. (2010). *Towards accurate recognition for children's oral reading fluency*. IEEE Spoken Language Technology Workshop: Berkeley, USA. https://doi.org/10.1109/SLT.2010.5700830.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *Ieee Access, 9*, 78368–78381. https://doi.org/10.1109/ACCESS.2021.3084050.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education, 42*(3), 309–328. https://doi.org/10.1080/15391523.2010.10782553

Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: Implications for policy and practice. *Policy insights from the behavioral and brain sciences, 3*(1), 54–61. https://doi.org/10.1177/2372732215624931

Cucchiarini, C., van Hamme, H., Driesen, J., Sanders, E. (2008). THE JASMIN-CGN: *CORPUS Design, recording, transcription and structure of the corpus.*

Cucchiarini, C., Neri, A., & Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication, 51*(10), 853–863. https://doi.org/10.1016/j.specom.2009.03.003

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. *Computer Speech & Language, 84*, 101567. https://doi.org/10.1016/j.csl.2023.101567

Fox, J. P., Klotzke, K., & Simsek, A. S. (2021). LNIRT: An R package for joint modeling of response accuracy and times. *arXiv preprint arXiv:2106.10144.* https://arxiv.org/abs/2106.10144

Fuchs, L., Fuchs, D., Hosp, M., And Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scient. Stud. Read. 5*, 239–256. https://doi.org/10.1207/S1532799XSSR0503_3

Grainger, J., & Segui, J. (1990). Neighborhood frequency effects in visual word recognition: A comparison of lexical decision and masked identification latencies. *Perception & psychophysics, 47*, 191–198. https://doi.org/10.3758/BF03205983

Groen, M. A., Veenendaal, N. J., & Verhoeven, L. (2018). The role of prosody in reading comprehension: evidence from poor comprehenders. *Journal of Research in Reading, 42*(1), 37–57. https://doi.org/10.1111/1467-9817.12133

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer. https://link.springer.com/book/10.1007/978-94-017-1988-9

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

Inspectorate of Education. (2024, June 19). *Schoolweging primair onderwijs* [Schoolweightprimary education]. https://www.onderwijsinspectie.nl/trends-en-ontwikkelingen/onderwijsdata/schoolweging-po.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000

Kheir, Y. E., Ali, A., & Chowdhury, S. A. (2023). *Automatic Pronunciation Assessment--A Review. arXiv preprint arXiv:2310.13974*. https://doi.org/10.48550/arXiv.2310.13974

Kim, Y. S. G., Quinn, J. M., & Petscher, Y. (2021). What is text reading fluency and is it a predictor or an outcome of reading comprehension? A longitudinal investigation. *Developmental Psychology, 57*(5), 718–732. https://doi.org/10.1037%2Fdev0001167

Kuhn, M., Schwanenflugel, P. & Meisinger, E. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 232–253. https://doi.org/10.1598/RRQ.45.2.4

Levis, J., & Suvorov, R. (2012). Automatic speech recognition. In: Chapelle, C. A. (2012). *The encyclopedia of applied linguistics.* Hoboken, NJ : John Wiley & Sons.

van der Linden, W. J. (2007). A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika, 72*, 287–308. https://doi.org/10.1007/s11336-006-1478-z.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Loukina, A., Klebanov, B. B., Lange, P. L., Qian, Y., Gyawali, B., Madnani, N., Misra, A., Zechner, K., Wang, Z., & Sabatini, J. (2019). *Automated Estimation of Oral Reading Fluency During Summer Camp e-Book Reading with MyTurnToRead*. INTERSPEECH 2019: Graz, Austria. https://www.iscaarchive.org/interspeech_2019/loukina19_interspeech.pdf

Malmberg (n.d.). *Lijn 3 aanvankelijk lezen groep 3 basisonderwijs [Line 3 initial reading grade 1 primary education]*. Malmberg.

Mastropieri, M. A., & Scruggs, T. E. (1997). Best practices in promoting reading comprehension in students with learning disabilities 1976 to 1996. *Remedial and Special Education, 18*, 197–213. https://doi.org/10.1177/074193259701800402

Meelissen, M. R. M., Maassen, N. A. M., Gubbels, J., van Langen, A. M. L., Valk, J., Dood, C., Derks, I., In 't Zandt, M., & Wolbers, M. (2023). *Resultaten PISA-2022 in vogelvlucht [Results PISA-2022-An overview]*. Enschede: Universiteit Twente. https://doi. org/10.3990/1.9789036559461.

Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading research quarterly, 43*(4), 336–354. https://doi.org/10.1598/RRQ.43.4.2

Morris, D., & Perney, J. (2018). Using a sight word measure to predict reading fluency problems in grades 1 to 3. *Reading & Writing Quarterly, 34*(4), 338–348. https://doi.org/10.1080/10573569.2018.1446857

Morrison, T. G., & Wilcox, B. (2020). Assessing expressive oral reading fluency. *Education sciences, 10*(3), 59. https://doi.org/10.3390/educsci10030059

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., & Tobin, B. (2003). Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research, 29*, 61–117. https://doi.org/10.2190/06AX-QW99-EQ5G-RDCF

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 International Results in Reading*. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342

Perfetti, C. (1985). *Reading ability.* New York: Oxford University Press.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of Reading, 18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Pikulski, J.J., & Chard, D.J. (2005). Fluency: Bridge between decoding comprehension. *The Reading Teacher, 58*(6), 510–519. https://doi.org/10.1598/RT.58.6.2

Posit team (2023). RStudio: Integrated Development Environment for R, version 4.3.1. [Computer software] Posit Software, PBC, Boston, MA. http://www.posit.co/.

Proença, J., Celorico, D., Candeias, S., Lopes, C., & Perdigão, F. (2015). *Children's Reading Aloud Performance: A Database and Automatic Detection of Disfluencies*. INTERSPEECH 2015: Dresden, Germany. https://doi.org/10.21437/Interspeech.2015-382.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., & Perdigão, F. (2017). Automatic evaluation of reading aloud performance in children. *Speech Communication, 94*, 1–14. https://doi.org/10.1016/j.specom.2017.08.006

Reeder, K., Shapiro, J., & Wakefield, J. (2007). *The effectiveness of speech recognition technology in promoting reading proficiency and attitudes for Canadian immigrant children*. Proceedings of the 9th European Conference on Reading.

Reutzel, D. R., & Hollingsworth, P M. (1993). Effects of fluency training on second graders reading comprehension. *Journal of Educational Research, 86*, 325–331. https://doi.org/10.1080/00220671.1993.9941225

Rokade, A. A. (2018). *Automated Grading System Using Natural Language Processing*. International Conference on Inventive Communication and Computational Technologies 2018: Coimbatore, India. https://doi.org/10.1109/ICICCT.2018.8473170.

Sabu, K., & Rao, P. (2018). Automatic assessment of children's oral reading using speech recognition and prosody modeling. *CSI Transactions on ICT, 6*, 221–225. https://doi.org/10.1007/s40012-018-0202-3

Seymour, P. H. K., Aro, M., Erskine, J. M., & COST Action A8 network. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143–74. https://doi.org/10.1348/000712603321661859

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: the perils of overreliance on an" outlier" orthography. *Psychological bulletin, 134*(4), 584–615. https://doi.org/10.1037/0033-2909.134.4.584.

Silva, W. A., Carchedi, L. C., Junior, J. G., de Souza, J. V., Barrere, E., & de Souza, J. F. (2021). A framework for large-scale automatic fluency assessment. *International Journal of Distance Education Technologies, 19*(3), 70–88. https://doi.org/10.4018/IJDET.2021070105.

Smith, A. C., Monaghan, P., & Huettig, F. (2021). The effect of orthographic systems on the developing reading system: Typological and computational analyses. *Psychological Review, 128*(1), 125–159. http://dx.doi.org/10.1037/rev0000257

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., Vloedgraven, J. & De Wijs, A. (2018a). *Wetenschappelijke verantwoording DMT [Scientific Justification DMT]*. Cito: Arnhem

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., & De Wijs, A. (2018b). *Wetenschappelijke verantwoording AVI [Scientific Justification AVI]*. Cito: Arnhem.

Torgesen, J. K., Wagner, R., & Rashotte, C. (1997). Test of word reading efficiency. Austin, TX: PRO-ED.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press.

University of Oregon (2020). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Administration and Scoring Guide*. Eugene, OR: University of Oregon. https://dibels.uoregon.edu

Veenendaal, N.J., Groen, M.A. & Verhoeven, L. (2015). What speech text reading fluency can reveal about reading comprehension. *Journal of Research in Reading, 38*(3), 213–225. https://doi.org/10.1111/1467-9817.12024

Veenendaal, N.J., Groen, M.A., & Verhoeven, L. (2016). Bidirectional Relations between Text Reading Prosody and Reading Comprehension in the Upper Primary School Grades: A Longitudinal Perspective. *Scientific Studies of Reading. 20*(3), 189–202. https://doi.org/10.1080/10888438.2015.1128939

van der Velde, M., Molenaar, B., Veldkamp, B. P., Feskens, R.C.W., & Keuning, J. (2024a). What do they say? Assessment of oral reading fluency in early primary school children: A scoping review. *International Journal of Educational Research, 128*, 102444. https://doi.org/10.1016/j.ijer.2024.102444

Van der Velde, M., Veldkamp, B. P., Keuning, J., Feskens, R. C. W., Swart, N. M., Harmsen, W. N. (2024b). The framework and development of SERDA: Speech enabled reading fluency assessment for Dutch. In Ranđelović B., Karalić E., Aleksić K., Đukić D. (Eds.), *E-testing and computer-based assessment. CIDREE Yearbook 2024* (pp. 99–123). CIDREE. https://cidree.org/wp-content/uploads/2024/11/cidree_yearbook-2024.pdf

Verhoeven, L. & Van Leeuwe, J. (2009). Modeling the Growth of Word-Decoding Skills: Evidence From Dutch. *Scientific Studies of Reading, 13*(3), 205–223. https://doi.org/10.1080/10888430902851356.

Wei, X., Cucchiarini, C., van Hout, R. W. N. M., & Strik, H. (2022). Automatic Speech Recognition and Pronunciation Error Detection of Dutch Non-native Speech: cumulating speech resources in a pluricentric language. *Speech Communication, 144*, 1–9. https://doi.org/10.1016/j.specom.2022.08.004

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition, 51*, 91–103. https://doi.org/10.1016/0010-0277(94)90010-8

Wools, S., Eggen, T. J. H. M., & Sanders, P. F. (2010). Evaluation of validity and validation by means of the argument-based approach. *Cadmo 18*(1), 63–82. https://doi.org/10.3280/CAD2010-001007.

Zwijsen Educatieve Uitgeverij. (2023). *Veilig leren lezen: KIM-versie [Learning to read safely: KIM-version]*. Zwijsen.

Chapter 5

Personalizing primary reading education: Detailed reading profiles through latent modelling.

van der Velde, M., Swart, N., Veldkamp, B. P., Feskens, R.C.W. (2026). *Personalizing Primary Reading Education: Detailed Reading Profiles Through Latent Modelling* [Manuscript submitted for publication]. University of Twente.

ABSTRACT

Personalized learning is crucial to facilitate the optimal development of students following formal education. Personalization might especially benefit reading education, as reading comprehension skills have been on the decline in many countries the world over. Given that reading fluency skills are crucial for developing adequate reading comprehension skills, the early identification of children's fluency strengths and weaknesses is of the utmost importance to improve reading development. A popular approach to identify and differentiate between such characteristics concerns profiling. However, research on reading profiles oftentimes do not focus on reading fluency, and do not profile all aspects of fluency. Therefore, we conducted Latent Profile Analysis to evaluated and validated the potential of creating practically and didactically relevant reading profiles that incorporate al aspects of fluency. Profile were validated by investigating the reliability of profile allocation, their generalizability towards subpopulations, their relationships with background characteristics and relevant reading outcomes, and their potential to personalise reading instruction. Six reading fluency profiles were identified for a sample of 534 Dutch second- and third-grade students. Profiles differentiated readers based on theoretically relevant reading proficiencies and difficulties that reflect developmental stages of reading, facilitating the personalisation of reading instruction. Furthermore, profile allocation was reliable, but generalizations towards subpopulation require further investigation, as the sample sizes for subpopulations were limited. Researchers are advised to investigate the subpopulations with a larger sample, to investigate the effect of task novelty, motivation and concentration on performance, and to evaluate the developmental benefits of using the reading fluency profiles in practice.

5.1 INTRODUCTION

Personalized learning has increasingly been identified as crucial to facilitate the optimal development of students during their journey throughout formal education (e.g. Connor & Morrison, 2016; Li & Wong, 2020). This trend appears justified, as personalization has been shown to improve student's achievements, motivation, understanding, satisfaction and learning efficiency (Falcão et al., 2018; Gómez et al., 2014; Zheng et al., 2021). At the same time, understanding and suiting individual needs and preferences can be time consuming and expensive, potentially leading learning benefits to more strongly depend on resource availability and teacher competency (Xu et al., 2024). Correspondingly, tremendous work has been conducted throughout the last decade to improve the implementation of personalization within education systems, especially through the utilization of technology (Chrysafiadi & Virvou, 2015; Nandigam et al., 2014; Scott et al., 2017; Xie et al., 2019) and AI (Bhutoria, 2022; Maghsudi et al., 2021; Tapalova & Zhiyenbayeva, 2022).

Personalization might especially benefit early reading education, as illustrated by Connor & Morrison (2016), who decisively argue that many children will fail to reach their reading potential unless literacy instruction is sufficiently differentiated. Such perspectives have gained substantial weight lately, as educational focus has increasingly centred on reading, following diminishing reading comprehension performances by primary and secondary school children in many countries the world over (Meelissen et al., 2023; Mullis et al., 2023).

As early reading skills tend to be predictive of reading comprehension skills (Verhoeven & van Leeuwe, 2008), improving reading proficiency through personalization, at an early stage, is likely to improve reading comprehension performances throughout formal education. To facilitate personalization, the identification of children's reading strengths and weaknesses is of the utmost importance. A popular approach to identify combinations of such characteristics concerns the specification of reading profiles. Reading profiles are quantifications or summaries of a learner's understanding, competencies, skills and attributes (Barthakur, 2023) with respect to reading. Especially automated, detailed, data-driven profiles tend to proliferate, as they reduce the amount of manual grading, provide relatively independent and objective information, and facilitate the provision of feedback (Barthakur, 2023).

While detailed reading profiles have shown promise within the context of reading education, as exemplified through the comprehensive profiling of reading comprehension skills (e.g. Keuning et al., 2019), limited research has focused on creating detailed reading fluency profiles. Reading fluency is defined as the ability to read with accuracy, speed and proper expression (Kuhn et al., 2010; Pikulski & Chard, 2005). As reading fluency is critical for developing proficient reading skills (National Institute of Child Health and Human Development, 2000), creating detailed reading fluency profiles might improve

reading development through the early identification and remediation of reading deficits (Catts et al., 2013; Grimm et al., 2018; Virinkoski et al., 2018). Hence, the current study will explore the potential of constructing detailed reading profiles for early primary education.

5.1.1 Reading fluency and comprehension in early primary education

The importance of reading fluency to reading development is often characterized in terms of its well-established relationship with reading comprehension (Amendum et al., 2021; Fuchs et al., 2001). This relationship can be described through two processes or paths which, when combined, incorporate all aspects of reading fluency.

The first process concerns the impact of the automaticity of reading on comprehending. Automaticity is the combined mastery of quick and accurate reading, at such a level that limited cognitive resources need to be expended (Kim et al., 2021). This low cognitive load allows the reader to focus on more complex tasks, such as comprehending (Aldhanhani & Abu-Ayyash, 2020; Morris & Perney, 2018). The second process exists through prosody, which reflects expressive aspects of reading (Miller & Schwanenflugel, 2008). Prosody is related to comprehension (Veenendaal et al., 2016) through the facilitation of attaining and maintaining meaning during the reading process (Cowie et al., 2002; Miller & Schwanenflugel, 2006). The interaction between these processes is relatively intuitive. As reading becomes more automatic, cognitive resources are freed up, allowing readers to focus on prosody and comprehension, both of which improve comprehension. This argumentation finds substantiation in earlier findings, which indicate that prosody and comprehension impact each other throughout primary education, even when controlling for automaticity (Groen et al., 2018; Veenendaal et al., 2015).

Although the importance of reading fluency is firmly ingrained within early reading literature, its assessment currently makes it difficult to differentiate instruction based on performance. To elaborate, most reading fluency assessment instruments only provide information on reading accuracy, speed and/or automaticity (van der Velde et al., 2024a), while prosody is often not incorporated or independently assessed using relatively subjective rating scales (Kuhn et al., 2010; Morrison & Wilcox, 2020). Furthermore, popular instruments such as the Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DORF; University of Oregon, 2020) and Gray Oral Reading Test-5 (GORT-5; Wiederholt & Bryant, 2012) only provide task-level information. This limits the amount, and level of detail, of the information available to practitioners, limiting their ability to differentiate instruction and to meet the individual needs of students learning to read.

In short, although developing fluent reading skills is essential, most reading fluency assessment instruments only provide reading diagnostics at the task level for a limited set of reading fluency components, a state of affairs that might explain the lack of research on detailed reading fluency profiles.

5.1.2 Profiling reading fluency performance

Profiling student's oral reading performance to help teachers individualize instruction has a long history within educational contexts. One early example concerns work by Argyle (1989), who described a method to record and analyse mistakes during oral reading tasks. More recently, research has focused on investigating the development and stability of reading profiles (Foorman et al., 2017; Psyridou et al., 2021; Risberg et al., 2024), and on comparing profiles of proficient readers to struggling readers (Dams et al., 2023; Grimm et al., 2018), readers with dyslexia (Miciak et al., 2022), and second language learners (Kim, 2024). In short, researchers have created reading profiles for many purposes, incorporating many different combinations of reading related skills.

However, the focus of most research is not on reading fluency performance. Instead, studies frequently create profiles using an array of reading performances. Oftentimes, the goal of such studies is the prediction of other reading competencies or outcomes (e.g. Foorman et al., 2017; Miciak et al., 2022). As such approaches provide little to no guidance for teachers to personalize their reading fluency instruction, specifically, the current study will aim to fill this gap by constructing profiles that solely reflect reading fluency skills.

Another issue, reflective of current assessment practices, concerns the limited level of detailed diagnostics used through the profiling of reading fluency. To elaborate, among the profiling literature discussed, no study has incorporated individual diagnostics for all reading fluency components. Indeed, prosody has generally been abandoned throughout reading fluency profiling practice, even though knowledge on all relevant domain competencies has previously been demonstrated to guide differentiation (Keuning et al., 2019).

Given the state of reading fluency assessment, novel reading fluency assessment instruments might be needed to supply detailed diagnostics (e.g. van der Velde et al., 2024b; 2025). As their usage is yet to be proven, however, evaluating the potential of generating detailed reading fluency profiles and using them to personalize reading instruction requires a thorough evaluation and validation.

5.1.3 The present study

Although the relevance of reading fluency to reading development has been well-established, limited research has focused on personalizing reading instruction based on the profiling of detailed diagnostics for all reading fluency components. Given that novel reading fluency assessment instruments are able to supply detailed reading fluency diagnostics for all relevant components, the current study will focus on evaluating and

validating the potential of creating reading fluency profiles that incorporate accuracy, speed and prosody. Specifically, the current study will investigate the following:

1. Can distinct, practically and didactically relevant reading fluency profiles be distinguished for a sample of early primary school children?
2. What reading fluency characteristics characterize the optimal set of profiles and differences between profiles?
3. How stable are the reading fluency profiles observed across different subpopulations?
4. How do the optimal set of profiles relate to relevant background characteristics and reading results?
5. Can the optimal set of profiles help guide personalized instruction?

5.2 METHODS

To evaluate whether detailed reading fluency profiles can be created, and used to personalize reading instruction, we administered the reading tasks of The Speech Enabled Reading Diagnostics App (SERDA; van der Velde et al., 2024b). To relate the profiles to relevant reading results, children's most recent scores on standardized word and passage reading tasks, the Three Minute Task [Drie-minuten-toets; DMT] (van Til et al. 2018a) and AVI [Analyse van Individualiseringsvormen; AVI] (van Til et al. 2018b), were also obtained. Finally, background characteristics were collected to improve profile interpretation and validation through a questionnaire filled in by the parents of the participants.

5.2.1 Participants

653 Dutch second- and third-grade primary school children participated in the current study. Participants were equally distributed over the second and third grade of primary education. Second- and third-grade participants were respectively, on average, seven and eight years of age, showing a sample mean of 7.6 (SD = 0.74). The sample included slightly more girls (52%) than boys.

Participants came from 19 regular primary schools in the Netherlands. Data was collected between October and February of schoolyears 2022-2023 and 2023-2024. Schools were randomly selected from all Dutch primary schools with the aim of representing children of different socio-economic backgrounds and dialectical regions. Within these schools, all second- and third-grade children were included in the study, including children with expected dyslexia, given that their parents provided written approval.

5.2.2 Materials

5.2.2.1 *SERDA: Reading fluency.*

SERDA is a Dutch reading fluency assessment instrument that allows for the evaluation of children's spoken reading fluency skills (see van der Velde et al., 2024b). SERDA consists of a word decoding and passage reading task, both containing three subtasks of increasing difficulty and complexity. During the word decoding task, children were instructed to read aloud three sets of 50 words, which were presented using a progressive demasking design (Grainger & Segui, 1990) to allow for accurate reading speed estimation. During the passage reading task, children were instructed to read aloud three passages of around 175 words, as quickly and accurately as possible. Both the word decoding and passage reading task were individually administered on a tablet, using a microphone to capture children's speech.

Information on children's fluency skills was extracted at the subtask and task level. Throughout this study, we primarily focused on profiling passage reading, as assessing prosodic features for word decoding was more complicated. Thus, we extracted task and subtask level diagnostics on participants' passage reading accuracy, speed, automaticity and prosody. For the purpose of modelling differences between good word decoders and passage readers, we also extracted the task and subtask automaticity diagnostics for the word decoding task. For a more thorough description of the extraction and validation of SERDA's metrics, see van der Velde et al. (2024b, 2025) and Harmsen et al. (2025).

5.2.2.2 *Accuracy, speed, automaticity.*

Accuracy, speed and automaticity estimates were derived for each completed subtask. For the accuracy and speed metrics, we applied the methodology described by van der Velde et al. (2025) to SERDA's subtasks. We used an Item Response Theory framework (Hambleton & Swaminathan, 1985) to determine subtask-level accuracy and speed estimates for each participant, based on their item-level accuracy and speed scores. Specifically, we utilized the Hierarchical Bayesian joint modelling approach (van der Linden, 2007), using the "LNIRT" (Fox et al., 2021) R-package. Automaticity was calculated as the number of words read correctly per minute per subtask. Task level metrics were followingly calculated by averaging over subtasks. Finally, we determined the variability of children's reading performances by calculating standard deviations (SD) over their accuracy, speed and automaticity subtask scores.

5.2.2.3 *Prosody.*

For each passage reading subtask, we estimated participants' speech-rate, articulation-rate, average duration of pauses, average pitch and average loudness. Here, speech-rate concerned the number of words read per minute over the entire audio file. Articulation rate concerned

the number of words per minute of voiced speech. Pauses concerned the average duration of pauses in seconds. These measures were calculated using the methodology discussed and validated by Harmsen et al. (2025), utilizing Whisper Timestamped with disfluency detection (Louradour, 2023). Meanwhile, average pitch and loudness were calculated using geMAPS (Eyben et al., 2015), thereby respectively reflecting the average F0 contours in semitones, and the average signal strength over the entire audio file. For pauses, pitch and loudness, standard deviations were calculated within each subtask to provide an indication of variability within the subtask. Task level performance and variability metrics were then determined by calculating averages over all completed subtasks.

5.2.2.4 *DMT: Word decoding.*

The DMT is a grade-normed reading examination, used to evaluate the word decoding development of Dutch primary school children (van Til et al., 2018a). During the DMT, children read word lists of increasing difficulty as accurately and quickly as possible for 1 minute per list. Based on the number of words read correctly per minute, children were provided with a normed, grade-specific score that reflects their word decoding level compared to other children in their grade. Specifically, children were placed in one of five categories, ranging from the 20% best to the 20% least proficient decoders.

5.2.2.5 *AVI: Passage decoding.*

The AVI is a grade-normed reading examination, used to evaluate the passage reading development of Dutch primary school children (van Til et al., 2018b). During the AVI, children read passages of increasing difficulty until they make too many mistakes or do not read quickly enough anymore. Based on the most difficult passage successfully completed, children were provided with a normed, grade-specific score that reflected their passage reading level compared to other children in their grade. Specifically, children were placed into one of twelve levels, which reflect the reading level expected at different primary school grades.

5.2.3 Background characteristics

To evaluate what background characteristics describe the generated reading profiles, and to further improve profile interpretability, we obtained children's grade, gender, information regarding expected dyslexia, and the languages they speak at home.

5.2.3.1 *Expected dyslexia.*

Parents were asked whether they expected that their child had dyslexia. Parents could indicate this to either be the case or not. While not as reliable as an official diagnosis, in the Netherlands the procedure to formally diagnose dyslexia is generally not initiated until

halfway through second grade. Further issues exist regarding privacy, which lead to the preference of expected dyslexia over a formal dyslexia diagnosis.

5.2.3.2 *Languages at home.*

Parents were asked what language, or languages, children spoke at home. Based on their response, children were categorized in one of two groups. The first group concerns children who only speak Dutch at home, while the second concerns children who speak multiple languages at home, or who do not speak Dutch at home.

5.2.4 Data analysis

Data analysis was separated into four steps. First, data was prepared for profiling analyses. Then, the optimal number of reading fluency profiles was determined. Afterwards, the optimal set of profiles was interpreted based on theoretical and analytical grounds. Finally, the validity of the optimal set of profiles was evaluated. Analyses were conducted in R-studio (Posit Team, 2025), using the Mclust and tidyLPA packages (Rosenberg et al., 2018; Scrucca et al., 2023).

5.2.4.1 *Data preparation.*

Before conducting LPA, data was cleaned. We removed observations with missing subtask scores, as their inclusion might bias the performance and variability metrics, given the gradual increase in difficulty over subtasks. No relevant differences were found in fluency features between children with complete data and missing subtask scores. Followingly, we removed extreme univariate outliers, defined as observations distanced more than three times the interquartile distance below the first or above the third quartile. Imputing extreme univariate outliers was attempted using the MICE R-package (van Buuren & Groothuis-Oudshoorn, 2011), but resulted in negligible differences in profiling choices and solutions. Thus, to reduce complexity, we only used complete data. Then, we investigated multicollinearity, removing the least theoretically and analytically relevant variable, if extreme correlations (> 0.9) were observed. Finally, we scaled all data before conducting LPA.

5.2.4.2 *Number of profiles.*

The optimal number of distinct reading fluency profiles was identified through Latent Profile Analysis (LPA). LPA concerns a latent variable modelling approach that groups individuals into mutually exclusive classes, based on combinations or patterns of observations from a set of variables (Collins & Lanza, 2013; Foorman et al., 2017). One of the benefits of LPA concerns the ability to evaluate many statistical fit metrics, allowing the users to identify an optimal model based on statistical, theoretical and content-related grounds (Spurk et al., 2020). Correspondingly, the optimal number of profiles was identified

by evaluating fit metrics, profile sizes, the distinctness of profiles, and the theoretical and practical utility of profiles. To obtain an optimal modelling approach, we compared Akaike Information Criterion (AIC), the Consistent AIC (CAIC), the Bayesian Information Criterion (BIC) and the sample size adjusted BIC (SABIC) for models that varied regarding their numbers of profiles (1:10) and variance and covariance constraints. Specifically, we evaluated models with equal variances and covariances fixed to zero (model 1), varying variances and covariances fixed to zero (model 2), equal variances and covariances (model 3), and varying variances and covariances between profiles (model 4).

For the optimal modelling approach, we evaluated the optimal number of profiles. This evaluation was conducted hierarchically, moving from the simplest solution containing the fewest profiles, to the most complex solution. For these purposes we used Bootstrapped Likelihood Ratio Tests (BLRT) to compare solutions of interest to solutions with one additional profile. Here, profile discrimination was explicitly taken into account by showing a preference for simpler solutions, excluding sets of profiles that include very small ($n < 25$), uninformative or poorly distinguishable profiles. Based on these qualities, the best analytical solutions were evaluated in depth, leading to the choice of the most theoretically and didactically relevant solution.

5.2.4.3 *Profile interpretation.*

Following their determination, the optimal set of profiles was interpreted, for which an expert on language and didactics was consulted. In essence, profiles were interpreted based on relevant differences in reading performance, performance variability and prosodic features, within and across profiles. Thereafter, profiles were named and provided with a short description that captures their essence.

5.2.4.4 *Profile validation.*

The final step within our data analysis concerned the validation of the optimal set of profiles. Our approach takes inspiration from the LPA "how to" guide (Spurk et al., 2020), the similarity analysis for latent profile solutions (Morin et al., 2015) and the argument-based approach to validation (Kane, 2012). The goal of this validation was to evaluate whether distinct, practically and didactically relevant reading fluency profiles could be generated for a sample of second- and third-graders.

The validation consisted of four steps. Firstly, we evaluated the reliability of cluster allocation by calculating posterior class probabilities, and the entropy (Celeux & Soromenho, 1996), of the optimal solution.

Secondly, we evaluated the generalizability of the profile solution by investigating LPA results for relevant subpopulations. Specifically, we evaluated whether separate LPA's for second- and third-grade children, and for girls and boys, would results in the same profiles as the optimal solution for all data, with regard to content and size.

Thirdly, we evaluated whether the profiles reflect reading fluency skills and development. Here, we investigated the relationship between profile membership, relevant background characteristics (grade, gender, expected dyslexia, and languages spoken at home), and relevant reading outcomes (DMT and AVI), using chi-squared tests of independence.

Given the large number of distinct AVI classifications (12+), and the limited number of observations within the highest and lowest performance groups, we constructed three proficiency classes based on the AVI classes. Namely, children showing AVI scores below the second primary grade were coded as "Less Proficient", while children with AVI scores above the third grade were coded as "Highly Proficient". Then, Children with AVI scores at the level of the second or third grade were coded as "Averagely Proficient".

Finally, we evaluated whether the profiles could be used to differentiate between readers with different types of reading proficiencies and difficulties, allowing practitioners to individualize their reading instruction. Therefore, we evaluate the theoretical relationship between the profiles and well-established developmental stages of reading fluency in alphabetic orthographies, which represent the most prevalent writing systems globally (Caravolas, 2022).

5.3 RESULTS

5.3.1 Descriptives

After data preparation, 534 participants were retained for profiling analyses, which is sufficient for conducting LPA (Spurk et al., 2020). Table 1 provides an overview of these participants, as well as frequency-distributions for relevant background characteristics. The multicollinearity analyses led to the removal of the speech- and articulation rate variables, as these highly correlated with one-another, and with the speed and automaticity metrics.

Table 1. Background characteristics for study participants.

	Grade			Gender		Age				Dyslexia		Language	
Participants	**2**	**3**	**2/3**	**Female**	**Male**	**<7**	**7**	**8**	**9+**	**No**	**Yes**	**NL**	**Multi**
N	242	264	28	278	256	20	180	237	44	492	28	400	133
%	45	49	5	52	48	4	37	49	9	95	5	75	25

Note. 2/3 = children in a combined class. Dyslexia = whether parents expected their child to have dyslexia. NL = only speak Dutch at home, Multi = Multiple languages are spoken at home.

5.3.2 Number of profiles

The optimal set of profiles was identified by comparing fit metrics, profile sizes, likelihood ratio tests, profile distinctness, and the theoretical and practical utility of profiles, resulting from LPA.

Figure 1 shows an overview of the AIC, CAIC, BIC and the SABIC for 40 models that differed with regard to the number of profiles, and their assumptions concerning variances and covariances. For each fit metric, the model with varying variances and covariances (model 4) was identified as the optimal solution. Within this modelling approach, the optimal number of profiles varied between two and nine based on fit metric. As none of these solutions contained fewer than 25 observations (5%), all eight solutions were deemed eligible for evaluation.

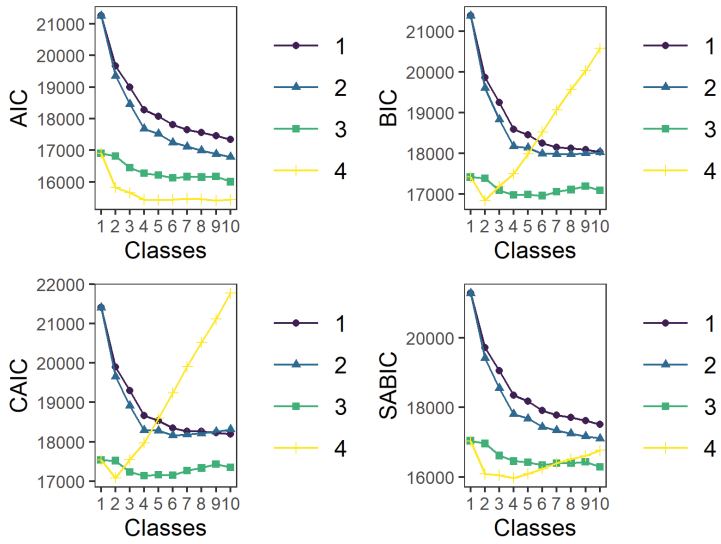


Figure 1. *Fit Metrics for Models Containing one to ten Classes, With Variances and Covariances set to 0 (1), Varying Variances (2), Equal Variances and Covariances (3) and Varying Variances and Covariances (4).*

Based on these findings, we evaluated the content of the profiles for solutions containing two to eight profiles based on their distinctness, the added value of the additional profile, and the theoretical and practical usefulness of the solution. To guide these decisions, we conducted Bootstrapped Likelihood Ratio Tests. As the likelihood ratio tests indicated that model performance no longer increases when a sixth profile is added ($p = 0.30$), we were especially strict when evaluating the added value provided by solutions containing more than five profiles. The solution containing six profiles was identified as the ideal set of profiles for the following reasons.

When moving from two to five profiles, the information provided developed from reflecting differences in general reading performance to reflecting nuanced differences in accuracy, speed and word and passage reading automaticity. In addition, relevant differences were found for prosodic features within the five-profile solution, especially regarding differences in pauses and loudness. However, the five-profile solution did not differentiate well between children who performed well on all tasks, while the six-profile solution did.

Specifically, the six-profile solution split up the most proficient readers into two profiles, differentiating them based on prosodic features. The first profile concerned readers with very high performance on all tasks, and a consistent high pitch and loudness, reflecting a lack of expression. Meanwhile, the second profile concerned readers with above average to high performance on all tasks, but with lower pitch and loudness and high variation in pitch, reflecting attention to expressiveness. Given that fluent readers can theoretically be separated from solely accurate and quick readers, the sixth profile was deemed sufficiently relevant to warrant inclusion. As the seventh profile did not provide a similar improvement, the six-profile solution was deemed optimal.

5.3.3 Profile interpretation

Each profile was interpreted and named based on differences in performance metrics, variability metrics, prosodic features and the combination of all metrics. Figure 2 shows the averages and confidence intervals for all variables included within the LPA, for each class of the optimal solution. Likewise, Table 2 and Table 3 show the scaled average and variability scores of each variable included within the LPA for each class of the optimal solution.

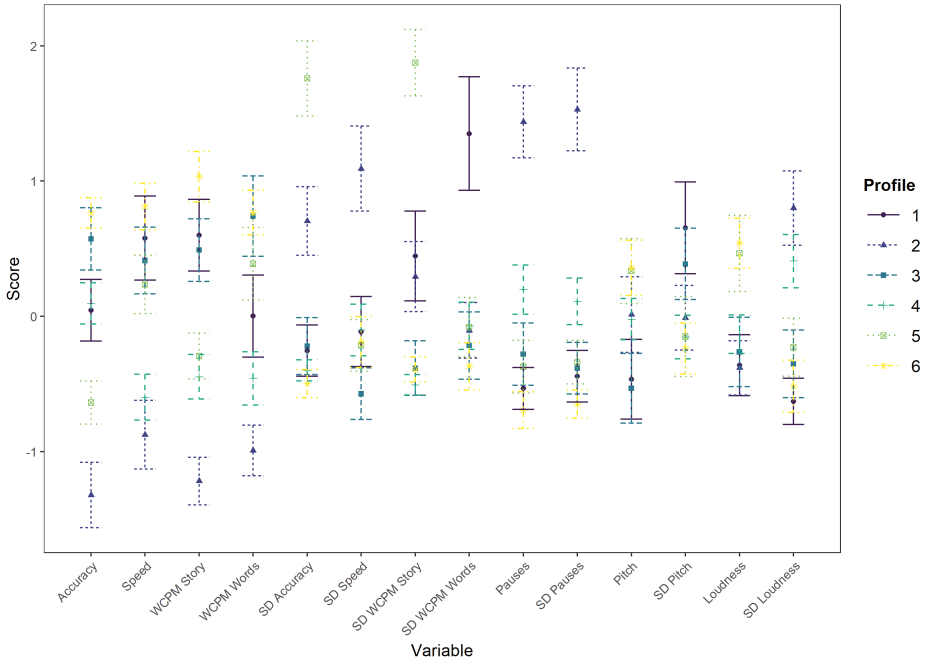


Figure 2. *Scaled Averages and 95% Confidence Intervals of Each Variable Included Within the LPA for Each Profile of the Optimal Solution.*

Note. WCPM = Words Correct Per Minute, SD = standard deviation.

Table 2. Average Scaled Scores of Each Variable Included Within the LPA for Each Profile of the Optimal Solution.

Profile	Accuracy	Speed	WCPM Story	WCPM Words	Pauses	Pitch	Loudness
1	0.07	0.61	0.62	0.01	-0.54	-0.49	-0.36
2	-1.33	-0.88	-1.23	-0.99	1.46	0.01	-0.38
3	0.57	0.40	0.48	0.71	-0.27	-0.53	-0.26
4	0.08	-0.62	-0.46	-0.49	0.20	-0.01	-0.14
5	-0.64	0.24	-0.29	0.39	-0.38	0.33	0.46
6	0.76	0.81	1.02	0.77	-0.69	0.36	0.54

Note. WCPM = Words Correct per Minute.

Table 3. Scaled Variability Metrics of Each Variable Included Within the LPA for Each Profile of the Optimal Solution.

Profile	Accuracy	Speed	WCPM Story	WCPM Words	Pauses	Pitch	Loudness
1	-0.25	-0.11	0.47	1.38	-0.45	0.67	-0.62
2	0.7	1.08	0.29	-0.11	1.54	0	0.80
3	-0.23	-0.58	-0.41	-0.22	-0.38	0.37	-0.36
4	-0.40	-0.10	-0.50	-0.05	0.11	-0.16	0.42
5	1.77	-0.21	1.88	-0.08	-0.34	-0.15	-0.23
6	-0.50	-0.17	-0.39	-0.37	-0.65	-0.24	-0.51

Note. WCPM = Words Correct per Minute.

5.3.3.1 *Profile 1: Context readers.*

The first profile [Context readers] contains 58 (11%) children and is primarily characterized by children that performed better while reading passages [WCPM Story], compared to reading isolated words [WCPM Words]. Correspondingly, their word decoding performance showed substantially higher variability than their passage reading variability [SD WCPM Words and SD WCPM Story]. Another finding concerns the high reading speed for the stories, relative to their average passage reading accuracy [Accuracy and Speed]. Regarding prosody, children in profile one show short and stable pauses, low pitch and low loudness [Pauses, SD Pauses, Pitch, Loudness]. While their loudness throughout the tasks is stable, participants in profile one tend to show a lot of variability in pitch, likely indicating that they attend to their expressiveness [SD Pitch and SD Loudness]. Based on these findings, profile one has been named "Context Readers", reflecting the relative ease with which these children read passages compared to reading isolated words. Children within this group appear to compensate for the difficulties they experience while reading isolated words by using the context provided within stories.

5.3.3.2 *Profile 2: Learning readers.*

The second profile [Learning readers] contains 83 (16%) children and is primarily characterized by relatively low accuracy, speed and automaticity, as well as high performance variability, for both the word decoding and passage reading tasks. Regarding prosody, children in profile two show lengthy, instable pauses and average pitch. While children in profile two are the least loud, they show the most variability in their loudness, potentially reflecting a lack of confidence or motivation. Based on these findings, profile two has been named "Learning Readers", reflecting their general difficulty with reading words and text.

5.3.3.3 *Profile 3: Fluent readers.*

The third profile [Fluent Readers] contains 61 (11%) children and is primarily characterized by average to high performance on all performance metrics with relatively low variability in performance. Regarding prosody, children in profile three show short and stable pauses, low pitch and low loudness. While their loudness is stable, participants in profile three tend to show a lot of variability in pitch, indicating that they attend to their expressiveness. Based on these findings, profile three has been named "Fluent Readers", reflecting stable average to high performance while attending to their expression.

5.3.3.4 *Profile 4: Accurate readers.*

The fourth profile [Accurate Readers] contains 157 (29%) children and is primarily characterized by relatively low reading speed, despite showing average accuracy. Likewise, they showed relatively low automaticity for both the word decoding and passage reading tasks. For all performance metrics, variability was low, indicating that the results are quite stable. Regarding prosody, children in profile four show relatively large pauses with high variability. Pitch is average, with relatively low variability, while loudness is relatively low and quite variable, suggesting that these children do attempt to attend to their expressiveness. Based on these findings, profile four has been named "Accurate Readers", reflecting the relative ease with which they read words accurately.

5.3.3.5 *Profile 5: Quick readers.*

The fifth profile [Quick Readers] contains 53 (10%) children and is primarily characterized by relatively low accuracy of reading, and relatively good word decoding automaticity. While children within profile five read passages quite quickly, their accuracy is relatively low and variable. As a result, their passage reading automaticity is below average, even though their word decoding automaticity is quite high. Regarding prosody, children in profile five show stable and short pauses. In addition, their pitch and loudness are quite high and relatively stable. Based on these findings, profile five is named "Quick Readers", reflecting their relatively high reading speed.

5.3.3.6 *Profile 6: Technical readers.*

The sixth profile [Technical Readers] contains 122 (23%) children and is primarily characterized by high performance and low variability on all tasks, but seemingly limited attention to prosody. Regarding prosody, pauses are very short and stable. However, pitch and loudness are quite high, with low variability, reflecting loud monotonous speech with little expressiveness. Based on these findings, profile six is named "Technical Readers", reflecting their high performance and stability, and limited regard for expressiveness.

5.3.4 Profile validation

To evaluate the reliability of cluster allocation, we calculated the average posterior class probabilities for each cluster and the entropy. Posterior class probabilities varied between 0.94 and 0.99, while entropy was 0.94. Both findings indicate high classification certainty, or high reliability of cluster allocation.

Then, to evaluate the generalizability of the optimal solution across relevant subpopulations, we evaluated whether the same profiles are extracted when conducting LPA separately for second- and third-graders, and for boys and girls. As the corresponding samples were half the size of the full dataset, the optimal modelling approach (model 4) did not converge. Therefore, the six-profile solution was specified using model 3 instead. Table 4 provides an overview of the profiles obtained for the solutions of the entire dataset and the subpopulations, as well as their proportions, ordered by developmental stage. An overview of the scaled average and variability scores for the LPA solutions of separate subpopulations is provided in Appendix A.

Table 4. Profiles, ordered based on developmental stage, and their relative sizes for the lpa solution for all data, and the separate solutions for second- and third-graders, boys and girls.

All Data	Grade 2	Grade 3	Boys	Girls
Learning: 16%	Learning: 14%	Learning: 5%	Learning: 8%	Learning: 9%
Accurate: 29%	Accurate: 26%	Accurate: 28%	Learning 2: 18%	Accurate: 22%
Quick: 10%	Quick: 15%	Quick: 17%	Accurate: 32%	Accurate 2: 10%
Context: 11%	Context: 8%	Average: 31%	Quick: 16%	Quick: 14%
Technical: 23%	Technical: 12%	Technical: 9%	Fluent Context: 7%	Average: 36%
Fluent: 11%	Fluent: 25%	Fluent: 9%	Technical: 19%	Technical: 9%

The results indicate that most profiles generalize to relevant subpopulations. However, second-graders showed lower performance and stability compared to third-graders. Furthermore, compared to the solution containing all data, children in second grade showed a higher proportion of Fluent Readers and a lower proportion of Technical Readers. In addition, context readers showed lower accuracy, and less severe differences in automaticity performance. Finally, pitch was observed to be somewhat higher for Fluent Readers.

The LPA solution for children in third grade showed relatively few Learning Readers and Technical Readers compared to the solution for all children. However, the main difference concerns the absence of a clear Context Reader profile. Instead, a profile was observed containing children with average performances across the board, termed “Average Readers”. While differences in automaticity performance were observed, these were reflected within the Fluent and Technical Reader profiles, showing higher variability in word decoding performance for especially Fluent Readers.

Meanwhile, the LPA's for boys and girls showed differences in prosodic features, showing higher pitch and lower loudness for girls. Compared to the solution containing all data, the LPA for boys did not contain Context and Fluent Reader profiles. Instead, these profiles were reflected within a singular profile, termed "Fluent Context Readers". To elaborate, Fluent Context Readers showed average to high performance and low variability on the passage reading metrics, but lower performance and very high variability on the word decoding tasks. The severe variability of word decoding observed within this group might indicate that the boys got bored or lost motivation during the word decoding task. In addition, a second Learning Reader profile was observed, containing less severely low and more stable performances. Finally, Quick Readers showed somewhat higher pitch and loudness, while Accurate Readers showed higher pitch.

For girls, we did not observe a Context or Fluent Reader profile. Instead, two Accurate Reader groups were observed, where one contained lower performance across the board and long, instable pauses, while the other contained high pitch, loudness and above average variability in pitch and loudness. Furthermore, girls showed an Average Reader profile, reflecting performances resembling Fluent Readers without a focus on expressiveness.

To evaluate whether the profiles reflect reading fluency skills and development, we conducted chi-square tests of independence to evaluate the relationship between profile membership, relevant background characteristics, and relevant reading results.

The chi-square test of independence showed a significant relationship between profile membership and Grade X^2 (5, n = 506) = 53.5, $p < 0.001$, Expected Dyslexia X^2 (5, n = 534) = 30.3, $p < 0.001$, the DMT classifications X^2 (20, n = 534) = 137.2, $p < 0.001$, and the AVI proficiency classes X^2 (10, n = 534) = 194, $p < 0.001$. Figure 3 shows balloon plots, visualizing the dependencies between profile membership, grade, expected dyslexia and the DMT and AVI classifications. The chi-square test of independence showed no significant relationship between profile membership and gender X^2 (5, n = 534) = 3.2, p = 0.67, nor Languages spoken at home X^2 (5, n = 534) = 2.9, p = 0.71.

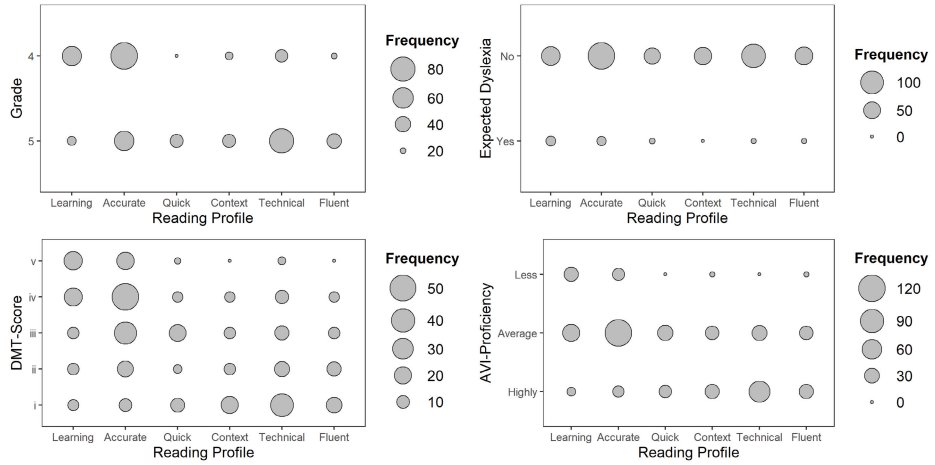


Figure 3. *Balloon Plots Showing the Distribution Of Reading Profiles for Grades, Expected Dyslexia, the DMT Scores and the AVI Proficiency Classes.*

Finally, we evaluated whether the profiles differentiate between children with various reading proficiencies and difficulties by evaluating their theoretical relationship with well-established developmental stages of reading fluency in alphabetic orthographies.

In the earliest phase, often described as the logographic or phonological recoding stage (Ehri, 1995; Share, 1995), reading is characterized by slow and effortful decoding. Here, children consciously retrieve phonological representations of letters, leading to increased cognitive load and a higher likelihood of errors (Ziegler & Goswami, 2005). As reading experience accumulates, children transition into more efficient decoding strategies. Correspondingly, automaticity begins to develop, reducing the mental effort required for grapheme-phoneme mapping, allowing for faster, more accurate word recognition (LaBerge & Samuels, 1974; Kuhn & Stahl, 2003).

Once decoding becomes largely automatized, cognitive resources can be reallocated to higher-order processes such as prosody and comprehension, which marks the onset of fluent reading (Wolf & Katzir-Cohen, 2001). Fluency, then, is not merely a matter of speed, but a multifaceted construct involving accuracy, rate, and expressive reading (Rasinski, 2012). However, not all children smoothly transition into fluent readers. Some struggle with acquiring consistent grapheme-phoneme correspondences, resulting in persistent inaccuracies, while others decode accurately but slowly, reflecting a lack of automaticity. Still others may decode fluently but fail to integrate prosodic features effectively, which impedes comprehension (Nation, 2005).

These diverse developmental trajectories are echoed in the six distinct profiles identified in our current study, each capturing a unique constellation of strengths and difficulties across the core components of reading fluency. Therefore, we argue that the profiles indeed

differentiate between children based on relevant reading proficiencies and difficulties, allowing practitioners to individualize their reading instruction.

5.4 DISCUSSION

The aim of the current study concerned evaluating and validating the potential of creating distinct, practically and didactically relevant reading fluency profiles that reflect aspects of accuracy, speed and prosody for a sample of second- and third-grade students, using LPA. We first determined, interpreted and named the optimal set of profiles based on analytical and theoretical grounds. Then, the optimal set of profiles was validated through combining perspectives from the "how to" guide for LPA (Spurk et al., 2020), the similarity analysis for latent profile solutions (Morin et al., 2015) and the Argument Based Approach to Validation (Kane, 2012).

The optimal solution contained six profiles, with freely estimated variances and covariances per profile. To elaborate, we identified children as Learning Readers, Accurate Readers, Quick Readers, Context Readers, Technical Readers and Fluent Readers. In essence, Learning Readers had relatively low reading accuracy and speed, while Accurate and Quick Readers, respectively, had relatively low reading speed and accuracy. Meanwhile, Context Readers read passages better than isolated words, while Technical Readers showed high performance on all tasks, but showed limited expressiveness. Finally, Fluent Readers showed relatively high performance on all tasks, but also showed attention to prosodic aspects.

To validate the optimal solution, we evaluated the stability of profile estimation, the stability of results for relevant subpopulations, whether the profiles reflect reading fluency skills and development, and whether the profiles differentiate between readers with different reading proficiencies and difficulties.

Based on the results of the validation, profile allocation was deemed reliable. Namely, high posterior probabilities were observed, as well as high entropy. While a limited number of children showed lower posterior probabilities, such cased are easily identified and dealt with through a more detailed inspection of their performance. Therefore, it can be concluded that profile allocation was stable.

The results also indicate that most profiles generalize to relevant subpopulations. Especially the Learning, Accurate, Quick and Technical Reader profiles were deemed generalizable, as these occurred within each solution. However, the Context Reader profile was often not separately observed, being incorporated within alternative reading profiles instead. Additionally, the solution for second-graders showed relatively many Fluent Readers, while third graders showed a limited number of Learning Readers. Furthermore,

the solutions of boys and girls did not contain a clear Fluent Reader profile. As these differences complicate generalizations, their origins require discussion.

Firstly, due to the limited number of observations for the separate LPA's, the optimal modelling approach did not converge, leading to the specification of a simpler model. While this is not surprising, given that the solution containing all children only barely contained a sufficiently large sample to conduct LPA, caution is warranted when interpreting or using the subpopulation solutions. To evaluate the impact of this computational shortcoming, we recommend validating the results with a sample of at least 500 observations per subpopulation (Spurk et al., 2020) before they are considered for use in practice.

Secondly, we argue that the subpopulations of interest are expected to differ. For example, second-graders are not expected to perform as well as third-graders. Therefore, second-graders that read relatively well and somewhat expressively are expected to be identified more easily when solely compared to second-graders, while children that read very well are less prevalent, potentially explaining the high prevalence of Fluent Readers and low prevalence of Technical Readers within this solution. Likewise, third-graders tend to have more reading experience and proficiency, possibly explaining their low proportion of Learning Readers. For boys and girls, solutional differences can mainly be attributed to a relative lack in prosodic variability. To elaborate, boys tend to speak relatively similarly to other boys, and relatively dissimilarly to girls, showing lower pitch and higher loudness. This lack in variability complicates the modelling of prosodic features, leading to the absence of a clear Fluent Reader profile, which primarily reflects prosodic elements.

Based on these findings, we conclude that the profiles are likely subpopulation dependent, explaining the differences in solutions. However, as the samples of the subpopulations were not sufficiently large to warrant their usage, we suggest to focus on utilizing the solution for all data, allowing for the nuanced modelling of all relevant reading fluency components.

The validation also substantiated that the optimal set of profiles represent fluent reading skills and development. Namely, we observed significant relationships between profile membership and grade, expected dyslexia and children's performances on the DMT and AVI. Specifically, profiles representing lower developmental stages mostly consisted of second-graders, while higher order profiles primarily contained third-graders. In addition, children expected to suffer from dyslexia were almost solely observed within profiles that reflect their well-documented difficulties with reading speed (e.g. Wagner et al., 2022). Furthermore, DMT and AVI performances were observed to increase based on the developmental stages reflected within the profiles.

However, Context Readers, characterized by low word decoding performances, performed well on the DMT. Therefore, their low word decoding performance might reflect the novelty of, or their difficulties with, SERDA's decoding tasks, specifically, potentially explaining their unstable word decoding performances. Substantiation for this claim was

found during the development of SERDA, where children complained about the length of the word decoding task, reporting issues regarding concentration (van der Velde et al., 2024b). As insufficient information is currently available regarding children's experiences with SERDA, we argue that future research should investigate the impact of children's motivation and concentration on performance.

Finally, the validation provides evidence that the profiles differentiate between different types of readers, reflecting relevant reading proficiencies and difficulties. To elaborate, the six profiles can be interpreted to describe children who have difficulty with reading in general, children who read relatively accurately but with limited speed, children who read relatively quickly, but not accurately, children who read passages well, but have difficulty with word decoding, children who read well but not expressively, and children who read well and expressively. As these profile interpretations are both simple and clear, they allow for relatively concrete suggestions regarding the personalization of reading instruction. To exemplify, Accurate Readers can specifically be supported to read more quickly, while Quick Readers can be aided in improving their accuracy. Likewise, Technical Readers can be taught how to read more expressively through focussing instruction on their variability in pitch. Given that the profiles reflect reading strengths and weaknesses that strongly resemble theoretical stages of reading development, in addition to providing simple and clear suggestions for improvement, the profiles are likely to facilitate the individualization of instruction though differentiation based on children's reading proficiencies and difficulties. However, the usability of the profiles, and their impact on reading development, require investigation before they are implemented in practice.

Based on the results of the current study, the following suggestions are made for future research. Firstly, researchers are advised to replicate the profiling analyses for relevant subpopulations with a sufficiently large sample, such that their potential usage can be evaluated. Secondly, future researchers are advised to investigate the impact of task novelty, motivation and concentration on children's performance, as the results suggest potential issues for a small subset of children. Thirdly, future researchers are advised to evaluate the usefulness of utilizing the optimal set of profiles to personalize instruction in practise, such that the impact on reading development can be evaluated. Finally, we stress that we aimed to remain pressingly conscious of the risks that can accompany personalisation, making sure teacher's workload is not increased or completely superseded, children's rights are retained and homogenization of education is thwarted. As such, we suggest that all future researches that tackle similarly relevant educational- or developmental contexts to keep these same risks in mind.

5.5 CONCLUSION

To conclude, the results of the current study provide evidence that distinct, practically and theoretically relevant reading fluency profiles that reflect all reading fluency components can be developed. The LPA analysis identified six profiles that differentiate second- and third-grade readers based on theoretically relevant reading proficiencies and difficulties, reflecting developmental reading stages. However, while the result support the reliability and generalizability of the optimal solution, the solutions for subpopulations contained to few observations to warrant generalization or usage. Therefore, future researchers are advised to replicate these analyses with a sufficiently large sample. Additionally, while the validation indicates that the profiles represent reading fluency skills and development, the interpretation of the context reader profile requires additional attention. Specifically, research should clarify whether children's reading performances might be the result of task novelty, motivation or boredom, instead of proficiency. Finally, future researchers are advised to evaluate whether using the profiles to guide reading instruction improves reading development. If their usage proves beneficial, this would allow teachers to individualize their reading education based on children's reading competencies and difficulties while reducing their workload, providing them with the means to personalize and improve the quality of their instruction.

REFERENCES

Aldhanhani, Z. R., & Abu-Ayyash, E. A. (2020). Theories and Research on Oral Reading Fluency: What Is Needed?. *Theory and Practice in Language Studies, 10*(4), 379–388. http://dx.doi.org/10.17507/tpls.1004.05

Amendum, S.J., Conradi, S.K., & Liebfreund, M.D. (2021). Explaining reading variance by student subgroup: Should we move beyond oral reading fluency? *Journal of Research in Reading, 44*(4), 757–786. https://doi.org/10.1111/1467-9817.12371.

Argyle, S. (1989). Miscue analysis for classroom use. *Reading Horizons, 29*(2), 93–102. https://scholarworks.wmich.edu/reading_horizons/vol29/iss2/2/.

Barthakur, A., Dawson, S., & Kovanovic, V. (2023, March). *Advancing leaner profiles with learning analytics: A scoping review of current trends and challenges.* LAK23: 13th international learning analytics and knowledge conference, New York, USA. https://doi.org/10.1145/3576050.3576083.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293–323. https://doi.org/10.1016/0010-0285(74)90015-2.

Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence, 3*, 100068. https://doi.org/10.1016/j.caeai.2022.100068.

van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R [Computer Software]. *Journal of Statistical Software, 45*(3), 167. https://doi.org/10.18637/jss.v045.i03.

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2013). Early Identification of Reading Disabilities Within an RTI Framework. *Journal of Learning Disabilities, 48*(3), 281–297. https://doi.org/10.1177/0022219413498115.

Caravolas, M. (2022). Reading and Reading Disorders in Alphabetic Orthographies. In M. J. Snowling, C. Hulme, & K. Nation (Eds.), *The Science of Reading: A Handbook* (2e ed., pp. 327–353). Wiley-Blackwell. https://doi.org/10.1002/9781119705116.ch15

Celeux, G., Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification 13*, 195–212. https://doi.org/10.1007/BF01246098.

Chrysafiadi, K., & Virvou, M. (2015). *Advances in personalized web-based education.* Springer International Publishing.

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. John Wiley & Sons.

Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: Implications for policy and practice. *Policy insights from the behavioral and brain sciences, 3*(1), 54–61. https://doi.org/10.1177/2372732215624931.

Cowie, R., Douglas-Cowie, E., & Wichmann, A. (2002). Prosodic characteristics of skilled reading: Fluency and expressiveness in 8–10-year-old readers. *Language and Speech, 45*(1), 47–82. https://doi.org/ 10.1177/00238309020450010301.

Dams, J. E., Schaars, M. M., Segers, E., & Blom, E. (2023). Understanding variation in prospective poor decoders: A person-centred approach from kindergarten to Grade c2. *Dyslexia, 29*(4), 312–329. https://doi.org/10.1002/dys.1750.

Ehri, L. C. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading, 18*(2), 116–125. https://doi.org/10.1111/j.1467-9817.1995.tb00077.x.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing, 7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417.

Falcão, P. T., e Peres, F. M. A., Morais, D. C. S., & da Silva Oliveira, G. (2018). Participatory methodologies to promote student engagement in the development of educational digital games. *Computers & Education, 116*, 161–175. https://doi.org/10.1016/j.compedu.2017.09.006.

Foorman, B. R., Petscher, Y., Stanley, C., & Truckenmiller, A. (2017). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of Research on Educational Effectiveness, 10*(3), 619–645. https://doi.org/10.1080/19345747.2016.1237597

Fox, J. P., Klotzke, K., & Simsek, A. S. (2021). LNIRT: An R package for joint modeling of response accuracy and times. *arXiv preprint arXiv:2106.10144*. https://arxiv.org/abs/2106.10144.

Fuchs, L., Fuchs, D., Hosp, M., And Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256. https://doi.org/10.1207/S1532799XSSR0503_3.

Gómez, S., Zervas, P., Sampson, D. G., & Fabregat, R. (2014). Context-aware adaptive and personalized mobile learning delivery supported by UoLmP. *Journal of King Saud University - Computer and Information Sciences, 26*(1), 47–61. https://doi.org/10.1016/j.jksuci.2013.10.008.

Grainger, J., & Segui, J. (1990). Neighborhood frequency effects in visual word recognition: A comparison of lexical decision and masked identification latencies. *Perception & psychophysics, 47*, 191–198. https://doi.org/10.3758/BF03205983.

Grimm, R. P., Solari, E. J., McIntyre, N. S., & Denton, C. A. (2018). Early reading skill profiles in typically developing and at-risk first grade readers to inform targeted early reading instruction. *Journal of School Psychology, 69*, 111–126. https://doi.org/10.1016/j.jsp.2018.05.009.

Groen, M. A., Veenendaal, N. J., & Verhoeven, L. (2018). The role of prosody in reading comprehension: evidence from poor comprehenders. *Journal of Research in Reading, 42*(1), 37–57. https://doi.org/10.1111/1467-9817.12133.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Springer. https://link.springer.com/book/10.1007/978-94-017-1988-9.

Harmsen, W.N., van Hout, R., Cucchiarini, C. and Strik, H. (2025). Can ASR generate valid measures of child reading fluency? *INTERSPEECH 2025, 2395-2399.* https://doi.org/10.21437/Interspeech.2025-306.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000.

Keuning, J., Swart, N., Scheltinga, F., Gruhn, C.S., Segers, E. & Verhoeven, L. (2019). Evaluatie en planning van leesleertrajecten: Een dynamisch perspectief (eindrapport NRO project 405-15-548). Cito: Arnhem. nro.nl/sites/nro/files/migrate/eindrapport-405-15-548.pdf.

Kim, E. J. (2024). Analysis of EFL Elementary School Students' English Reading Ability Profiles and Their Learning Backgrounds. *The Journal of AsiaTEFL, 21*(3), 661–678. http://dx.doi.org/10.18823/asiatefl.2024.21.3.9.661.

Kim, Y. S. G., Quinn, J. M., & Petscher, Y. (2021). What is text reading fluency and is it a predictor or an outcome of reading comprehension? A longitudinal investigation. *Developmental Psychology, 57*(5), 718–732. https://doi.org/10.1037%2Fdev0001167.

Kuhn, M. R., Schwanenflugel, P. & Meisinger, E. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 232–253. https://doi.org/10.1598/RRQ.45.2.4.

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95*(1), 3–21. https://doi.org/10.1037/0022-0663.95.1.3.

Li, K. C., & Wong, B. T. M. (2020). Features and trends of personalised learning: a review of journal publications from 2001 to 2018. *Interactive Learning Environments, 29*(2), 182–195. https://doi.org/10.1080/10494820.2020.1811735.

van der Linden, W. J. (2007). A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika, 72*, 287–308. https://doi.org/10.1007/s11336-006-1478-z.

Louradour, J. (2023). *Whisper-timestamped [Computer Software].* GitHub Repository: https://github.com/linto-ai/whisper-timestamped.

Maghsudi, S., Lan, A., Xu, J., & van Der Schaar, M. (2021). Personalized education in the artificial intelligence era: what to expect next. *IEEE Signal Processing Magazine, 38*(3), 37–50. https://doi.org/10.1109/MSP.2021.3055032.

Meelissen, M. R. M., Maassen, N. A. M., Gubbels, J., van Langen, A. M. L., Valk, J., Dood, C., Derks, I., In 't Zandt, M., & Wolbers, M. (2023). *Resultaten PISA-2022 in vogelvlucht.* Enschede: Universiteit Twente. https://doi.org/10.3990/1.9789036559461.

Miciak, J., Ahmed, Y., Capin, P., & Francis, D. J. (2022). The reading profiles of late elementary English learners with and without risk for dyslexia. *Annals of Dyslexia, 72*(2), 276–300. https://doi.org/10.1007/s11881-022-00254-4.

Miller, J., & Schwanenflugel, P.J. (2006). Prosody of syntactically complex sentences in the oral reading of young children. *Journal of Educational Psychology, 98*(4), 839–853. https://doi.org/10.1037/0022-0663.98.4.839.

Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading research quarterly, 43*(4), 336–354. https://doi.org/10.1598/RRQ.43.4.2.

Morin, A. J. S., Meyer, J. P., Creusier, J., & Biétry, F. (2015). Multiple-Group Analysis of Similarity in Latent Profile Solutions. *Organizational Research Methods, 19*(2), 231–254. https://doi.org/10.1177/1094428115621148.

Morris, D., & Perney, J. (2018). Using a sight word measure to predict reading fluency problems in grades 1 to 3. *Reading & Writing Quarterly, 34*(4), 338–348. https://doi.org/10.1080/10573569.2018.1446857.

Morrison, T. G., & Wilcox, B. (2020). Assessing expressive oral reading fluency. *Education sciences, 10*(3), 59. https://doi.org/10.3390/educsci10030059.

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 International Results in Reading.* Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342.

Nandigam, D., Tirumala, S. S., & Baghaei, N. (2014, December). *Personalized learning: Current status and potential.* In 2014 IEEE Conference on e-Learning, e-Management and e-Services, Hawthorne, VIC, Australia. https://doi.org/10.1109/IC3e.2014.7081251.

Nation, K. (2005). Children's reading comprehension difficulties. In Snowling, M. J., &Hulme, C. (Eds.), *The science of reading: A handbook* (pp. 248–265). Blackwell. https://doi.org/10.1002/9780470757642.ch14.

National Institute of Child Health and Human Development. (2000). Report of the National Reading Panel. Teaching children to read: *An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. U.S. Government Printing Office: Washington, DC. https://www.nichd.nih.gov/publications/pubs/nrp/smallbook.

Pikulski, J.J., & Chard, D.J. (2005). Fluency: Bridge between decoding comprehension. *The Reading Teacher, 58*(6), 510–519. https://doi.org/10.1598/RT.58.6.2.

Posit team (2025). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

Psyridou, M., Tolvanen, A., de Jong, P. F., Lerkkanen, M. K., Poikkeus, A. M., & Torppa, M. (2021). Developmental profiles of reading fluency and reading comprehension from grades 1 to 9 and their early identification. *Developmental Psychology, 57*(11), 1840–1854. https://doi.org/10.1037/dev0000976.

Rasinski, T. V. (2012). Why reading fluency should be hot. *The Reading Teacher, 65*(8), 516–522. https://doi.org/10.1002/TRTR.01077.

Risberg, A. K., Widlund, A., Hellstrand, H., Vataja, P., & Salmi, P. (2024). Profiles of reading fluency and spelling skills: Stability and change across the early school years. *Scandinavian Journal of Educational Research, 68*(6), 1231–1246. https://doi.org/10.1080/00313831.2023.2228822.

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., Van Lissa, C. J., Schmidt, J. A. (2018). tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software [Computer Software]. *Journal of Open Source Software, 3*(30), 978. https://doi.org/10.21105/joss.00978.

Scott, E., Soria, A., & Campo, M. (2017). Adaptive 3D Virtual Learning Environments—A Review of the Literature. *IEEE Transactions on Learning Technologies, 10*(3), 262–276. https://doi.org/10.1109/TLT.2016.2609910.

Scrucca L, Fraley C, Murphy TB, Raftery AE (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R [Computer Software].* Chapman and Hall/CRC. https://doi.org/10.1201/9781003277965.

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.

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of vocational behavior, 120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445.

Tapalova, O., & Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIEd for personalised learning pathways. *Electronic Journal of e-Learning, 20*(5), 639–653. https://eric.ed.gov/?id=EJ1373006.

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., Vloedgraven, J. & De Wijs, A. (2018a). *Wetenschappelijke verantwoording DMT.* Cito: Arnhem.

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., & De Wijs, A. (2018b). *Wetenschappelijke verantwoording AVI.* Cito: Arnhem.

University of Oregon (2020). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Administration and Scoring Guide*. Eugene, OR: University of Oregon. https://dibels.uoregon.edu.

Veenendaal, N.J., Groen, M.A. & Verhoeven, L. (2015). What speech text reading fluency can reveal about reading comprehension. *Journal of Research in Reading, 38*(3), 213–225. https://doi.org/10.1111/1467-9817.12024.

Veenendaal, N.J., Groen, M.A., & Verhoeven, L. (2016). Bidirectional Relations between Text Reading Prosody and Reading Comprehension in the Upper Primary School Grades: A Longitudinal Perspective. *Scientific Studies of Reading. 20*(3), 189–202. https://doi.org/10.1080/10888438.2015.1128939.

van der Velde, M., Harmsen, W., Veldkamp, B. P., Feskens, R.C.W., Keuning, J., & Swart, N. (2025). Speech enabled reading fluency assessment: A validation study. *International Journal of Artificial Intelligence in Education,* 1-27. https://doi.org/10.1007/s40593-025-00480-y.

van der Velde, M., Molenaar, B., Veldkamp, B. P., Feskens, R.C.W., & Keuning, J. (2024a). What do they say? Assessment of oral reading fluency in early primary school children: A scoping review. *International Journal of Educational Research, 128*, 102444. https://doi.org/10.1016/j.ijer.2024.102444

Van der Velde, M., Veldkamp, B. P., Keuning, J., Feskens, R. C. W., Swart, N. M., Harmsen, W. N. (2024b). The framework and development of SERDA: Speech enabled reading fluency assessment for Dutch. In Ranđelović B., Karalić E., Aleksić K., Đukić D. (Eds.), *E-testing and computer-based assessment. CIDREE Yearbook 2024* (pp. 99–123). CIDREE. https://cidree.org/wp-content/uploads/2024/11/cidree_yearbook-2024.pdf

Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology, 22*(3), 407–423. https://doi.org/10.1002/acp.1414.

Virinkoski, R., Lerkkanen, M.-K., Holopainen, L., Eklund, K., & Aro, M. (2018). Teachers' ability to identify children at early risk for reading difficulties in grade 1. *Early Childhood Education Journal, 46*(5), 497–509. https://doi.org/10.1007/s10643-017-0883-5.

Wagner, R. K., Zirps, F. A., & Wood, S. G. (2022). Developmental dyslexia. In M. J. Snowling, M.J., Hulme, C., & Nation, L. (Eds.), *The science of reading: A handbook* (2nd ed., pp. 416–438). Wiley-Blackwell. https://doi.org/10.1002/9781119705116.ch19.

Wiederholt, J. L., & Bryant, B. R. (2012). *Gray Oral Reading Test–Fifth Edition* (GORT-5): Examiner's manual. Austin, TX: Pro-Ed.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading, 5*(3), 211–239. https://doi.org/10.1207/S1532799XSSR0503_2.

Xie, H., Chu, H. C., Hwang, G. J., & Wang, C. C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education, 140*, 103599. https://doi.org/10.1016/j.compedu.2019.103599.

Xu, X., Li, Z., Hin Hong, W. C., Xu, X., & Zhang, Y. (2024). Effects and side effects of personal learning environments and personalized learning in formal education. *Education and Information Technologies, 29*(15), 20729–20756. https://doi.org/10.1007/s10639-024-12685-0

Zheng, L., Zhong, L., Niu, J., Long, M., & Zhao, J. (2021). Effects of Personalized Intervention on Collaborative Knowledge Building, Group Performance, Socially Shared Metacognitive Regulation, and Cognitive Load in Computer-Supported Collaborative Learning. *Educational Technology & Society, 24*(3), 174–193. https://www.jstor.org/stable/27032864.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*(1), 3–29. https://doi.org/10.1037/0033-2909.131.1.3.

Chapter 6

Personalizing primary reading education: Can Data-to-Text generation guide reading instruction?

van der Velde, M., Swart, N., Veldkamp, B. P., Feskens, R.C.W. (2026). *Personalizing Primary Reading Education: can Data-to-Text Generation Guide Reading Instruction?*[Manuscript submitted for publication]. University of Twente.

ABSTRACT

The benefits of providing students with fitting feedback throughout formal education has been well-documented, proving promising for optimizing developmental trajectories. However, constructing personal feedback is time-consuming, especially when done regularly, increasing the work-load of teachers when systematically implemented. As a result, this role is increasingly being fulfilled by automatic feedback systems. Especially LLM's have proven useful recently, showing potential in reducing teacher's teaching burdens and in improving student's learning performances and experiences. Given the current popularity of LLM's, assessing the quality of their feedback is of the upmost importance, especially within life determining context such as education. Thus, the current study will focus on evaluating to what degree hallucination-free, practically and didactically relevant feedback can be generated using a data-to-text LLM approach to feedback generation, and to what degree contextualization impacts feedback quality. The results indicate that the quality of LLM feedback is generally high, especially with regard to readability. However, LLM feedback was less coherent and didactically sound than human feedback, providing overtly general or vague feedback. The current study underlines the importance of furthering knowledge on optimal modes of data delivery to LLM's within complex data situations and contexts, and on furthering knowledge on optimalization processes within LLM data-to-text generation approaches.

6.1 INTRODUCTION

The benefits of providing students with fitting feedback throughout formal education has been well-documented, proving promising for optimizing developmental trajectories (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Newman et al., 2021). Especially the provision of personalized and instructional feedback, aimed at improving task performance based on previous and current successes, has established advantages (Hattie, 1999; Kluger & DeNisi, 1996). These advantages were quantified in work by van der Kleij et al. (2015), who investigated effect of feedback type on student's learning outcomes through a meta-analysis. Their results indicate that elaborative forms of feedback were more effective than feedback where information was provided on the correctness of the answer, and feedback where the right answer is provided, especially with regard to higher order learning outcomes.

Furthermore, constructing elaborate feedback based on performance suits data-driven instruction. Data-driven instruction focusses on utilizing student-specific statistics to guide, differentiate and personalize teaching approaches (Campbell & Levin, 2009; Schildkamp et al., 2017), which has shown promise in supporting the improvement of teacher's instruction and student learning (e.g. Schildkamp et al., 2013; van Geel et al., 2016). In addition, increasing the quality and feasibility of personalized feedback furthers the personalization of education. Personalized education, in turn, has increasingly been identified as crucial to facilitate student development (Connor & Morrison, 2016; Li & Wong, 2020), showing beneficial effects on student performance, motivation, understanding, satisfaction and learning efficiency (Falcão et al., 2018; Gómez et al., 2014; Zheng et al., 2021).

However, although the benefits of personalization have been thoroughly discussed and demonstrated, the potential risks of personalising education have also received a lot of attention as of late. Firstly, constructing personal feedback can be time-consuming, especially when done regularly, leading to an increase in teacher's workloads. In addition, concerns exist regarding ethical issues related to data privacy, stemming from the increased availability and utilization of children's personal data by, at or for schools (Merino-Compos, 2025). Others voice concerns regarding the quality of education, as learning might become more homogenized when (generative) artificial intelligence (AI) is used to facilitate personalisation, reducing the potential of students to develop critical thinking and collaborative skills (Sack & Little, 2025).

In short, while personalized feedback fits currently popular educational perspectives, we should remain pressingly conscious of the risks that can accompany the personalisation of education. Especially in a time where generative AI runs rampant, it is of the utmost importance to investigate if, how and where to further facilitate its implementation. Therefore, we shall focus on evaluating the potential of utilizing generative AI, which we shall refer to as Large Language Models (LLM's), to automatically generate personalized

feedback throughout the remainder of this article. However, before we detail how these evaluations are to be conducted, we will first discuss the concept of feedback and previous research on automatic feedback generation.

6.1.1 Feedback in education

Educational theory describes feedback as the dual process of observing or noticing student's errors and remediating these errors through the provision of information (Butler & Winne, 1995). Following Brown and VanLehn's (1980) repair theory, these errors are not viewed as random, but instead follow a student-specific mental representation of the task, containing personal misrepresentations and misconceptions. Correspondingly, error remediation should transcend identifying errors. Instead, effective feedback ought to be aimed at guiding students through the "repair" of their misconceptions, providing meaningful revisions to their mental models (Carless, 2006; VanLehn, 2006). Based on this theoretical framework, we refer to feedback as the provision of performance- or understanding-related information, by any agent (Hattie & Timperley, 2007), with the aim of reducing discrepancies between what a student understands or masters and that which is required to be understood or mastered (Sadler, 1989).

While the agents providing feedback in educational settings have traditionally been teachers, limited resources, a focus on more continuous forms of assessment, and an increase in available student data have complicated the provision of fitting feedback (Boud & Molley, 2013). In response, this role is increasingly being fulfilled by automatic feedback systems, such as intelligent tutoring systems, online learning platforms and LLM's (Dai et al., 2023; Kochmar et al., 2020; Razzaq et al., 2020). This increased interest is not surprising, as automatic feedback has often been shown to improve student performance (Cavalcanti et al., 2021). Especially LLM's have proven useful recently, showing potential in reducing teacher's teaching burdens (Agostini & Picasso, 2024; Seo et al., 2025), and in improving student's learning performances and experiences (Meyer et al., 2024; Wiboolyasarin et al., 2024).

To understand the application of LLM's within the context of feedback generation, as well as its benefits and shortcomings compared to alternative approaches, we provide the following overview of feedback generation approaches, based on which our approach is founded.

6.1.2 Feedback generation

A field that has made considerable progress regarding the automatic provision of feedback concerns the domain of Natural Language Generation (NLG). Here, NLG refers to AI applications that convert different types of data into comprehensive, understandable and human-like texts (Osuji et al., 2024; Reiter, 1997). NLG applications traditionally focus on tasks related to summarizing, simplifying or translating texts, typically following a text-

to-text design. However, data-to-text generation (e.g. McKeown, 1992), which focusses on generating easy to understand text based on non-textual data (Reiter & Dale, 2000), such as tables or databases, has gained increased interest over the last decades (Reiter, 2007; Nan et al., 2022; Parikh et al., 2020).

Traditional models for data-to-text generation (McKeown, 1992) have focused on using specific rules or templates. However, as such approaches tended to lack scalability, and provide limited generalization towards alternative domains (Lin et al., 2023), traditional methods were gradually replaced by probabilistic statistical techniques (e.g. Angeli et al., 2010). More recently, these statistical techniques have made way for end-to-end neural models, providing more naturally reading, fluent, coherent and scalable textual output (Lin et al., 2023). More recently still, the rapid onset of LLM's has redefined what seemed possible within the field of NLG (Li et al., 2024; Ouyang et al., 2022; Zhao et al., 2023). Especially educational settings, where the complexity and magnitude of available information can complicate effective utilization (Schildkamp et al., 2017), might benefit from LLM-based data-to-text generation.

While the generation of personal feedback could prove beneficial within various educational settings, early reading education is an educational area pressingly in need of aid. This perspective is not necessarily new, as Connor & Morrison (2016) argued that sufficiently differentiated literacy instruction might help children who would otherwise fail to reach their reading potential, one decade ago. However, recent trends, showing reduced reading performances by primary and secondary school children in countries the world over (Meelissen et al., 2023; Mullis et al., 2023), have given such perspectives renewed interest and increased weight. Here, improving reading instruction at an early stage, through personalized feedback, is expected to prove especially effective, as the reading skills of children in the early years (e.g. Grade 1-3) of primary education tend to be predictive of reading skills in the latter (e.g. Grade 4-6) years (Verhoeven & van Leeuwe, 2008).

Nevertheless, automating feedback generation through LLM's within educational settings warrants caution. Some applications have led to an overreliance on AI assistance, negatively impacting student's critical thinking ability and creativity (Darvishi et al., 2024). In addition, while generated texts are often fluent, coherent and believable, hallucinations, which concern the generation of nonsensical or unfaithful text, are still a prominent problem (Ji et al., 2023). Moreover, multiple pedagogical challenges remain, as LLM's show a general lack of specificity and often lack the ability to adequately incorporate relevant educational contexts (Ji et al., 2023; Seo et al., 2025). Given these shortcomings, it is of the utmost importance to investigate how LLM's can be optimized to perform within the relevant educational context of early reading.

6.1.3 The present study

Given the current popularity of LLM's within the area of NLG, is of the upmost importance to investigate if and how these models can most optimally perform, especially within life determining context such as education. Therefore, the current study will focus on evaluating to what degree hallucination-free, practically and didactically relevant feedback can be generated using a data-to-text LLM approach to feedback generation. In addition, we investigate whether the provision of relevant contextual information impacts feedback quality. The current study will focus on answering the following questions:

1. To what degree can data-to-text LLM's generate hallucination-free, practically and didactically relevant feedback within the context of early reading education?
2. How does the inclusion of contextual information impact the quality of feedback generation?

6.2 METHODS

Throughout this study we evaluated to what degree an LLM data-to-text generation approach can generate hallucination-free, practically and didactically relevant feedback based on primary children's reading fluency skills. For these purposes, we utilized the Speech Enabled Reading Diagnostics App (SERDA; van der Velde et al., 2024) dataset. First, a feedback template was constructed in collaboration with an expert on reading and didactics to support humans in providing structured feedback. Then, the template was applied to a subset of the data to create examples of didactically sound feedback. Finally, an LLM was prompted to generate feedback based on (1) the SERDA data, (2) the SERDA data and the template, and (3) the data, template and feedback examples.

6.2.1 Data

The current study utilized a subset of the SERDA dataset, for which complete data was observed (van der Velde et al., 2026). This subset contained 534 second- and third Grade children from 19 different primary schools in the Netherlands. Within the data, children were, on average, 7.6 (SD = 0.73) years of age. Second graders were generally a year younger (M = 7.0) than third graders (M = 8.14). The dataset contains slightly more girls (52%) than boys.

6.2.2 Materials

6.2.2.1 *SERDA: Reading fluency metrics.*

SERDA is a reading fluency assessment instrument that evaluates the oral reading fluency skills of Dutch children (van der Velde et al., 2024). SERDA contains three passage reading

tasks and three word decoding tasks, constructed to match the reading difficulties and complexities relevant for the second and third grade of Dutch primary education (Dutch: Groep 4 and Groep 5). Throughout the passage reading task, children read three passages of about 175 words aloud. Children were instructed to read as quickly and accurately as possible. The word decoding task consisted of the reading of three 50-word sets, which were presented through a progressive demasking design (Grainger & Segui, 1990) to improve speed estimation and ASR consistency (van der Velde et al., 2024). Both the passage reading and word decoding tasks were administered, one-on-one, on a tablet in a separate room.

Based on children's reading performance, accuracy, speed and automaticity scores were extracted at the task level for the passage reading and word decoding task, which were aggregated based on all completed subtasks. Here, accuracy reflected the average number of words read correctly, while speed concerned the average number of words a child was able to read per minute. Automaticity was followingly calculated as the average number of words read correctly per minute. In order to obtain an estimate of variability for the performance metrics, standard deviations were calculated for each of the performance metrics over their subtasks scores.

Furthermore, prosodic features were extracted based on children's performance on the passage reading task. These concerned the average duration of pauses, the average volume and average pitch of children, as well as their average variability in pauses, volume and pitch. While a more thorough discussion of the extraction and validation of SERDA's fluency features is beyond the scope of this article, these can be consulted in their corresponding articles (van der Velde et al., 2024, 2025; Harmsen et al., 2025).

6.2.2.2 *SERDA: Reading fluency profiles.*

In addition to fluency metrics, the SERDA dataset also contains a reading profile for each child, reflecting the types of reading qualities and difficulties children were observed to experience (van der Velde et al., 2026). There were six different profiles, which, in developmental order, were referred to as "Learning Readers", who experience difficulties with accurate and quick reading, "Accurate Readers", who read relatively accurately but have difficulty with speed, "Quick Readers", who read relatively quickly but have difficulties with accuracy, "Context Readers", who read the passages relatively well but have difficulty with the word decoding task, "Technical Readers", who read both accurately and quickly but relatively inexpressively and "Fluent Readers", who read relatively accurately, quickly and expressively.

6.2.3 Data analysis

6.2.3.1 *Feedback generation.*

During this study we utilized Open-WebUI (Open WebUI, 2025) to generate the feedback in the LLM conditions. However, given the personal nature of the data, we worked within a protected environment, such that the data was not shared with other parties. To evaluate the impact of contextualization, we have applied three data-to-text generation approaches with various levels of didactical contextualization, to a random sample of 40 children. An overview of the feedback generation and evaluation process is provided in Figure 1.

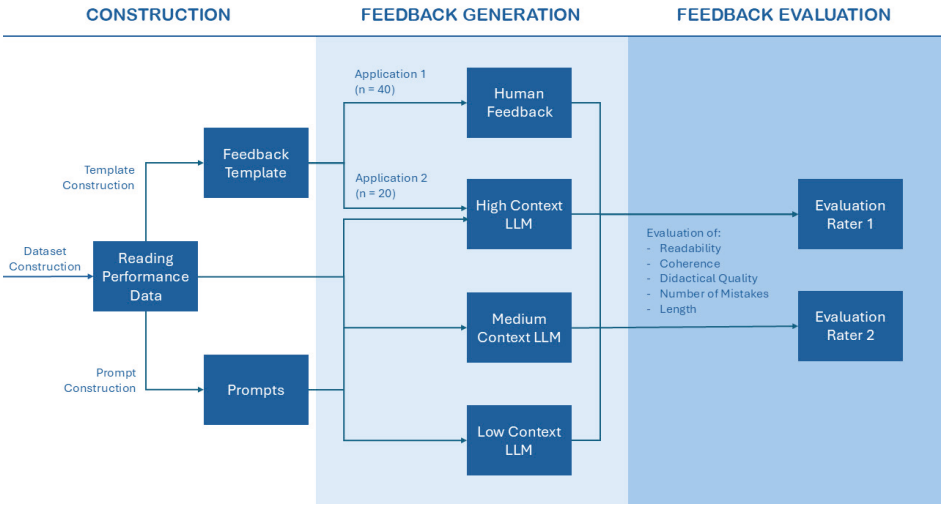


Figure 1. *Conceptual Representation of the Feedback Generation and Evaluation Pipeline*

The first approach, termed low context, fed SERDA data and a feedback generation prompt to the LLM, resembling a Retrieval-Augmented Generation (RAG) setup. RAG enhances feedback generation by providing LLM's with relevant external information, facilitating the evaluation of truthfulness through a comparison with source data (Ji et al., 2023).

The second approach, termed medium context, furthers model contextualization by incorporating a didactically grounded feedback template, as discussed below, which was constructed with an expert on reading and didactics. Using a feedback template to enhance truthfulness took inspiration from traditional data-to-text methodology (McKeown, 1992), as well as more recent two-stage and neural-template approaches to data-to-text, which have shown promise regarding controllability and truthfulness (see Lin et al., 2023). In essence, we aimed to investigate whether we could combine the best aspects of traditional and

modern approaches to construct quickly and easily available, yet didactically informed, feedback.

The third approach, termed high context, took the didactical contextualization a final step further by including human applications of the template to a separate random sample of 20 children. In essence, each child in this new sample was provided with human feedback, based on the template and the data of these children. To evaluate to what degree the LLM copies examples, one child from the original sample of 40 children was included within the new sample, for which the generated feedback was compared to the human feedback.

The exact prompts fed to the LLM's are provided in Appendix A.

6.2.3.2 *Template construction and application.*

The feedback template was constructed in two steps. First, we evaluated how the data of the children could most fittingly be summarized. Then, we determined how potential summaries could logically result in suggestions for improving reading development. The resulting template, and the instructions related to its implementation, can be consulted in Appendix B.

The most relevant information to be summarized was determined to be information that reflected relevant educational aspects of reading development. For these purposes, following the Dutch learning goals for early primary education was assessed to be the most optimal solution. Fittingly, the measures and profiles present within the SERDA dataset have been argued to reflect relevant aspects of early reading and developmental stages of reading (van der Velde et al., 2025, 2026). Therefore, the summary primarily focused on the reading profiles and children's deviations from their profile.

As such, the summarization of the template was specified as follows. First, the child was described as belonging to a specific profile (e.g. Charlie is an Accurate Reader). Second, a short description of the profile is provided (e.g. Accurate Readers tend to make relatively few reading mistakes, but often have difficulty with reading quickly). Third, children's deviations from their profile are detailed (e.g. While Charlie read the stories as well as other accurate readers, they had more difficulty with the word reading task. In addition, Charlie made more mistakes on the word reading task than other accurate readers, but did read the words more quickly). Here, deviations from the profile concerns differences compared to the mean of the profile regarding accuracy, speed or automaticity, with a magnitude of at least one standard deviation.

Correspondingly, the template describes how and based on what to construct suggestions. Specifically, suggestions are provided in two stages. First, advice is provided for the student based on profile membership. Then, more personal suggestions are made based on children's deviations from their profile. The finalized template and its implementation instructions are provided in Supplementary Appendix A.

The application of the template was relatively straightforward. For the random sample of 20 children, we followed the specified template to construct a summary based on children's scores, and specify didactical suggestions based on the summary. After finetuning the template, it was applied to the random sample of 40 children.

6.2.3.3 *Feedback evaluation.*

To investigate the quality of the generated feedback and the impact of contextualization we used human evaluations. Specifically, two experts on reading fluency and -didactics from the Expertise Center for Dutch (Expertisecentrum Nederlands) were consulted to evaluate all feedback on four dimensions. To elaborate, both evaluators evaluated the readability, coherence, didactical quality and presence of mistakes for the feedback generated by the LLM's, as well as the human feedback. In addition to these four dimensions, we also determined the length of the feedback, resulting in five dimensions.

Here, readability concerned the general conciseness and quality of the text of the feedback, reflected whether the feedback felt natural and was free from grammatical and syntactical errors. Coherence referred to the completeness, consistency and structure of the statements in the feedback, and whether the feedback provides a coherent story. The didactical quality of the feedback was evaluated by investigating the correctness of the didactical suggestions, as well as the corresponding methods, to current standards. Finally, the presence of mistakes was determined through comparing the specific statements that were made about the children's performance to their data.

For each of the forty sets of feedback, these aspects were provided with a score from 0-5, where 0 indicates very poor and 5 indicates very good. Meanwhile, length concerned the total number of words in the feedback. To control for ordering effects, the order in which the feedback was presented to the experts was randomized. The final dimension scores for each piece of feedback was determined by averaging over the rater's scores. To evaluate the usability of these scores, we evaluated the inter-rater reliability, using the weighted Kappa with quadratic weights to account for both differences in scores and their magnitude.

After evaluating the feedback, differences and similarities between feedback approaches were investigated. First, for each of the evaluation dimensions, we investigated whether their final scores showed significant differences between feedback conditions through ANOVA analyses, including Tukey post-hoc analyses if significant differences were found. Then, we investigated whether the feedback approaches varied in feedback quality over time, through linear regression analyses. These linear models were separately specified for each feedback condition, took the order of feedback generation as outcome, and all evaluation dimensions as predictors. Significance was evaluated with an alpha of 0.05.

Assumptions were checked, leading to the application and description of robust alternatives if assumptions were not met, and the alternative analyses lead to relevant changes in interpretation.

6.3 RESULTS

6.3.1 Descriptives

Throughout this study, two raters evaluated 160 pieces of feedback (forty for each of the four feedback conditions). Table 1 shows the final dimension scores for each feedback condition, obtained through averaging the scores of both raters. Weighted Kappa's estimates, calculated between the scores from both raters, ranged between 0.60 and 0.84, indicating moderate to strong agreement between the raters.

Table 1. Average Dimension Scores and Standard Deviations (SD) for Each Evaluated Dimension, for Each Feedback Generation Condition.

Condition	Readability	Coherence	Didactical Quality	Mistakes	Length
Human	2.90 (0.22)	4.31 (0.59)	4.81 (0.25)	0.03 (0.16)	186.38 (12.3)
Low Context	3.40 (0.59)	3.54 (0.54)	2.80 (0.71)	0.51 (0.90)	73.7 (14.5)
Medium Context	3.26 (0.68)	3.42 (0.75)	2.79 (0.76)	0.38 (0.95)	89.4 (27.1)
High Context	3.56 (0.61)	3.61 (0.82)	3.71 (0.83)	0.86 (1.19)	71.4 (21.8)

Note. Length concerns the average number of words in the feedback. Bracketed numbers concern SD's.

We evaluated whether the LLM copies the provided contextual examples by comparing the feedback provided to the child that was included within the original and new sample. Based on a direct comparison of the feedback for this child in the human and LLM conditions, and the corresponding scores, we conclude that the LLM did not copy the example.

6.3.2 One way ANOVA's

Table 2 shows the results of the one-way ANOVA analyses used to evaluate differences in dimension scores between conditions, including Post-Hoc Results.

Table 2. One-way anova results for each dimension, including post-hoc results.

Dimension	F-score	P-value	Eta^2	Post-Hoc Results
Readability	9.92	< 0.001	0.16	Human < Low, Medium and High
Coherence	13.84	< 0.001	0.21	Low, Medium and High < Human
Didactical Quality	80.21	< 0.001	0.61	Low, Medium and High < Human, Low, Medium < High
Number of Mistakes	6.14	< 0.001	0.11	Human < High
Length	303.9	< 0.001	0.85	Low, Medium and High < Human, Low, High < Medium

Note. Low, Medium and High refer to the degree of contextualization in the LLM conditions.

The results of the one-way ANOVA analyses indicate that there are significant differences in the readability, coherence, didactical quality, number of mistakes and length of the feedback, between conditions.

The post-hoc analyses indicated that feedback in all LLM conditions was significantly more readable than human feedback, but less coherent.

The feedback in all LLM conditions showed significantly lower didactical quality than human feedback. In addition, the feedback in the high context LLM condition showed higher didactical quality than the feedback in the low and medium context LLM conditions.

The number of mistakes was significantly higher in the high context LLM condition, in comparison to Human feedback.

The length of the feedback was significantly lower in all LLM conditions, compared to human feedback. In addition, the length of feedback was substantially lower in the low and high context LLM conditions, compared to the medium context LLM condition.

6.3.3 Feedback stability

Table 3 shows the results of the linear regression analyses, used to investigate the effect of generational order on the Readability, Coherence, Didactical Quality, the Number of Mistakes and Length of the feedback in each feedback condition.

Table 3. Effect of generational order of feedback on readability, coherence, didactical quality, the number of mistakes and length.

Feedback Condition	F-score	P-value	R^2	Effect of Generational Order
Human	0.90	0.49	0.12	No effect
Low Context	10.72	< 0.001	0.61	Lower Didactical Quality, Lower Length
Medium Context	15.97	< 0.001	0.70	Lower Didactical Quality, Lower Length
High Context	5.30	0.001	0.43	Lower Length

Note. Low, Medium and High refer to the degree of contextualization in the LLM conditions.

The results of the linear regression analyses indicate that the specified models significantly predict the generational order of feedback in all LLM conditions, but not in the human feedback condition.

For the low context condition, significant negative relationships were observed for didactics (t = -3.49, p = 0.001) and length (t = -5.73, p < 0.001), indicating that the didactical quality and length of the feedback diminished over time.

For the medium context condition, significant negative relationships were observed for didactics (t = -2.44, p = 0.02) and length (t = -5.34, p < 0.001), indicating that the didactical quality and length of the feedback diminished over time.

For the high context condition, a significant negative relationship was observed for length (t = -4.66, p < 0.001), indicating that the length of the feedback was diminished over time.

Figure 2 shows the distribution of the readability, coherence and didactical quality scores, as well as the number of mistakes, in generated order, for each of the feedback generation condition. Correspondingly, Figure 3 shows the length of the feedback, in generated order, for each feedback generation condition.

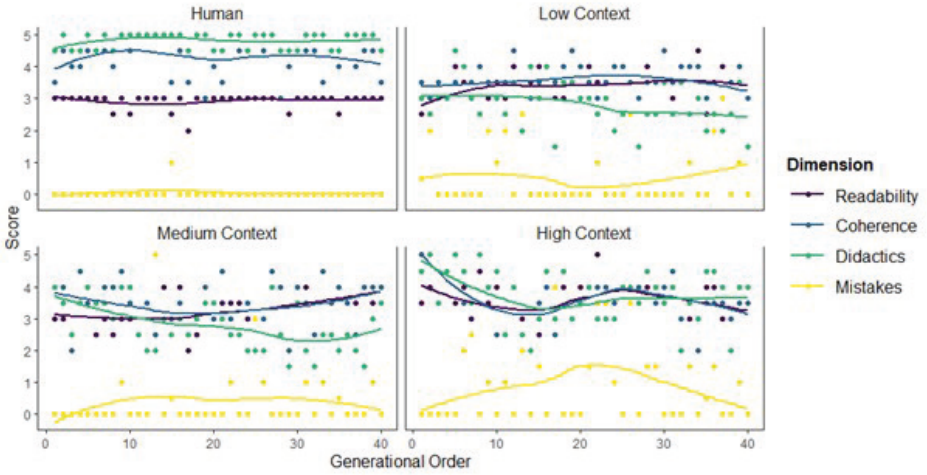


Figure 2. *Readability, coherence, didactical quality and number of mistakes of the feedback, in order of generation, for each generational approach.*

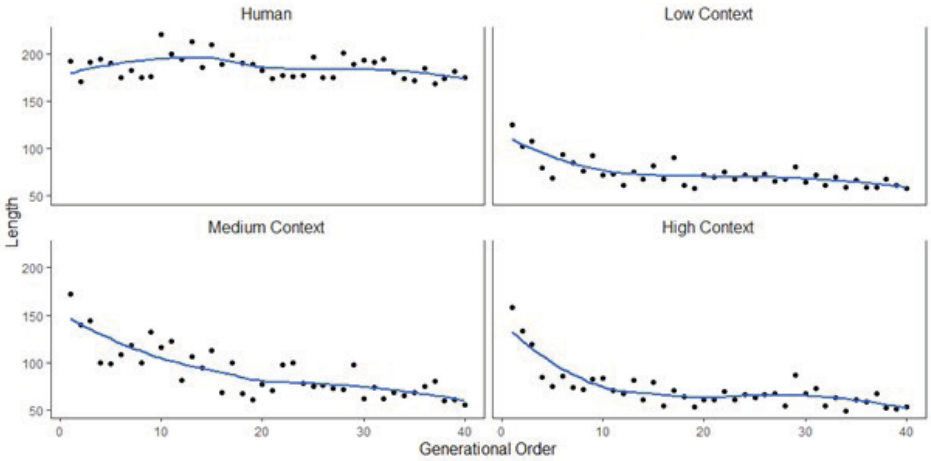


Figure 3. *Length of the Feedback in Order of Generation, for Each Feedback Condition.*

6.4 DISCUSSION

The aim of the current study was to investigate to what degree LLM's can generate hallucination-free, practically and didactically relevant feedback within the context of early reading education, and how the inclusion of a didactically grounded template, and applied examples of this template, impact feedback generation. For these purposes we generated feedback for 40 primary school children, in one human and three data-to-text generation conditions. The 160 resulting pieces of feedback were evaluated by two experts, based on their readability, coherence, didactical quality, number of mistakes and length.

The results show that both raters rated the feedback relatively favorably. To elaborate, over three-fourth of all evaluations concerned scores between three and five, indicating that most of the feedback was of decent to good quality. In fact, no feedback was provided with a score below two for readability and coherence, while only a few ($n < 10$) low scores were provided for the didactical quality of the feedback. At the same time, one-fourth to one-fifth of the feedback texts contained at least one mistake, indicating that the feedback is not free from hallucinations. These results match well known patterns in the literature, which suggest that LLM feedback generation tends to provide readable and coherent, yet hallucinatory feedback (Ji et al., 2023).

Further contextualization is provided by the ANOVA analyses, which indicate that LLM feedback is more readable than human feedback. An explanation for this difference concerns the accompanying difference in the length of the feedback, which was lower in all LLM conditions when compared to human feedback.

At the same time, none of the LLM conditions were able to match the human feedback with regard to coherence. This differences in human and LLM feedback is primarily related to the completeness of the feedback. While the human feedback structurally discussed all datapoints, specifically following up on any negative finding, the LLM's tended to provide short summaries with less coherent follow ups. In that sense, the reduced length of the feedback might have increased the readability of the feedback at the cost of its coherence.

Similarly, human feedback showed higher didactical quality than LLM feedback. These results are not all that surprising, given that the human feedback was provided by an expert on language and didactics. This expertise was not available to the LLM within the low and medium context conditions, forcing the LLM to produce this instead. As discussed in earlier research (Ji et al., 2023; Seo et al., 2025), this tends to result in overly general and vague suggestions, which were also observed in this study (e.g. "The teacher can get to work on this by explicitly incorporating word-level work" and "using multi-sensory reading strategies"). Fittingly, the high context condition showed higher didactical quality than both the low and medium context conditions. This improvement is likely due to the fact that the high context condition did have access to didactic expertise, through the applied examples

of the template. These results underline the importance of providing explicit examples of contextualized didactical information, if an LLM is expected to provide it.

This increase in didactical quality does not appear to be without cost, as the feedback in the high context condition also showed more hallucinations. These hallucinations primarily concerned mistakes in the naming of deviations from the profile, where the LLM mixed up the variable names when interpretating scores. While these results appear problematic, suggesting that additional context might increase instead of reduce hallucinations, some nuance is required.

First of all, the didactical suggestions that followed from the mistakes were sometimes, surprisingly, correct. To elaborate, instead of following the incorrectly provided feedback, the didactical suggestions were oftentimes based on children's scores, resulting in correct didactical suggestions. This indicates that the mistakes might primary concern an issue regarding accurate naming in a complex data situation.

In addition, the number of mistakes in the high context condition might be a result of the limited number of texts being generated. To elaborate, Figure 2 shows that the number of mistakes for the high context condition appears to be in sharp decline near the end of the generational process. These findings indicates that the LLM might still have been in the process of optimizing the feedback. This is especially likely when taking into account that the complexity of this condition, containing an additional dataset, surpasses the complexity of the low and medium context conditions.

This optimization argument brings us to a final point of interest: the stability of the feedback quality over time. For all but the human feedback, the length of the feedback was significantly reduced throughout the generational process. At the same time, the didactical quality of feedback decreased over time for the low and medium context conditions. This finding indicates that the quality of the feedback actually diminishes over time if no correct examples of didactical suggestions are provided. These results further substantiate the importance of providing correct examples of didactical suggestions, if LLM's are ever expected to learn how to provide them.

Based on these finding, a few suggestions are made regarding future research. First, given the importance of including didactically sound examples within an already complex data scenario, future researchers are advised to further clarify optimal LLM specification procedures for complex data conditions. Specifically, we argue that experiments should be conducted that inform how practically relevant characteristics of feedback, such as readability, coherence and didactical quality, behave for various LLM and RAG approaches, prompting designs and contexts. Especially the investigation of differences based on modes of information presentation, such as through data, prompting, a combination of both or alternatives, as well as their qualitative nuances, might create a more thorough understanding of how to implement LLM's within relevant educational contexts.

In addition, future research should more thoroughly investigate how feedback generation processes optimize themselves over their runtime, especially when applied to complex, impactful contexts where mistakes are life-altering. For such purposes, we suggest the investigation of burn-in periods before evaluations are conducted, as can be borrowed from Bayesian approaches.

6.5 CONCLUSION

Throughout this study, we investigated to what degree LLM's can generate hallucination-free, practically and didactically relevant feedback within the context of early reading education, and how providing context impacts feedback quality. The results indicate that the quality of the LLM feedback is relatively high, showing improvements regarding readability and text length when compared to human feedback. However, LLM feedback was found to be less coherent and didactically sound than human feedback, showing less complete and overtly general or vague feedback. In addition, LLM feedback contained more hallucinations, mostly in the form of misnaming data. In short, it appears that LLM's can generated qualitative feedback within education settings, but improvements need to be made regarding coherence, didactical quality and hallucination reduction before implementation is warranted. Further improving knowledge on optimal modes of data delivery within complex data situations and contexts, and a thorough investigating of optimalization processes within LLM data-to-text generation approaches, might guide the way towards more practically applicable feedback within data and consequence rich contexts.

REFERENCES

Agostini, D., & Picasso, F. (2024). Large language models for sustainable assessment and feedback in higher education: Towards a pedagogical and technological framework. *Intelligenza Artificiale, 18*(1), 121–138. https://doi.org/10.3233/IA-240033.

Angeli, G., Liang, P., & Klein, D. (2010). A simple domain-independent probabilistic approach to generation. *Conference on Empirical Methods in Natural Language Processing, USA, EMNLP 2010,* 502-512. aclanthology.org/D10-1049.pdf.

Van der Velde, M., Veldkamp, B. P., Keuning, J., Feskens, R. C. W., Swart, N. M., Harmsen, W. N. (2024). The framework and development of SERDA: Speech enabled reading fluency assessment for Dutch. In Ranđelović B., Karalić E., Aleksić K., Đukić D. (Eds.), *E-testing and computer-based assessment. CIDREE Yearbook 2024* (pp. 99–123). CIDREE. https://cidree.org/wp-content/uploads/2024/11/cidree_yearbook-2024.pdf

van der Velde, M., Harmsen, W., Veldkamp, B. P., Feskens, R.C.W., Keuning, J., & Swart, N. (2025). Speech enabled reading fluency assessment: A validation study. *International Journal of Artificial Intelligence in Education*, 1-27. https://doi.org/10.1007/s40593-025-00480-y.

van der Velde, M., Swart, N., Veldkamp, B. P., Feskens, R.C.W. (2026). *Personalizing Primary Reading Education: Detailed Reading Profiles Through Latent Modelling* [Manuscript submitted for publication]. University of Twente.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in higher education, 38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462.

Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive science, 4*(4), 379–426. https://doi.org/10.1016/S0364-0213(80)80010-3.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research, 65*(3), 245–281. https://doi.org/10.3102/003465430650032.

Campbell, C., & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability, 21*(1), 47–65. https://doi.org/10.1007/s11092-008-9063-x.

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in higher education, 31*(2), 219–233. https://doi.org/10.1080/03075070600572132.

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027.

Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: Implications for policy and practice. *Policy insights from the behavioral and brain sciences, 3*(1), 54–61. https://doi.org/10.1177/2372732215624931.

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. *ICALT: IEEE international conference on advanced learning technologies*, USA, 323–325. https://doi.org/10.1109/ICALT58122.2023.00100.

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education, 210*, 104967. https://doi.org/10.1016/j.compedu.2023.104967.

Falcão, P. T., e Peres, F. M. A., Morais, D. C. S., & da Silva Oliveira, G. (2018). Participatory methodologies to promote student engagement in the development of educational digital games. *Computers & Education, 116*, 161–175. https://doi.org/10.1016/j.compedu.2017.09.006.

van Geel, M., Keuning, T., Visscher, A. J., & Fox, J. P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *American Educational Research Journal, 53*(2), 360–394. https://doi.org/10.3102/000283121663734.

Gómez, S., Zervas, P., Sampson, D. G., & Fabregat, R. (2014). Context-aware adaptive and personalized mobile learning delivery supported by UoLmP. *Journal of King Saud University - Computer and Information Sciences, 26*(1), 47–61. https://doi.org/10.1016/j.jksuci.2013.10.008.

Grainger, J., & Segui, J. (1990). Neighborhood frequency effects in visual word recognition: A comparison of lexical decision and masked identification latencies. *Perception & psychophysics, 47*, 191–198. https://doi.org/10.3758/BF03205983.

Harmsen, W.N., van Hout, R., Cucchiarini, C. and Strik, H. (2025). Can ASR generate valid measures of child reading fluency? *INTERSPEECH 2025, Netherlands*, 2395–2399. https://doi.org/10.21437/Interspeech.2025-306.

Hattie, J. (1999). *Influences on student learning* [Inaugural Lecture]. University of Auckland. Retrieved on 6 February 2026 from https://www.researchgate.net/publication/237248564.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishi, E., Bang, Y. J., Madotto, A., Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730.

van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research, 85*(4), 475–511. https://doi.org/10.3102/003465431456488.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin, 119*(2), 254.

Kochmar, E., Vu, D.D., Belfer, R., Gupta, V., Serban, I.V., Pineau, J. (2020). Automated Personalized Feedback Improves Learning Gains in An Intelligent Tutoring System. *Artificial Intelligence in Education, AIED 2020, Lecture Notes in Computer Science*, 12164, 140-146. https://doi.org/10.1007/978-3-030-52240-7_26.

Li, K. C., Wong, B. T. M. (2020). Features and trends of personalised learning: a review of journal publications from 2001 to 2018. *Interactive Learning Environments, 29*(2), 182–195. https://doi.org/10.1080/10494820.2020.1811735.

Li, Z., Xu, X., Shen, T., Xu, C., Gu, J. C., Lai, Y., Tao, C., Ma, S. (2024). Leveraging Large Language Models for NLG Evaluation: Advances and Challenges. *arXiv preprint arXiv:2401.07103.* https://doi.org/10.48550/arXiv.2401.07103.

Lin, Y., Ruan, T., Liu, J., & Wang, H. (2023). A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering, 36*(4), 1431–1449. https://doi.org/10.1109/TKDE.2023.3304385.

McKeown, K. (1992). *Text generation*. Cambridge University Press.

Meelissen, M. R. M., Maassen, N. A. M., Gubbels, J., van Langen, A. M. L., Valk, J., Dood, C., Derks, I., In 't Zandt, M., & Wolbers, M. (2023). *Resultaten PISA-2022 in vogelvlucht.* Enschede: Universiteit Twente. https://doi.org/10.3990/1.9789036559461.

Merino-Campos, C. (2025). The impact of artificial intelligence on personalized learning in higher education: A systematic review. *Trends in Higher Education, 4*(2), 17. https://doi.org/10.3390/higheredu4020017.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199.

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 International Results in Reading.* Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342.

Nan, L., Hsieh, C., Mao, Z., Lin, X. V., Verma, N., Zhang, R, Kryściński, W., Schoelkopf, H., Kong, R., Tang, X., Mutuma, M., Rosand, B., Trindade, I., Bandaru, R., Cunningham, J., Xiong, C., Radev, D., Radev, D. (2022). FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics, 10*, 35–49. https://doi.org/10.1162/tacl_a_00446.

Newman, M., Kwan, I., Schucan-Bird, K., & Hoo, H. T. (2021). *The impact of feedback on student attainment: A systematic review*. Education Endowment Foundation. Retrieved on 6 February 2026 from discovery.ucl.ac.uk/id/eprint/10138571/1/Systematic-Review-of-Feedback-EPPI-2021.pdf.

Open WebUI. 2025. open-webui/open-webui. https://openwebui.com/.

Osuji, C. C., Ferreira, T. C., & Davis, B. (2024). A systematic review of data-to-text NLG. *arXiv preprint arXiv:2402.08496.* https://doi.org/10.48550/arXiv.2402.08496.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askel, A., Welinder, P., Christiano, P. F., Leike, J., Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems, USA, NeurIPL 2022, 35*, 27730–27744. proceedings.neurips.cc/paper_files/paper/2022.

Parikh, A., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., & Das, D. (2020). ToTTo: A controlled table-to-text generation dataset. *Conference on Empirical Methods in Natural Language Processing, EMNLP 2020,* 1173–1186. https://doi.org/10.18653/v1/2020.emnlp-main.89.

Razzaq, R., Ostrow, K.S., Heffernan, N.T. (2020). Effect of immediate feedback on math achievement at the high school level. *Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science,* 12164, 263–267. https://doi.org/10.1007/978-3-030-52240-7_48.

Reiter, E. (2007). An architecture for data-to-text systems. *The eleventh European workshop on natural language generation, Germany, ENLG 07,* 97–104. aclanthology.org/W07-2315.pdf.

Reiter, E., Dale, R. (1997). Building applied natural language generation systems. *Natural language engineering, 3*(1), 57–87. https://doi.org/10.1017/S1351324997001502.

Reiter, E., Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press.

Sack, N., Little, B. (2025). *The Risks of Personalising Higher Education with Artificial Intelligence. Ethical Risk Report*. Retrieved on 6 February 2026 from surf.nl/files/2025-04/the-risks-of-personalising-higher-education-with-artificial-intelligence-nati-sack-ben-little.pdf.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119–144. https://doi.org/10.1007/BF00117714.

Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School effectiveness and school improvement, 28*(2), 242-258. https://doi.org/10.1080/09243453.2016.1256901.

Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based decision making in education.* Springer.

Seo, H., Hwang, T., Jung, J., Kang, H., Namgoong, H., Lee, Y., & Jung, S. (2025). Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy. *Applied Sciences, 15*(2), 2076–3417. https://doi.org/10.3390/app15020671.

VanLehn, K. (2006). The behavior of tutoring systems. *International journal of artificial intelligence in education, 16*(3), 227–265. https://doi.org/10.3233/IRG-2006-16(3)02.

Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology, 22*(3), 407–423. https://doi.org/10.1002/acp.1414.

Wiboolyasarin, W., Wiboolyasarin, K., Suwanwihok, K., Jinowat, N., & Muenjanchoey, R. (2024). Synergizing collaborative writing and AI feedback: An investigation into enhancing L2 writing proficiency in wiki-based environments. *Computers and Education: Artificial Intelligence, 6*, 100228. https://doi.org/10.1016/j.caeai.2024.100228.

Zhao, Y., Zhang, H., Si, S., Nan, L., Tang, X., & Cohan, A. (2023). Investigating table-to-text generation capabilities of LLMs in real-world information seeking scenarios. *arXiv preprint arXiv:2305.14987.* https://doi.org/10.48550/arXiv.2305.14987.

Zheng, L., Zhong, L., Niu, J., Long, M., & Zhao, J. (2021). Effects of Personalized Intervention on Collaborative Knowledge Building, Group Performance, Socially Shared Metacognitive Regulation, and Cognitive Load in Computer-Supported Collaborative Learning. *Educational Technology & Society, 24*(3), 174–193. https://www.jstor.org/stable/27032864.

Chapter 7

Discussion

Throughout this dissertation, we investigated whether and how improving the assessment of reading fluency can reduce some of the issues that trouble the reading landscape. Specifically, we investigated options to reduce the heavy time burden placed on teachers by current reading assessments, to improve the objectivity and flexibility of assessment, to include prosodic features within assessment, and to increase the availability of detailed and informative diagnostics. These ambitions lead to the construction of the Speech Enabled Reading Diagnostics App (SERDA), the development and evaluation of which stands at the forefront of this dissertation.

To inform the construction of SERDA, we investigated how reading fluency is currently defined and assessed within the literature, as well as which instruments are available and how valid and reliable they are. Based on the results of the review, we developed and validated SERDA, a reading fluency assessment instrument that incorporates Automatic Speech Recognition (ASR) and Learning Analytics (LA) with the aim of automatically providing objective and detailed diagnostics on all reading fluency components. In addition, we investigated the potential of personalizing reading instruction by constructing reading profiles based on these diagnostics, and by evaluating whether these profiles could be used to automatically generate instructional feedback tailored towards children's learning needs.

We believe that the results of the dissertation indicate that it is possible to (eventually) reduce current reading fluency assessment issues. Firstly, we observed that applying ASR allowed for the automatic translation of speech into reading fluency diagnostics. In addition, we found that the measures resulting from SERDA are quite detailed, allowing for investigations at item, subtask, task and child level, as well as the variability of performances over tasks and subtasks. Furthermore, the validation provided evidence that, while many aspects of the generational process can be improved, SERDA's diagnostics have a reliable and valid foundation.

The profiling analyses provided evidence that SERDA's diagnostics can be used to construct didactically and practically useful profiles, providing means for teachers to personalize their reading instruction. The feedback generation study indicated that qualitatively sound feedback can be constructed based on SERDA's reading profiles and diagnostics, even though improvements are necessary before practical implementations can be considered. In short, we believe that SERDA shows promise in solving some of the major issues that currently plague the reading fluency landscape, and might meet this promise if sufficient attention, time and money is invested in furthering its development.

While the first steps towards realizing this framework have been described within this dissertation, additional work is required to breath further life into this foundation. As such, the remainder of this discussion will focus on this potential for growth, and the necessary future endeavors required to live up to this potential.

First, while a large sample of second- and third graders was obtained, this sample primarily allows for conclusions regarding SERDA's implementation within these specific

grades. Researchers would do well to expand the SERDA dataset, such that a representative sample becomes available for all relevant primary grades in the future. In addition, it might be worthwhile to consider the inclusion of alternative tasks and norms for specific subpopulations, such as for children with dyslexia or learning disorders.

Another way to improve SERDA would be to investigate its utility with longitudinal data. As things currently stand, it is only possible to obtain a snapshot of children's reading difficulties and qualities. In order for development to be charted, repeated measures should be collected from students throughout their entire early educational journey. As such, future research should focus on evaluating whether SERDA's diagnostics can be used to inform teachers about children's reading qualities over time. Specifically, it is interesting to investigate whether developmental reading profiles can be reliably and validly generated through SERDA. If attempted, it is likely that alternative or additional profiles would be observed, be it through timepoints, grades or an interaction of both. As for the question of whether SERDA should primarily prioritize on providing teachers with instructions based on snapshots or development: I will leave that up to a future discussion section.

Once the sample is sufficiently supplemented, it would be a useful addition to obtain norm scores for SERDA. As these are currently missing, it is difficult to make concrete comparisons between children in different grades and between children who conducted different tasks. These norm-based grades would allow teachers to make more general conclusions about children's reading performance, whereas they mostly learn about children's relative reading difficulties and qualities with the current version of SERDA. To exemplify, a child could be defined as an *Accurate Reader*, irrespective of how well the child reads compared to other children in their grade, as long as their reading accuracy is substantially better than their reading speed. As such, differentiating between a child that reads with great difficulty and a child that reads quite well would require additional work by the teacher. In a new version of SERDA, such information would be provided automatically.

Second, there are the design choices made during the construction of SERDA. While a rationale has been provided for favoring an online application, research has not clearly demonstrated to what degree online assessment is preferable to pen-and-paper alternatives. In addition, the effects of children's familiarity with technology, and SERDA specifically, should be further investigated. This is especially relevant here, as none of the children within the SERDA dataset had any previous experience with SERDA. Therefore, future research should investigate the behavior of SERDA's metrics over multiple administrations and time, such that familiarity effects can be clarified.

In addition, SERDA's use of ASR and LLM's within a relevant educational setting requires discussion. While ASR and generative AI have become increasingly popular over the years, some argue that they should never be used without some type of human supervision. This concept of keeping a "human in the loop" brings us to the question of

whether we are even happy that a machine can, perhaps, eventually relieve teachers of their assessment related duties.

Given that teachers tend to be in high demand and short supply, and given that the quality of instruction likely declines as their schedules overflow, reducing teacher's burdens was brought up as one of the primary reasons to automate reading fluency assessment. However, we believe teachers should remain active within the process of evaluating and instructing children's reading. In the first place, this is to reduce the potential of developing a dependency on an automatic system, but it is also to make sure teachers do not lose the skills to evaluate reading performances themselves, thereby losing the ability to correct the system. As such, we highly suggest using SERDA within approach that incorporate and value teachers. Indeed, it is in interaction with teachers where SERDA can truly shine.

In short, while promising results are presented within this dissertation, caution is warranted with regard to the implementation of SERDA in practice. Although the required improvements appear, and are, vast, they can be viewed as limitations of time. As such, future work can build upon the foundation provided here by improving each module of SERDA, one project at a time. Where alternative items or (sub)tasks are required, they can be constructed. Where additional data is required, it can be collected. Where ASR and LLM performances have undoubtedly faltered here, they will unquestionably succeed, in time, given the popularity of these approaches and the scientific progress made in these fields. As such, that which currently limits SERDA's implementation is likely to be resolved in time. Indeed, it can be viewed it as a matter of when, not whether, an instrument like SERDA will find its place in educational practice.

To finally and truly conclude, the work conducted throughout this dissertation provides evidence that combining ASR and LA can improve the assessment of reading fluency, potentially reducing some of the issues that exist within the reading landscape. Suggestions are provided to further the work discussed here, such that an implementation within educational settings can be realized. The most important among these concerns the expansion of available speech data, as this improves the potential informative breadth and depth, and the likelihood of an eventual practical implementation, for an instrument like SERDA. In addition, reducing teacher's burdens while not outright removing teachers from the developmental and assessment process is of the utmost importance if meaningful implementation is ever to be realized. While this provides the future of the field with a lot of work, together, we might reduce the increasing risk of functional illiteracy, hopefully improving children's ability to appreciate reading once more. An important endeavor I would label well worth the effort. After all, is it with reading that we truly feed the brain. Until it "is wider than the sky[49]".

[49] Dickenson, E. (1886-1896). *The Brain — is wider than the Sky*. https://acdc.amherst.edu/view/EmilyDickinson/ma00167-16-05-00049#page/1

Samenvatting

FOUTEN RECHT V(L)ECHTEN: VERBETER HET VROEGE LEESONDERWIJS MET AUTOMATISERING EN PERSONALISATIE.

Het belang van lezen is moeilijk te overschatten. Kunnen lezen is noodzakelijk om volwaardig deel te nemen aan onze zo geletterde samenleving, is een voorwaarde voor diepere en bredere scholing en ontwikkeling, en biedt de beheerser de mogelijkheid deuren te openen tot prachtige werelden en perspectieven. Toch laten de resultaten van grootschalige onderzoeken zien dat leesvaardig worden lang niet overal even goed gaat. Zo zien we in Nederland dat het risico op laaggeletterdheid flink is toegenomen de afgelopen jaren, ondanks dat de Nederlandse jeugd enkele decennia geleden nog relatief goed leek te kunnen lezen. Deze resultaten hebben tot veel ophef geleidt, wat op zijn beurt vertaald is in een scala aan onderzoeken en interventies.

Waar veel van deze aanpakken zich direct richten op begrijpend lezen, gaat minder aandacht naar vroegere vormen van lezen, zoals vloeiend lezen. Vloeiend lezen betreft de vaardigheid om, hardop, accuraat, snel en met expressie te kunnen lezen en wordt veelal gezien als een voorwaarde voor en voorspeller van begrijpend lezen. Deze relatie bestaat grotendeels via het automatiseren van lezen, welke een reflectie is van de accuratesse en snelheid van lezen. Zodra leerlingen bekwaam raken met accuraat en snel lezen, begint lezen automatisch te worden. Dit zorgt ervoor dat leerlingen minder aandacht hoeven te spenderen aan het lezen zelf, wat vervolgens kan worden ingezet om de tekst te begrijpen. Ook expressie faciliteert begrip. Dit gebeurt voornamelijk via de focus en aandacht die een kind plaatst op het leesproces, wanneer een kind probeert expressief te lezen.

In het kort indiceren deze bevindingen dat het verbeteren van de vloeiende leesvaardigheid een mogelijkheid biedt om de huidige en toekomstige leesvaardigheid van leerlingen te verbeteren. Er zijn momenteel echter enkele praktische redenen die het moeilijk maken om in te spelen op het opkrikken van de vloeiende leesvaardigheid. Zo bieden veel instrumenten geen informatie over expressie, resulteert de afname van veel instrumenten niet in gedetailleerde informatie, en kost het leerkrachten vaak veel tijd om leesinstrumenten af te nemen. Het gevolg hiervan is dat het lastig is om de individuele leeskwaliteiten- en moeilijkheden van leerlingen scherp in beeld te krijgen, en om een persoonlijk ontwikkelplan te bewerkstelligen.

Binnen deze dissertatie onderzoeken we of het mogelijk is om deze problemen te reduceren via de toepassing van diverse technologische ontwikkelingen binnen het meet- en instructieproces van vloeiend lezen. De dissertatie is een onderdeel van het overbruggende ASTLA project, welke door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO: NWO 406.20.TW.009) is gefinancierd. Specifiek richten de dissertatie en het ASTLA project zich op het optimaliseren van de leesontwikkeling door middel van Automatische Spraakherkenning (ASR) en Learning Analytics (LA). Het doel is deze aanpakken te combineren binnen een nieuw meetinstrument, the Speech Enabled Reading

Diagnostics App (SERDA). Deze dissertatie richt zich dan ook specifiek op de constructie en validatie van SERDA.

In hoofdstuk twee presenteren we een scoping review, wat een specifieke vorm van een systematisch literatuuronderzoek betreft. Binnen deze review onderzochten we hoe vloeiend lezen in de literatuur wordt gedefinieerd en gemeten, en hoe betrouwbaar en valide huidige instrumenten zijn. De review resulteerde in een systematisch overzicht van veelgebruikte meetinstrumenten voor vloeiend lezen, evenals hun kwaliteiten en tekortkomingen. Dit overzicht heeft tevens als inspiratie gediend voor de ontwikkeling van SERDA. Een ander relevant resultaat betreft de bevinding dat huidige meetinstrumenten hoge betrouwbaarheid en validiteit rapporteren, maar zelden expressiviteit evalueren. Dit geeft een gat in de literatuur bloot, betreffende hoe vloeiend lezen wordt gedefinieerd en gemeten, wat zorgen oproept met betrekking tot de werkelijke validiteit van huidige methoden.

In hoofdstuk drie beschrijven we het raamwerk achter, en de ontwikkeling van, SERDA. Specifiek bespreken we hoe de incorporatie van ASR en LA mogelijk het meten van vloeiend lezen kunnen verbeteren. Ook informeren we over de constructie van de leestaken, en de verzameling van de SERDA dataset. De studie resulteerde in de constructie van SERDA, en de afname van SERDA bij een steekproef van 653 Nederlandse basisscholieren. Verder worden enkele vroege bewijzen aangeleverd over de bruikbaarheid, betrouwbaarheid en validiteit van SERDA, welke diens potentie voor implementatie in de praktijk informeren.

In hoofdstuk vier bieden we een discussie aan over de validiteit van SERDA's accuratesse, snelheid en automaticiteit maten. Hiervoor gebruiken we de Argument-Based-Approach to validation (ABP) van Kane. Specifiek beantwoorden we of een instrument wat vloeiend lezen meet, door middel van ASR, valide leesvaardigheid scores oplevert. De resultaten indiceren dat betrouwbare en valide maten kunnen worden gegenereerd door een op ASR gebaseerd meetinstrument voor vloeiend lezen.

In hoofdstuk vijf evalueren we de potentie van het gebruik van de gedetailleerde diagnostische gegeven van SERDA om distinctieve, praktisch en didactisch relevante leesprofielen te construeren. Ook hier hanteren we de ABP als validatiemethode, ditmaal om de kwaliteit van de optimale set aan profielen te beoordelen. Tijdens deze beoordeling speelde de praktische bruikbaarheid van de profielen een grote rol. De resultaten van het onderzoek bieden bewijs dat het mogelijk is om distincte, praktisch en theoretisch relevante leesprofielen te construeren op basis van SERDA's leesgegevens.

In hoofdstuk zes evalueren we of Large Language Models (LLM's) gebruikt kunnen worden om SERDA's leesgegevens, en de leesproefielen uit hoofdstuk vijf, gebruikt kunnen worden om instructionele feedback te construeren voor leerkrachten. Daarnaast hebben we onderzocht of de kwaliteit van feedback wordt beïnvloed door de mate van contextualisatie, zoals aangeboden aan de LLM's. Specifiek vergeleken we de leesbaarheid, coherentie, didactische kwaliteit, het aantal fouten en de lengte van menselijke feedback met feedback

zoals geconstrueerd in drie LLM condities, welke varieerde met betrekking tot de mate van contextualisatie. De resultaten van het onderzoek indiceren dat LLM's feedback genereren van relatief hoge kwaliteit. Feedback van LLM's waren relatief prettig leesbaar, en waren van bescheiden lengte. Wel toonde feedback uit de LLM condities een lagere coherentie en didactische kwaliteit dan de feedback van mensen. Om deze reden is het aan te raden de LLM's verder te optimaliseren, voordat praktische implementatie wordt overwogen.

In hoofdstuk zeven bespreken we, ter conclusie, het geheel aan werk dat heeft plaatsgevonden om de dissertatie tot stand te brengen, inclusief enkele suggesties voor vervolgonderzoek. Op basis van de resultaten van de diverse onderzoeken concluderen we dat de dissertatie bewijs toont dat het mogelijk is om het meten en instrueren van vloeiend lezen te verbeteren door middel van een combinatie van ASR en LA. Wel is hiervoor een diversiteit aan vervolgonderzoek nodig, welke als voorwaarde dienen voor de praktische implementatie van instrumenten zoals SERDA. Ten eerste dient de data op diverse manieren uitgebreid te worden, zodat zowel de diepgang als breedte van de aangeleverde informatie kan worden verrijkt. Ten tweede suggereren we dat het belangrijk is om een goede balans te vinden tussen het ontlasten van leerkrachten, zonder dat dit ze vervangt binnen het ontwikkel- en meetproces van hun leerlingen, mocht een betekenisvolle implementatie vorm willen krijgen.

Personal Acknowledgements / Dankwoord

Before making things personal, I would like to show my deep appreciation for the Dutch Research Council (NWO). In a time of mis-information, the importance of science ever-increases. Without an organisation like the NWO, none of the work conducted here could have ever been made possible. As such, I would like to once more thank the NWO for making the research within this dissertation possible through one of their research grants (ID: NWO 406.20.TW.009).

Now, on to the personal side. Let me start of by thanking you, the reader, for making the work gathered within this potentially overly lengthily written book more meaningful. After all, a problem shared is a problem halved, and the reading crisis is quite the beast of a problem. Thankfully, if you are reading this (even if only to find out if I mentioned you by name) you are already actively participating in combating this horrid situation, good on you! Through the remainder of this section I will provide my personal thanks to colleagues, friends and family in Dutch, such that I can most accurately articulate my appreciation. Thank you for understanding.

Ten eerste zou ik graag mijn begeleiders bedanken. Bernard, ik kan je niet genoeg bedanken voor alle mogelijkheden die je me hebt geboden om te leren, en de vrijheid die je me hebt gegeven om dit leerproces mij eigen te maken. Je hebt me alle kanten van de wetenschap laten zien, de mooie en minder mooie, de leuke en minder leuke. Daarnaast kan ik je interesse om te vertellen en vragen over ook niet wetenschappelijke zaken enorm waarderen. Zeker als het gaat over je diverse (variabel succesvolle) schaats-tripjes. In die zin ben ik een beetje jaloers op je toekomstige PhD-kandidaten. Dat ze door mogen hebben dat ze in goede handen zijn. Ik weet dat ik het was.

Remco, ook jou ben ik dankbaar voor de goede en geduldige begeleiding. Van stage tot masterthesis tot promotie tot mijn eerste “grote mensen” baan, het lijkt bijna alsof jij altijd een klein oogje in het zeil weet te houden. Mijn dank is dan ook groot voor de vele mogelijkheden die je me hebt blijven durven geven, en de kritische blik die jij altijd zo snijdend werpt op mijn (veel) te fladderige teksten. Daarnaast waardeer ik het enthousiasme wat je overdraagt voor alles wat je leuk acht, ware het specifieke psychometrische modellen, dan wel het elo-systeem achter ons pingping klassement. Ik hoop je deze dank de komende jaren verder te kunnen bieden in het werk wat ik bij cito mag gaan doen, hetgeen me een eer blijven zal.

Jos, ook jou noem ik graag bij naam hier. Ondanks dat alles anders is gelopen dan verwacht en gehoopt, heb ook jij een aanzienlijke bijdrage geleverd aan mijn proefschrift en ontwikkeling. Op inhoudelijk, praktisch en technisch vlak heb je het project in goede banen weten te leiden, maar ook in de begeleiding heb je me enorm veel perspectief geboden. Ja, misschien is het juist op persoonlijk vlak waar ik het meeste van je geleerd heb. Ik acht zelf dat juist die lessen mij zullen helpen in de lange carrière die ik beoog te hebben.

Daarnaast zijn er nog diverse collega’s binnen het ASTLA project die een flinke bijdrage hebben weten te leveren aan de goede afloop van dit boekje. Nicole, jou wil ik

als eerste bedanken. Jouw enthousiasme, energie, kennis en humor hebben de meest saaie taken dragelijk weten te maken. Ik hoop dat we elkaar nog veel tegen komen in leesland, en dat we nog vele projecten samen tot een goed einde mogen brengen. Eerst maar eens kijken of we nog een keer toe komen aan die veld-publicaties… Op dat er altijd maar een reden mag zijn dat je de slingers uit hangt.

Bo, enorm bedankt voor de prachtige start die ik mede dankzij jou heb mogen maken aan deze draak van een opdracht. Ik heb onze gesprekken en gedeelde werkzaamheden enorm gewaardeerd en had de review vast nooit overleefd als ik jou niet als tweede auteur had. Ik ben dankbaar voor de vriendschap die we onderhouden en hoop dat we weer veel te bespreken hebben over een maand of 6.

Wieke, soortgelijk zou ik jou graag bedanken voor alles wat jij gedaan hebt om ook mijn proefschrift tot een soepele afronding te brengen. Ik kan me enkel voorstellen hoeveel druk er op de positie stond die jij hebt weten te vullen en ben enorm dankbaar dat binnen die druk ook aandacht bleef bestaan voor mijn deel van het project. Het resultaat is een diversiteit aan mooie publicaties en wat mij betreft een goede basis voor meer. Maar eerst maar even afstuderen, of niet? Sterkte.

Catia, Helmer en Christian, ook jullie wil ik expliciet bedankten voor alles wat jullie binnen het ASTLA project hebben gedaan. Zonder jullie zou het project nooit hebben bestaan, en zou ik mogelijk nooit de kansen hebben gekregen die ik nu heb gehad. Jullie duidelijke passie voor het vak was aanstekelijk en jullie hebben mij elk veel geleerd over hoe een succesvolle wetenschapper te werk gaat. Op een mooie toekomst voor ASTLA, SERDA en leesvaardigheid.

Vervolgens zou ik graag mijn dank uiten aan mijn collega's buiten ASTLA. Collega's van CODE, enorm bedankt voor het warme welkom en de vele enerverende gesprekken. Specifiek wil ik Ellen nog bedanken. Man man man wat heb ik jou vaak moeten storen met het één of het ander. Enorm bedankt voor al je geduld. Ook Jolien en Sandra bedank ik graag voor het warme welkom, de vrijheid om op mijn manier les te geven en jullie passie voor doceren, deze was enorm aanstekelijk. Johannes en Erik, bedankt voor de mogelijkheden binnen jullie projecten, de hilarische en leerzame gesprekken en jullie perspectief op "de wetenschapper". "Kebab Max" moeten we absoluut nog een keer werkelijkheid maken.

Jitske, het spijt me nog steeds dat ik je niet heb weten te herkennen, hopelijk gaat dit tijdens de verdediging, dan wel in de toekomst, beter. Zonder dollen tho, enorm bedankt dat ik deel heb mogen uitmaken van jouw prachtige project. Daarnaast ben ik dankbaar voor je humor, oprechtheid en relativeringsvermogen. Ik zal er eeuwig jaloers op zijn hoe alles van jou af lijkt te glijden en hoop er een les uit te mogen leren. Misschien wel binnen een volgend project?

Collega's van citolab, ook jullie dank ik graag voor het warme bad waar ik de afgelopen 5 jaar in heb mogen rondzwemmen. De gezelligheid en bereidheid elkaar te helpen acht ik echt een pracht, hetgeen me altijd trots heeft gemaakt met jullie te mogen werken

(hoe bescheiden dat tot op kort ook was). Specifiek zou ik graag Karen, Hendrik en Bas bedanken, waar mijn naschoolse carrière ooit is begonnen. Jullie hebben me enorm veel geleerd in een periode waarin ik nog koppiger was dan ik nu ben. Het is een understatement om te zeggen dat ik in goede handen was en mijn passie voor de psychometrie is mede dankzij jullie zo sterk. Daarnaast bedank ik graag Patrick, Marcel en Tijmen, wie samen met mij een draak van een dashboard tot iets moois hebben gebracht. Ik hoop dat we nog veel mooie dingen mogen ontwikkelen samen.

Ook bedank ik graag alle studentassistenten die mij hebben geholpen met de toch aardig gigantische dataverzameling gedurende de eerste twee jaar van mijn proefschrift. Zonder jullie was het niet gelukt! Tegelijkertijd bedank ik graag alle scholen, schoolleiders, leerkrachten en leerlingen die hebben bijgedragen aan het onderzoek. Ik had jullie graag direct veel teruggegeven, maar weet dat ik het mijn doel zal blijven maken om de leesvaardigheid in Nederland te verbeteren.

Vervolgens dank ik graag mijn PhD collega's. Eva en Aranka, dank jullie voor de veiligheid die jullie me hebben geboden in de rare wereld van het promoveren. Lara, dank voor de goede gesprekken, de heftige schrikreacties, en de begrijpende blik. Ik waardeer jouw passie voor de wetenschap en je onverwacht scherpe meningen over de wereld van vandaag. Ik hoop dan ook dat we nog veel samen mogen kletsen over hoe het allemaal moet en zou en niet. Je zult vast een prachtig proefschrift in elkaar zetten, ik kijk er alvast naar uit het te lezen. Luca, I did not forget you, how could you even think about it! Just joking, thanks for being the best Italian a guy could know when visiting the lovely city of Bologna, let's do Rome next, or would you like to visit Arnhem again? Also thank you for our many interesting discussions regarding statistics, data and everything life has to offer. I hope we get many chances to work together in our new, ever so similar, jobs. Keep up the grind. Een klein bedankje gaat ook uit naar de nieuwe PhD's die ik heb mogen leren kennen (specifiek noem ik graag Senne (dus toch!), Joost en Paul), jullie kersverse interesse en nieuwsgierigheid heeft me geholpen de streep te halen en deze te passeren. Weet me te vinden als je hulp nodig hebt of gewoon even wil nerden over je nieuwste idee.

Ten slotte zijn er natuurlijk nog de mensen buiten mijn werkomgeving die een flinke bijdrage hebben geleverd aan dit proefschrift. De mensen die mij door de turbulente delen van het proefschrift hebben gedragen en de mensen waarmee ik hoop de rest van mijn leven te delen. Lars, wat een prachtige reis hebben we gemaakt de afgelopen tien jaar en wat zijn we naar elkaar toegegroeid. De kers is natuurlijk Japan. Wat een rijkdom aan ervaringen hebben we ook daar weer op gedaan. Ik kijk er naar uit de komende tien jaar mee te maken, maar dan zul je me wel dr. Max moeten noemen. Je wist dat deze dag zou komen. Florian, mijn medewandelaar op het pad wat we de wetenschap noemen. Jou bedank ik graag voor de rijkdom aan gesprekken en de hopelijk eeuwig blijvende wandelingen. Van onze favoriete H.P. tot de definitie van altruïsme, mijn kritische blik naar alles wat is, is toch ook tezamen met (door?) jou ontwikkeld. En wat zijn we ver gekomen. Ja, ik stap nu eerst door

de poort des dr., maar vrees niet. Ik zal em op een kiertje houden en zie je graag (zonder twijfel) snel aan de andere kant. Patrick, nog zo'n bron van kritische noten en blikken. Van perspectieven en enthousiasme, sequentiële hyperfocus zullen we het maar noemen? Ik ben jou, naast onze toch al best durende vriendschap, dankbaar voor de diversiteit aan perspectieven die jij altijd weet aan te nemen, waar we het ook over hebben. Ondanks dat we nu allebei een nieuwe fase ingaan hoop ik dat we elkaar kunnen blijven vinden, onze gesprekken zijn en blijven me enorm waardevol. Danny, de enige overgeblevene van een web aan klimmers. Opmerkelijk wat een mooie vriendschap we hebben weten te creëren de afgelopen paar jaren en hoeveel tijd we voor elkaar weten blijven te vinden. Ondanks dat jij ook een ander leven tegemoet gaat hoop ik dat we die tijd, hoe we het ook voor elkaar krijgen, aan elkaar kunnen blijven geven. Vraag en ik sta klaar!

Als laatste dank ik graag mijn familie, beginnende bij zij die er niet meer zijn. Opa van der Velde, u bent het geweest die altijd het trotste wist te zijn op mijn academische prestaties. Hoe oppervlakkig het misschien ook is, zo belangrijk was het voor mij tijdens mijn matige tienerjaren. Het is altijd een doel geweest u trots te maken, hetgeen ik hiermee hoop (wederom) te hebben bereikt (ondanks dat ik nog steeds geen oesters lust…). Opa Gonlag, zonder u was ik hier ook nooit gekomen. Niet zonder uw liefde, uw enthousiasme, uw acceptatie van alles ik. De liefde voor presenteren en kletsen heb ik van u, ik hoop dat mijn collega's u ooit zullen vergeven.

Oma Gonlag, als er iets is waar ik u dankbaar voor ben, dan is het wel het goede voorbeeld dat u altijd weet te zijn. Nee, niemand is perfect, zo zegt u zelf, en ik kan niet anders dan instemmen. Maar toch, de manier waarop u altijd eerst aan een ander weet te denken is en blijft inspirerend. Ik hoop zelf ook ooit zo'n goed persoon te zijn. Gelukkig maar dan dat het voorbeeld me altijd scherp zal blijven. Oma van der Velde, ook u ben ik al een leven lang dankbaar. Altijd die: hoe gaat het nou, waar ben je mee bezig, hoest met je meisje, zet em op he. Die eeuwige interesse, acceptatie van alles wat ik uitkraam en oneindige steun verklaren in grote mate hoe ik dit boekwerk voor elkaar heb weten te schoppen. Maar ook de kritische blik. Nee, ik kwam bij jullie nooit zomaar ergens mee weg. Dat is, zo zullen mijn promotoren ook wel beamen, zeker blijven plakken.

Sanne, zuslief. We hebben elkaar mogelijk niet zoveel gezien als we zouden willen de afgelopen jaren. Ik door mijn proefschrift, jij door die fantastische dochters van je. Weet dat ik veel bewondering heb voor wat je allemaal weet te doen, hoe moeilijk het je soms ook af gaat. Het is oprecht inspirerend. Verder ben ik je dankbaar voor de mooie gesprekken die we hebben als we de tijd wel weten te vinden, want wat lijken we stiekem toch veel op elkaar en wat hebben we toch maar met dezelfde (en verschillende) dingen te dealen. Ik zie (spreek) je snel.

Arjen, jij ontspringt de dans ook niet. Hoe jij de ballen allemaal in de lucht weet te houden is me ook wat. 2 dochters, mijn zus en een klas vol kinderen. Ik doe het je niet na. Weet dat ik jouw kritische blik enorm waardeer en je gewilligheid mee te denken

(en spreken) over alles wat is en moeten zou. Ik zal je (perspectief) vast nog vaak nodig hebben, nu bij cito werkende, om een beetje te aarden in de praktijk. Daarnaast, boven alles, natuurlijk ook voor het goede gesprek. Waarde op zich.

Om toch bijna tot een einde te komen benadruk ik graag dat het werk wat hier is gedaan nooit (door mij) had kunnen worden uitgevoerd als mijn ouders er niet waren geweest. Anita, Edwin, paps en mams. Dank jullie dat jullie me hebben opgevoed naar het beste van jullie kunnen. Dank jullie dat jullie, en jullie liefde, er altijd is geweest. Het is alles waar een kind om wensen kan. Dank ook dat jullie me hebben ondersteund in (bijna) alles waar ik mijn hoofd maar naar heb willen zetten. Jullie hebben me de veiligheid en zekerheid geboden om vrijwel alles wat het leven te bieden heeft te overkomen, uiteindelijk. Ik hoop dat we nog veel wandelingen mogen maken en veel prachtige films mogen ervaren, samen.

Tot slot, Dees. Het mag duidelijk zijn dat we iets speciaals hebben. Geen van ons had ooit makkelijk kunnen worden genoemd, ware het niet dat we elkaar hebben gevonden (en dan nog). Je hebt me veel gegeven in die toch korte tijd dat we elkaar kennen. Misschien is het belangrijkste wel een stukje rust. Je hoeft niet lang naar dit document te kijken om te realiseren dat ik niet graag stil zit, en toch is het af en toe enorm belangrijk. Waar ik hier, voor jou, zelden gelegenheid voor wist te vinden, gaat het nu als vanzelf. Zoals veel als vanzelf gaat met jou. Wat kan ik anders zeggen dan, ja, 1hersencel hé. Ik kan zo wel even doorgaan, en dat zal ik zeker doen, maar je ligt nog te slapen. Weet echter, dat terwijl jij in dromenland verkeerd ik droom van de toekomst die ik met je voor me zie. Want ondanks dat mijn toekomst, nu na dit proefschrift, weer even oogt als een ongepaveerd pad, moge één ding helder zijn. Jij staat er middenin.

Dank voor ook iedereen die niet expliciet genoemd is, voor de wereld waarin we verkeren en voor het feit dat lezen bestaat. Perspectief is mij alles en nergens is zoveel te vinden als in een goed boek. Dat iedereen maar de mogelijkheid krijgen mag om dat te ondervinden.