Social Functioning and Reading Proficiency:

Validity of Educational Assessments Used in Norwegian Elementary Schools

Anne Arnesen

Thesis for the degree of PhD

Department of Special Needs Education
Faculty of Educational Sciences

UNIVERSITY OF OSLO

2017

Series of dissertations submitted to the
Faculty of Educational Sciences, University of Oslo
No. 284

ISSN 1501-8962

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Acknowledgements

Decades of clinical meetings focusing on the developmental challenges and concerns experienced by struggling students, their parents and teachers have inspired the research presented in this thesis resulting from a four-year PhD project at the Department of Special Needs Education (ISP), University of Oslo (UiO).

First of all, I am very thankful to all participating students, to their parents for completing the questionnaire, to the principals and teachers for completing the survey, and to the teachers for administering reading tests and rating their students' social skills. Hopefully, the results may serve to promote knowledge about the educational assessment practice in Norwegian elementary schools and to improve approaches for early identification and support of the difficulties many students are facing.

I will express my sincere gratitude to my supervisors Monica Melby-Lervåg, ISP and Terje Ogden, the Norwegian Center for Child Behavioral Development (NUBU). They are the most eminent Norwegian experts in the fields of reading and social behavior. It is a great honor to be included to their world of knowledge and I am honored for their contributions and different perspectives to my work. Their insights and extensive academic experiences supported and encouraged my project and a four-year cycle of manuscript revisions. I have been very privileged to have them on my team and as co-authors.

I am deeply thankful to NUBU and ISP for providing me opportunities to do the PhD project. Additionally, to the National Graduate School in Educational Research (NATED) and the Child, Language and Learning group (CLL) for providing me with excellent research conditions, seminars and courses. Many thanks to NATED and the Norway-America Association which provided me visiting scholar funding for two semesters at the Center on Teaching and Learning (CTL) in the College of Education at the University of Oregon (UO) and the Oregon Research Institute (ORI). I am indebted to Hank Fien and Scott Baker, CTL and Keith Smolkowski, ORI who hosted me and made my scholarly visits valuable and productive. Special thanks to Scott and Keith for joining me as co-authors, outstanding supervision, sharing excellent substantial and methodological knowledge, and valuable discussions of data analyzing. I will also express my deepest gratitude to my significant long-lasting friend and colleague at UO, Anne Todd, who has edited several of my manuscripts. Anne and her husband, Stuart Perlmeter, cared safely for my partner Tor and me while we were in Oregon. I am grateful to Nancy Knutson, my Viking sister and inspiring mentor, who introduced me to the highly

respected pioneers at the Oregon Social Learning Center and UO, namely Gerald Patterson, Hill Walker, Rob Horner and Jeffrey Sprague. Collectively, they have created unforgettable Oregon memories and important contributions to my work.

Special thanks to my workmate, Wilhelm Meek-Hansen, for walking the line with me at NUBU for many years, collecting data and co-authoring. He motivated me while establishing the basic ideas which this PhD project is founded. Many, thanks to Johan Braeken serving as a co-author and for patience while guiding me through data analyzing, psychometrics and growth modeling. Also, appreciation to Ronny Scherer for the initial modeling and discussions, and to Ernst Ottem and Jørgen Frost who inspired this work.

I am honored to Thorleif Lund, my first statistics teacher and supervisor, who from a source of wisdom stimulated my research curiosity, and to Thormod Idsøe, who provided valuable comments at the midway and final review phases. Thanks to Øivind Hoff for IT-support, Siva Rajah and Tora Monsrud for research assistance, Elisabeth Askeland, Terje Christiansen, Monica Dalen, Linda Larsen, Ingrid Madslien, Heidi Mjelve, Kathrine H.-Omdal, Anne-Lise Rygvold, Marika Vartun, and many other colleagues at NUBU, ISP and CLL for inspiring words and hugs. Finally, to the best fellow PhD students, and in particular to Hanne Hjetland for fabulous company across the finishing line, Arne Rødvik for awesome coffee breaks, Anette Andresen for wisdom and understanding, Linn Guttormsen and Silje Systad for joyful room company, and May-Britt Monsrud for humor and energy keeping the “old girls” going both inside and outside the monastery.

Last, but not least, I am very honored and thankful to the incredible pillars that made the hard work more fun than work: Heidi, Ida, Linn-Cathrin, Julie, Rikke Sofie, Oda, Maja, Sofia and Alma Luna, from whom I have learned the main purpose of life as they experienced developmental pathways to become successful, strong and healthy members of our society. The family´s great boys and many good friends for helping me focus on the life outside of work during my academic breaks. My parents and big brother, Per, for giving me a safe base for growth and for encouraging me to be a life-long learner. Finally, Tor, who not only encouraged me to go for the PhD, but also believed that I could do it. Thank you for your love, patience, understanding, respect, and for making our life so much fun and open minded.

Oslo, September 11, 2017
Anne Arnesen

Abstract

Social functioning and reading proficiency are critical for success in school and society. However, many students struggle in one or both of these areas. It is widely known that accurate identification of the specific difficulties that students are facing is a key to preventing social and academic failure. This requires valid and efficient educational assessments to guide effective interventions and monitor students' progress. We examined the quality of assessments for use in Norwegian elementary schools in three studies:

First, in a survey about schools' use of assessments targeting children's social functioning and reading proficiency, we found that reading assessments were used three times more often than assessments of social functioning. Moreover, there were more assessments available for reading than for social functioning. The review based on the survey showed that the psychometric quality of most of the used assessments was overall weak or undocumented, while the assessments' material quality was generally good.

Second, in the validation study of the Elementary Social Behavior Assessment (ESBA) using an 8-week test-retest design, the Explorative Factor Analysis (EFA) and the Confirmatory Factor Analysis (CFA) established construct validity finding that both one and two factors may be useful for Grades 1 to 6. High score and test-retest reliabilities were also found. Correlation analyses between teachers' ratings on the ESBA and the Social Skills Rating System (SSRS-T) established criterion validity which was consistent after controlling for students' background.

Third, the study of the Norwegian adapted Oral Reading Fluency (ORF) measure using a second-order latent growth curve model, showed measurement invariance across one school-year for Grades 2 to 5. Even though, initial individual differences varied more than growth rates, growth was positive for all participating students. We found relatively high stability in ORF scores within and across Grades 2 to 5. Criterion validity was established and the ORF and the National tests and assessments in reading correlated moderate to strong.

The weak and undocumented psychometric quality of the assessments used to identify students at risk in Norwegian elementary schools demonstrates a need for improvement of the assessment practice. The Norwegian adapted ESBA and ORF screeners will probably contribute to the pool of high-quality assessments for use within and across school-years to identify specific difficulties, guide interventions and monitor students' growth in social functioning and reading.

Contents

ACKNOWLEDGEMENTS ..... **III**
ABSTRACT ..... **V**
LIST OF PAPERS ..... **XI**
1. INTRODUCTION **1**
1.1 STRUCTURE AND OBJECTIVES 2
2. SOCIAL FUNCTIONING AND READING PROFICIENCY **4**
2.1 THE ROLE OF INDIVIDUAL AND ENVIRONMENTAL FACTORS 4
2.2 PATHWAYS TO SOCIAL FUNCTIONING 5
2.2.1 Social skills - social behaviors - social competence *6*
2.2.2 Learning-related social skills *8*
2.2.3 Difficulties in social functioning *8*
2.3 PATHWAYS TO READING PROFICIENCY 9
2.3.1 Reading comprehension and the Simple View of Reading *10*
2.3.2 Decoding and fluency *11*
2.3.3 Difficulties in reading *12*
2.4 COMORBIDITY IN DIFFICULTIES OF SOCIAL FUNCTIONING AND READING 14
2.5 PREVALENCE OF DIFFICULTIES RELATED TO SOCIAL FUNCTIONING AND READING 14
3. EDUCATIONAL ASSESSMENT PRACTICES AND CHANGES **17**
3.1 ASSESSMENT FOUNDATION 17
3.2 ASSESSMENT APPROACHES 17
3.3 ASSESSMENT PRACTICES 19
3.4 THEORY OF CHANGE MODEL FOR IMPROVING ASSESSMENT PRACTICE 20
4. FOUNDATION OF ASSESSMENT VALIDITY **24**
4.1 EVIDENCE-BASED ASSESSMENTS 24
4.2 RELIABILITY 25
4.3 VALIDITY AND VALIDATION 25
4.3.1 Construct validity *26*
4.3.2 Content validity *27*
4.3.3 Criterion-related validity *27*

4.3.4 The unified concept of validity 28

4.3.5 Domain model of construct validity 29

5. METHODOLOGICAL CONSIDERATIONS 31

5.1 DESIGNS 31

5.2 PARTICIPANTS, SAMPLES AND SELECTION PROCEDURES 31

5.3 METHODS AND MEASURES 34

5.3.1 Evaluation of assessment quality (Study 1) 34

Survey 34

Systematic review of literature 35

The EFPA evaluation review model 35

5.3.2 Social functioning (Study 2) 36

The ESBA 37

The SSRS 37

5.3.3 Reading proficiency (Study 3) 38

The ORF 38

The National Tests of Reading Proficiency 38

5.4 STATISTICAL METHODS OF ANALYSIS 39

5.4.1 Descriptive orientation 39

5.4.2 Inter-rater agreements 39

5.4.3 Multilevel analysis – intra-class correlation 40

5.4.4 Common factor model – structural equation modeling 40

EFA and CFA 41

Longitudinal growth factor structural equation modeling 41

Estimation methods 42

Model fit 43

5.5 MISSING DATA 44

5.6 ETHICAL PERSPECTIVES 45

6. SUMMARIES AND DISCUSSION OF MAIN FINDINGS 47

6.1 ASSESSING SOCIAL FUNCTIONING AND READING PROFICIENCY (STUDY 1) 47

6.1.1 Common use of assessments without documented evidence 47

6.1.2 Weak theory-based constructs and lack of pschometric evidence 48

6.1.3 The gaps in assessment practice and competence 49

6.2 VALIDATION OF THE ELEMENTARY SOCIAL BEHAVIOR ASSESSMENT (STUDY 2) 50

6.2.1 Academic engagement and peer social relations 50

6.2.2 Consistency in teachers' ratings of students' social skills 50

6.2.3 The ESBA is a valid screener to guide specific social skills intervention 52

6.3 GROWTH IN ORAL READING FLUENCY (STUDY 3) 52

6.3.1 Longitudinal invariance and a measure of growth in reading 53

6.3.2 The ORF measure - an indicator to identify reading dificulties for interventions 53

6.4 LIMITATIONS OF THE STUDIES 54

6.5 CONCLUSION AND FURTHER PERSPECTIVES 56

REFERENCES 59

PAPERS I - III

APPENDICES A- H

ERRATA

List of Papers

Paper I: Arnesen, A., Braeken, J., Ogden, T., & Melby-Lervåg, M. (2017). Assessing Students' Social Functioning and Reading Proficiency: A Systematic Review of the Quality of Educational Assessment Instruments used in Norwegian Elementary Schools. *Resubmitted for publication to Scandinavian Journal of Educational Research*

Paper II: Arnesen, A., Smolkowski, K., Ogden, T., & Melby-Lervåg, M. (2017). Validation of the Elementary Social Behavior Assessment: Teacher ratings of students' social skills adapted to Norwegian, Grades 1 to 6. *Emotional and Behavioural Difficulties*. doi: 10.1080/13632752.2017.1316473

Paper III: 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. doi:10.1002/rrq.159

1. Introduction

Education is the most significant factor related to future possibilities for at-risk students (Gustafsson et al., 2010; Stipek, 2001). It is well known that social functioning and reading proficiency are critical to students' success in school and society (Cooper, Moore, Powers, Cleveland, & Greenberg, 2014; Duncan et al., 2008). However, many students face difficulties in one or both of these areas (Miles & Stipek, 2006; The Norwegian Ministry of Education, 2009, 2017; Walker, Ramsey, & Gresham, 2003). Difficulties in social functioning, which include internalizing (keeping to oneself) and externalizing (being heard and seen) behavioral disorders, are frequently reported to coexist with severe reading difficulties, such as dyslexia (Dahle, Knivsberg, & Andreassen, 2011; Terras, Thompson, & Minnis, 2009; Undheim, Wickstrøm, & Sund, 2011). Moreover, early reading difficulties are risk factors for internalizing and externalizing behavioral disorders (Carroll, Maughan, Goodman, & Meltzer, 2005; McIntosh, Horner, Chard, Boland, & Good, 2006; Trzesniewski, Moffitt, Caspi, Taylor, & Maughan, 2006). That is, students who struggle with social functioning and/or reading may be at risk for social and academic failure that negatively impacts their education.

Research over the past decades has provided knowledge regarding the development of students' social functioning and reading proficiency and the importance of preventing difficulties in these areas (Fuchs & Fuchs, 1986; Stecker, Fuchs, & Fuchs, 2005; Quirk, Dowdy, Goldstein, & Carnazzo, 2017). Studies have demonstrated the importance of identifying students who do not respond as expected to universal interventions that aim to promote students' social behavior and academic achievements (e.g., Elliott, Huai, & Roach, 2007; Jones, Greenberg, & Crowley, 2015). To identify and understand the difficulties these students face, it is critical that the assessments are of high quality and are used as intended in terms of their underlying theory-based constructs, purpose and target group (Merrell, 2009).

The early identification of risk factors and learning difficulties among students has been widely recommended (Walker, Colvin, & Ramsey, 1995; Loeber & Farrington, 2001; Heckman, 2013; Kautz, Heckman, Diris, Weel, & Borghans, 2014). In sum, the research emphasizes the importance of (a) identifying students' specific needs for support at an early stage, (b) informing instructional decisions for students at risk, and (c) continuously monitoring students' growth and whether and how they are responding to instructions. In Norway, however, converting early identification into practice is an ongoing challenge

(The Norwegian Ministry of Education, 2011, 2017). Based on this knowledge, three groups of students who face either (a) difficulties in social functioning, (b) difficulties in reading, or (c) difficulties in both social functioning and reading are the focus of the present thesis, which examines the quality of the educational assessments used in a Norwegian elementary school context to identify and guide instructional decisions for these difficulties.

1.1 Structure and Objectives

The present thesis consists of two parts: (a) an extended abstract and (b) three studies reported in three separate publications (Papers I - III). The overall objectives are to (a) contribute to the body of knowledge of educational assessments for both social functioning and reading proficiency, (b) uncover new insights regarding the quality of the educational assessment practices and approaches used in Norwegian elementary schools that, in turn, can impact students' learning and development in relation to social functioning and reading, and (c) inform decision-makers, practices and policies with evidence that can influence needed revisions in Norwegian schools' assessment practices.

Throughout the thesis, *social functioning* is used as an umbrella term for several social subskills that rely on a variety of social behaviors that influence social competence. In turn, these subskills affect students' opportunities to be motivated by and engaged in the learning situation and social activities with peers in school (Sutherland & Wehby, 2001; Beauchamp & Anderson, 2010). The umbrella term *reading proficiency* encompasses decoding skills; reading fluency, in terms of accuracy, automaticity and prosody; and reading comprehension of connected texts (Breznitz, 2006; García & Cain; 2014).

Figure 1 provides an overview of the three studies. Each study contributes unique aspects to the primary focus on the quality of assessments that aim to identify students' specific needs and inform instructional decisions, followed by the findings of each of the studies. The aim of Study 1 was twofold. First, we identified which assessments are used in Norwegian elementary schools (Grades 1 to 7) to screen, monitor progress and inform instructional decisions regarding students' social functioning and reading proficiency. Second, we evaluated the quality of the available information regarding the psychometric properties of the identified instruments. The evaluation aimed to derive precise information about the assessments' material descriptions and psychometric properties based on international standard procedures described in the European Federation of Psychology Associations' (EFPA's) review model (Evers, Hagemeister, & Hostmaelingen, 2013).

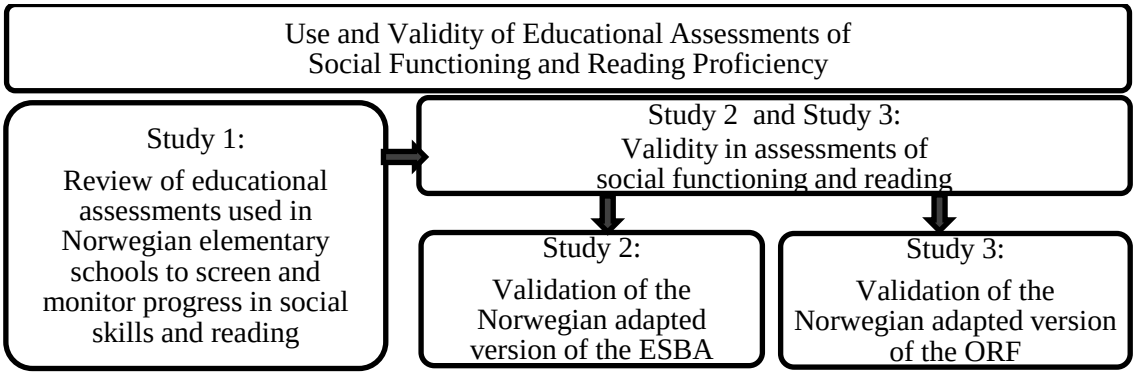


Figure 1. Overview of the thesis's main focus and the relationships among the three studies

In Study, 2 we examined the adapted Elementary Social Behavioral Assessment (ESBA: Pennefather & Smolkowski, 2015) and its relationship with the teacher's form of the Social Skills Rating Scale (SSRS-T: Gresham & Elliot, 1990) in Norwegian elementary Grades 1 to 6. The psychometric properties were analyzed. Additionally, we investigated whether there were differences in the teachers' ratings within a single school and between schools. The differences could explain the variability in students' scores that could imply cultural differences in the assessment's functioning.

In Study 3, the adapted Norwegian version of the curriculum-based measure Oral Reading Fluency (ORF: Good & Kaminski, 2002) was used to examine students' growth in oral reading fluency within each grade and across Grades 2–5. In addition to the longitudinal measurement's invariance and the relative stability of the measure, we examined the criterion validity and the relationship between the ORF measure and the National Tests of Reading Proficiency (NTRP).

2. Social Functioning and Reading Proficiency

In this chapter, I briefly present the overarching theoretical perspectives and evidences of the constructs of social functioning and reading proficiency. These relate to the role of individual and environmental factors regarding how students learn and develop in these areas, and why both are important. The constructs, however, are complex and are limited within this thesis to maintain the focus on the use and validity of educational assessments. Also, the comorbidity and prevalence of difficulties in one or both areas is outlined.

2.1 The Role of Individual and Environmental Factors

The extent to which students have equal opportunities to learn and develop their abilities and motivation to attend school is a complex issue. This complexity is due to several developmental factors and risk factors associated with learning difficulties. Studies have demonstrated that several aspects of students' individual capacity (i.e., genetic or hereditary, temperamental, neurological, psychological) and environment (i.e., family, home literacy, socio-economic status (SES), parenting, neighborhood, school, and society), in addition to the intensity and the duration of difficulties, may impact students' social and academic learning and development (Capaldi, DeGarmo, Patterson, & Forgatch, 2002; Heckman, 2011; Rutter 1989; van Bergen, van Zuijen, Bishop, & de Jong, 2017; Walker & Sprague, 1999). Teachers' perceptions of their relationships with students are also shown to predict school success (Hamre & Pianta, 2001; Jennings & DiPrete, 2010). Taken together, both individual and environmental factors play important roles in several isolated skills that impact how students learn and develop social competence and reading proficiency. The underlying concept is an understanding of how these factors build strong foundations for students' social and academic development to success and how the two developmental areas are inter-related (Heckman & Kautz, 2012; McEvoy & Welker, 2000; Organisation for Economic Co-operation and Development: OECD, 2015). These factors are important to consider when assessing students' social functioning and reading proficiency.

Several theories have provided knowledge of the importance of how the social environmental context contributes to a child's social and academic development. First, Bronfenbrenner's (1979, 1997) ecological system theory describes four inter-related systems (i.e., micro, meso, exo, and macro) that directly and indirectly impact how a child experiences and learns. Second, learning theories on cognitive development incorporate Vygotsky's (1978) two main principles ("more-knowledgeable others" and "the zone of

proximal development") and Bruner's (1996) "concept of scaffolding" to conceptualize the importance of significant others (i.e., parents, siblings, teachers, peers) for supporting students' success in terms of academic learning and social well-being. Third, Bandura's (1977, 1986) triadic reciprocally theory describes how individual factors (e.g., cognitive, affective, biological) and environmental factors influence each other and impact students' behavioral patterns and social, cognitive and emotional development. Finally, Patterson's (1982, 2016) social interaction learning (SIL) theory expands our understanding of why and how destructive social behaviors or problem behaviors in students develop as a consequence of dysfunctional social interaction and coercive relationships at the micro level (i.e., family, preschool/kindergarten, school). These theoretical perspectives have influenced the understanding of how and why cognitive and social learning and development are the outcomes of several learned subskills acquired through experiences, positive responses and support from the environment (i.e., parents, teachers, peers).

The overarching theory- and evidence-based knowledge derived from the presented studies contributes to an understanding of the main pathways to social functioning and reading proficiency, how specific difficulties develop, and inequities in abilities, engagement, and motivation to learn among students. Moreover, it underlines why students' social functioning and reading proficiency are important to assess as a foundation for success in school (Beswick, Sloat, & Willms, 2008; Rutter, 1983). Understanding both multiple protective factors and risk factors that underlie typical and atypical development of social functioning and reading is important for understanding how and why assessment information can lead effective instructional decisions that prevent the growth of learning difficulties in at-risk students. In addition, as explained in Chapters 4 and 5, well-defined, theory-based constructs are an important basis for the examinations of construct validity and statistical fit within structural equational modeling that are used in Studies 2 and 3 (see Papers II and III).

2.2 Pathways to Social Functioning

The pathways to appropriate social functioning reflect ongoing processes of developing social competence as a result of several learned social skills in interaction and communication with others in different contextual settings and environments (Bronfenbrenner, 1979, 1997; Vygotskij, 1978; Bruner, 1996; Bandura, 1977, 1986). Several studies have shown that children's social behavior patterns, emotions, feelings, attitudes, social adjustment and self-regulation impact their social functioning with age

(e.g., Schaffer, 1999). Patterns of social behavior established at home are transferred to other social contexts and influence how students succeed in school (Patterson, DeBaryshe, & Ramsey, 1989; Patterson, Reid, & Dishion, 1992). Consequently, behavioral responses and interactions with other students are functions of how children process the social information through which they establish relationships (Crick & Dodge, 1994).

Social adjustment includes how a child develops and masters self-regulation which is a critical underlying skill that contributes to success in the social and academic setting of school (Blair & Raver, 2015; Ogden & Hagen, 2013). That is, social adjustment reflects the child's abilities to regulate attention, emotions and behavior during interactions with others. Basically, self-regulated learning is an executive function that includes underlying cognitive factors such as impulse control, attention regulation, emotional control, flexibility, planning, problem-solving, and working memory. Hence, executive functioning is related to the acquisition of social skills that impact social functioning and academic achievement (Beauchamp & Anderson, 2010; Zelazo, & Müller, 2010).

2.2.1 Social skills - social behaviors - social competence

Given that the use of appropriate social skills is critical to a student's educational success, it is important to teach social skills that facilitate appropriate social functioning within the learning environment. The construct *social skills* is commonly defined as the skills that are required to develop social relationships, emotional and academic engagement, and school motivation (Beauchamp & Anderson, 2010; Cordier et al., 2015; Gresham, 2007). In addition, social skills are defined as specific, observable *social behaviors* that predict outcomes that in turn promote positive social relations with peers and school success (Elliott & Gresham, 1991; Gresham, 1986; Gresham, Elliott, Cook, Vance, & Kettler, 2010). *Social competence* refers to the competent use of the social skills that a person is expected to use in different environments (Gresham, 2002). In this view, social competence in a school context is the result of a student's achievement of fluency in several learning-related social skills.

Studies have demonstrated that students' social competence has a significant impact on the social, emotional and cognitive development and learning that drive their future life outcomes (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Gresham, 2007; Jones et al., 2015; OECD, 2015). In accordance with Spence (2003), the range of responses and behaviors (e.g., verbal, non-verbal, imitation, gestures, eye contact) that occur between individuals during social interactions are micro-level aspects of social skills that influence

the development of social functioning and are highly important to how we cope during interactions with others. Furthermore, in the development of appropriate social functioning, a student's ability to integrate micro-level skills into more complex social tasks at the macro-level is important. For example, asking for help, initiating conversations, inviting others to join in, and listening to others are responses and behaviors that impact the outcomes of social interactions and promote social functioning. In sum, several micro- and macro-level subskills influence the pathway to social functioning as a result of important emotional, cognitive and environment factors and, therefore, social competence (Spence, 2003, p. 84). Given that school is an important micro- and macro-level environment during a student's life-span, appropriate social functioning within this context is critical for student achievement and is highly important to monitor (Bronfenbrenner, 1979; Spence, 1995).

In a longitudinal study of Norwegian students in Grades 8 to 10, both social competence and problem behavior were found to be stable dimensions of a student's social functioning (Sørlie, Hagen, & Ogden, 2008). That is, students who scored high on measures of either social competence or problem behavior at the first time-point tended to score high on the same dimension two years later. Furthermore, initially low social competence scores predicted high scores for problem behavior two years later. These findings strengthen the importance of the early identification of social skills difficulties to promote social competence. However, initially high problem behavior scores were not found to relate significantly to low social competence scores two years later. This said, students who struggle with self-regulation and demonstrate problem behavior are not necessarily lacking in social competence per se, but they may need support that can help them to develop self-control and self-regulation skills.

Findings from a 19-year longitudinal study demonstrated a significant association between social competence ratings in kindergarten and positive and negative outcomes for education, employment, mental health, and crime (Jones et al., 2015). In addition, in a longitudinal study of 14- and 15-year-old in nine countries, OECD (2015) found that social and emotional skills (defined as a sense of responsibility, locus of control self-esteem, self-calming, respect and emotionally stability) improved the students' cognitive skills, although cognitive skills had limited impact on the development of social skills and emotional behavior. The study demonstrated that students can compensate for academic shortcomings through their use of learned social skills.

2.2.2 Learning-related social skills

In a study of students' risk for early academic difficulties, McClelland, Morrison, and Holmes (2000) included both *interpersonal skills* and *work-related social skills* in the concept of *learning-related social skills*. The researchers found that students who showed poor work-related skills at the beginning of kindergarten had low academic achievement at the end of second grade. In other words, early work-related social skills are important to how students cope during their transition to school and early academic achievement (McClelland, Morrison, & Holmes, 2000). Hence, the distinction between learning-related social skills and interpersonal skills is important for grasping the complex relationship between social competence and academic achievement. Moreover, appropriate social functioning is shown to impact students' motivation and engagement in academic learning and social activities with peers in school (Al-Hendawi, 2012; Sutherland & Wehby, 2001; Welsh, Parke, Widaman, & O'Neil, 2001).

In line with the abovementioned distinctions among learning-related social skills, the ESBA scale considers both work-related social skills and interpersonal social skills (Paper II). The items represent social behavior skills (see Paper II, Table 2) that are related to *academic engagement* and *peer social relationships* and that teachers find important for successful learning (Gresham & Elliot, 2008; Walker et al., 2015). Moreover, the content of the ESBA scale reflects the social skills that contribute to a student's ability to pay attention, comply, be self-confident, have self-control and impulse-control, communicate, and solve problems (e.g., Gresham & Elliot, 1990; Spence, 1995, 2003).

Taken together, these findings provide evidence that the social skills that drive individual success and social progress in students' learning and development in school are multi-dimensional and include both cognitive and social-emotional elements (e.g., Heckman, 2011; Norwegian Ministry of Education and Research, 2015; OECD, 2015). These are important factors to consider when assessing students' development and learning to promote their pathway to appropriate social functioning.

2.2.3 Difficulties in social functioning

Many students struggle with social functioning when they enter school (Cummings, Kaminski, & Merrell, 2008). The students' difficulties do not only impact a student's well-being and academic learning in school; they also contribute to competing problem behavior and interfering feelings, which include internalizing and externalizing behaviors (Gresham

& Elliot, 1990). The accurate identification of the specific difficulties in social functioning that students are facing is a key to preventing social and academic failure.

When assessing students' social functioning in school to identify those at-risk and guide instructional decisions that lead to improved social functioning, it is important to distinguish between social skill difficulties, motivational difficulties and fluency difficulties (Gresham, 2002, p. 408). *Social skill difficulties* refer to acquisition difficulties: "can't do". That is, the student is lacking the social skills that are needed to cope with challenges of school. Such students misbehave because they try but do not know what is needed to execute a social skill, even under optimal conditions. *Motivational difficulties* refer to performance difficulties - "won't do". The "won't do" student may have sufficient social skills but fails to perform these as expected in particular situations. Hence, they may engage in undesired behavior at school because they "won't do" academic work as expected of them. *Fluency difficulties* refer to skills a student may know and want to perform but may execute incorrectly and/or without fluency due to lack of instruction and/or practice. This distinction is important because the information derived from teacher rating scales, such as the ESBA (Paper II), requires different instructional approaches depending on the identified difficulties. For instance, "won't do" students may need interventions that promote their academic attitudes, engagement and motivation rather than social skills instruction, while "can't do" students may need social skills and/ or reading-related skills interventions.

2.3 Pathways to Reading Proficiency

Reading is a complex activity in which the student acquires several underlying skills on the path to becoming a proficient reader (Hoover & Gough, 1990). In accordance with Ehri, Barron and Feldman (1978) and Ehri (1997, 2005), learning to read depends on several simultaneous and essential processes that help students acquire reading-related skills. These skills, which are briefly outlined in the following sections, have different functions although they are tightly interwoven. They are involved in the verbal and visual cognitive processes that lead to the ultimate goal of understanding and interpreting the content of connected texts. When assessing and monitoring students' growth in reading, it is important to understand how and why students are struggling in these processes and how each skill develops and contributes to the path to reading proficiency (Rack, Hulme, & Snowling, 1993).

Reading abilities in 1st Grade are shown to be a strong predictor of reading proficiency ten years later (Cunningham & Stanovich, 1997). These findings are in line with those of Duncan and colleagues (2008), whose meta-analysis of longitudinal studies found that students' reading skills at school entry were consistently associated with higher levels of academic performance in later grades. Moreover, the relationship between students' attitudes toward reading and their reading proficiency has been shown to be stronger for elementary school students than for middle school students (Petscher, 2010).

2.3.1 Reading comprehension and the Simple View of Reading

Reading comprehension is the basis for overall academic achievements, which in turn may impact how a student will cope in school and as an adult in society (García-Madruga, Vila, Gómez-Veiga, Duque, & Elosúa, 2014; Hoover & Gough, 1990). In other words, because most school subjects (e.g., math, literacy/language, geography, biology, history, art, social science) require reading skills, learning to read is the gateway to reading to learn in these subjects (O'Reilly & Sabatini, 2013; Snow, Burns, & Griffin, 1998). Becoming a comprehensive reader requires the use and fusion of two basic processes when learning to read: First, students acquire the skills to decode printed letters and blend those letters into words accurately and fluently. Next, students acquire skills to understand the meaning of the decoded words in connected text (Ehri, 2005). These processes are widely known as the Simple View of Reading.

In accordance to the theoretical framework of the Simple View of Reading, reading comprehension is the product of word decoding and listening comprehension (Gough & Tunmer, 1986; Hoover & Gough, 1990). Although there are studies that challenge the Simple View of Reading model (e.g., Kershaw & Schatschneider, 2012; Tunmer & Chapman, 2012), strong evidence supports it (see e.g., García & Cain, 2014). For instance, in a longitudinal study from preschool to 4th grade, Storch and Whitehurst (2002) demonstrated (a) a strong relationship between code-related precursors and oral language in preschool, (b) a high degree of continuity of decoding and oral language over time, and (c) how the level of letter knowledge and phonological awareness in kindergarten influences the level of reading ability during the early grades. From Grades 1 to 6, most students begin the process by learning how to read and then learn how to use reading skills to acquire more knowledge. In a Norwegian longitudinal study that began when the participants were 7.5 years old and followed them across 5 school years, the findings were in line with the Simple View of Reading model: listening comprehension and word

decoding explained 96% of the variation in early reading comprehension (Lervåg, Hulme, & Melby-Lervåg, 2017).

2.3.2 Decoding and fluency

As posited in the Simple View of Reading model, *decoding* skills are critical for reading comprehension. That is, decoding - the process of translating printed letters into words by blending sounds to pronounce the letters and words - is the foundation of all other reading skills (Paper III). This assumes that the student knows how single letters typically symbolize sounds that blend to form words. When students have developed skills in blending letters to form words, they can decode the words in a connected text correctly, with *accuracy* and *automaticity,* and at a *fluent speed,* which makes reading efficient and understandable (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Becoming a fluent reader requires frequent practice of both decoding-related and language-related skills (see e.g., Biancarosa & Shanley, 2016; Rose, 2006; Smolkowski, Cummings, & Strycker, 2016).

The foundation for reading decoding and fluency is thoroughly discussed in the introduction section (pp. 177-180) of Paper I and is not reiterated here. However, the main issues are the understanding of the importance of a student's pathway to reading proficiency as it proceeds through learning to read and reading to learn. Learning to decode with fluency requires sufficient practice to build the fluency with which students retrieve sight words from memory (Ehri & Wilce, 1985). It has been shown that beginning readers (an average age of 7 years and 7 months) can keep sight words in memory after reading them as few as four times (Reitsma, 1983). This means that students gradually remember words that they already know by sight because they have heard and seen them through previous experiences. This process results in reading fluency because it replaces the need to decode each word slowly. Through the student's memories of repeatedly decoding the same word, he or she will master word fluency, and most words will become sight words that the student can read automatically. When the mind is processing words automatically (fluently) despite an intention to ignore them, the student combines the sounds at a glance as a whole unit without pausing between the sounds (LaBerge & Samuels, 1974). It is critical that students learn to accurately decode words before building fluency and automaticity. This is a pivotal time in a student's reading development.

Typically, students quickly acquire *automatization* of specific correspondences between patterns of letters in spelling and pronunciation as they start learning to read (see, e.g., Stanovich, 1980). Automaticity in decoding becomes stable after the third grade. At

that point, practice in decoding builds accuracy and speed, which become the major factors in *reading fluency* (Perfetti & Hogaboam, 1975). Additionally, fluency in reading connected text has been shown to significantly predict reading comprehension in 7th and 10th graders after controlling for decoding, listening comprehension and language comprehension (Kershaw & Schatschneider, 2012). As students become more fluent, decoding becomes more automatic; less time and effort are required for word recognition, and the focus of instructional time shifts to reading comprehension (Johnston & Kirby 2006). Thus, automatization and accuracy are pivotal skills for fluent and efficient reading. Moreover, reading fluency serves as a bridge from decoding to reading comprehension (Pikulski & Chard, 2005).

Accuracy and fluency are important developmental indicators of reading comprehension, and each plays a different role in the path to reading proficiency (Biancarosa & Shanley, 2016; Bretznitz, 2006). However, different languages have different levels of orthographic transparency. Studies have shown that even though the predictors for learning to read are the same across orthographies, the time it takes for children to learn to read varies in different languages (Caravolas et al., 2012; Caravolas, Lervåg, Defior, Seidlová Málková, & Hulme, 2013). In contrast to languages with a transparent orthography (i.e., Spanish, Finnish), languages with a non-transparent orthography (i.e., English, French) or a semi-transparent orthography (e.g., Norwegian) cannot be read accurately by translating letters into sounds (for more details, see Paper III, p. 179).

2.3.3 Difficulties in reading

Because reading is a continuous variable that is normally distributed in the population, a number of students struggle in learning to read (Hulme & Snowling, 2011; OECD, 2013). However, the definitions and degrees of difficulties may depend on the cut-off criteria's sensitivity and specificity of the assessments used to identify students' difficulties in reading. For instance, as seen in Chapter 2.5, the prevalence of dyslexia may vary from 5% to 17%. Moreover, schools may have different base rates for the expected benchmarks and levels of students' achievements (Cummings & Smolkowski, 2015; Smolkowski, Cummings, & Strycker, 2016).

Children with oral language difficulties or specific language impairment are at risk for later reading difficulties resulting from poor reading comprehension (e.g., Bishop & Adams, 1990). The Simple View of Reading model provides evidence that students'

reading development varies within its two dimensions (listening comprehension and decoding). That is, struggling readers can have poor decoding skills and good listening comprehension, good decoding skills and poor listening comprehension, or both poor decoding skills and poor listening comprehension (see e.g., Hogan, Adlof, & Alonzo, 2014; Hulme & Snowling, 2011; Woolley, 2011). The early identification of a student's specific difficulties paired with immediate targeted instruction is crucial for preventing students from failing in school and society as they age. However, different approaches are needed to meet the broad types of reading difficulties (Bishop & Snowling, 2004).

Several studies have demonstrated that the gap in social and academic achievements between typically developing and at-risk children increases with age and is influenced by genetics, family SES, and environmental risk factors (Hart & Risley, 1995, 2003; Olson, Keenan, Byrne, & Samuelsson, 2014; van Bergen et al., 2017). Genetics are shown to account for most of the variance in students' reading abilities in early grades (see e.g., Olson et al., 2014; van Bergen et al., 2017). Despite the strong influence of genetics, the importance of environmental factors for supporting struggling readers while decreasing the social-academic gap is highlighted. For instance, identifying students at-risk for difficulties in reading, encouraging the environment to model positive reading attitudes and enhance students' motivation for reading (e.g., access to books of interest at home, social interaction with books at school) might be a contribution for support.

Along with genetics, school-level SES, which is an environmental factor, has been shown to be a factor in reading comprehension (Hart, Soden, Johnson, Schatschneider, & Taylor, 2013). Soden and colleagues (2015) reported a longitudinally stable change in genetic and environmental influences on reading comprehension across Grades 1 to 6. These findings indicate that although reading is a learned skill, the environment remains important to reading development. Individual differences in reading comprehension are influenced by a core of genetic stability that persists throughout the developmental course of reading.

When assessing students' growth in reading, it is important to use measures that distinguish among the types of difficulties students experience to guide effective intervention. For instance, curriculum-based screening measures such as the ORF have demonstrated both significant correlations with other high-stakes reading tests and informal inventories for identifying students at risk for reading difficulties and greater diagnostic accuracy than other inventories for correctly identifying the difficulties (see e.g., Paper III; Parker et al., 2015; Yeo, 2010).

2.4 Comorbidity in Difficulties of Social Functioning and Reading

Decades of studies have examined students' comorbidity in difficulties in the areas of social functioning and reading proficiency and how skills in one area impact or co-occur with skills in the other area (see e.g., Algozzine, Wang, & Violette, 2011; Hinshaw, 1992; Lane, Barton-Arwood, Nelson, & Wehby, 2008; McEvoy & Welker, 2000; Russell, Ryder, Norwich, & Ford, 2015). As described in Paper I, reading difficulties may trigger difficulties in other academic and social functioning skills (Terras et al., 2009). For instance, students may be frustrated and anxious in the classroom because of their specific reading difficulties and fail to regulate these negative feelings that in turn may lead to behavior problems (McIntosh, Sadler, & Brown, 2012). Moreover, Rhoades, Warren, Domitrovich, and Greenberg (2011) found that students with higher social competence are more attentive toward learning and thus gain more from academic instruction. It is also likely that students who get better along with teachers receive more positive attention and thus are both more engaged and motivated to learn and then gain academically (Denham & Brown, 2010).

In accordance with McIntosh, Horner, Chard, Dickey, and Braun (2008), some combinations of difficulties concerning social functioning and reading can be explained in several ways. That is, a student can either show patterns of (a) social behavior difficulties (e.g., pushes another away from favorite toys on playground) and appropriate academic proficiency, (b) academic difficulties (e.g., struggles in reading-related skills) while getting along with peers, (c) inter-related social behavior difficulties and academic difficulties (e.g., engages in conflicts with peers to escape a reading task and is reading below grade level), or (d) unrelated social behavior and academic difficulties (e.g., engages in bullying behavior to get attention from peers and has difficulty in reading). Facing social and/ or academic difficulties over time have negative impacts on students' motivation and engagement to attend and learn in school (Al-Hendawi, 2012; Durlak, et al., 2011; OECD, 2015). In turn, students who fall behind academically are more likely to find academic work aversive and may find escape-maintained problem behaviors reinforcing (OECD, 2015; McIntosh, Horner, Chard, Dickey, & Braun, 2008).

2.5 Prevalence of Difficulties Related to Social Functioning and Reading

Depending on the cut-off point between typical and atypical development, the prevalence of students who are facing difficulties in a specific area may vary considerably. Additionally, the distribution of specific difficulties will vary depending on, for instance,

the specific developmental area, the categorical definitions used, the sample size, and the cohort of students measured.

It has been reported that 15% to 25% of the students in Norwegian schools (Grades 1 to 10) face academic (math, reading) and socio-emotional (anxiety, conduct disorders, depression) difficulties that impact the benefits of attending school (The Norwegian Ministry of Education, 2009, 2017). Moreover, 20% of students need more intensive support than their peers to succeed socially and/or academically (Haug, 2014; The Norwegian Ministry of Education, 2017). In fact, this number is twice the number of students who receive special education services. The total number of students receiving special education services in Grades 1 to 10 has decreased slightly, from 8.3% (2013-14 school year) to 7.8% (2016-17 school year), and 70% are male. In 2013, twice as many students in Grades 8 to 10 (11.2%) than in Grades 1 to 4 (5.6%) received special education services. These data show a systematic increase in the number of students who struggle as they age; the percentage of students with difficulties is twice as high in Grades 8 to 10 (10.2%) as in Grades 1 to 4 (5.1%).

Students requiring special education services typically present difficulties in both social functioning and reading proficiency (The Norwegian Ministry of Education, 2011, 2017). Although a relatively low percentage of students qualifies for special education services, teachers in Grades 2-5 have reported concerns about semantic, reading, and social functioning skills in far more students (16%) than are receiving individualized support through special education services (6.7%) (Arnesen, Meek-Hansen, Ottem, & Frost, 2013; Statistics Norway, 2016).

Behavior presents itself in two basic forms: internal and external. Approximately 10% to15% of Norwegian students are identified as demonstrating internalizing behavior that takes the form of anxiety, depression, social withdrawal, and/or negative self-attribution (Bru, 2011). The prevalence of diagnosable external behavioral difficulties is estimated at 1.7% for students diagnosed with severe conduct disorder and 1.8% for those diagnosed with oppositional defiant disorder, with a predominance of boys for both disorders (Skogen & Torvik, 2013). The prevalence of dyslexia varies from 5% to 17% depending on the sample selection criteria and the definition used (Morken & Helland, 2013).

Reflecting on these data, a few patterns become evident: when we ignore the early onset of academic and social behavior problems, students fall further behind. When students fall behind, they experience an achievement gap that fosters disruptive and

challenging behavior, requiring more intensive interventions, and dropping out is a possible outcome. The research is clear; early identification of problems followed by effective instruction leads to positive student outcomes.

3. Educational Assessment Practices and Changes

This chapter addresses some perspectives on educational assessment foundations, approaches and practices which aim to (a) monitor students' social functioning and reading proficiency development, (b) identify specific challenges for students at risk, and (c) guide instructional decisions and evaluate progress. Furthermore, due to probable needs for changes in the assessment practices, I present a Theory of Change (ToC) model. The ToC model is presented to obtain an understanding of why and how improvements can be conceptualized in future practice based on the evaluation of the quality of the assessments used.

3.1 Assessment Foundation

Assessments used to identify specific difficulties and guide effective instruction need to be accompanied by theoretical perspectives that explain how difficulties in social functioning and reading proficiency grow (Merrell, 2009). When assessments are based on vague theories and definitions of constructs, it may be difficult to interpret their findings in ways that support the students' learning and development. For instance, defining and measuring the construct of *social functioning* consistently in practices is shown to be challenging (Cordier, et al., 2015; Cummings, Kaminski, & Merrell, 2008; OECD, 2015; Sutherland, McLeod, Conroy, & Cox, 2013). Inconsistencies in definitions may impact how struggling students' difficulties are understood and interpreted in different school contexts. Moreover, as earlier mentioned (e.g., Chapter 2.2.3, 2.3.3, 2.5), the understanding of typical and atypical development of social functioning and reading may differ widely depending on the students' age and language, contextual expectations, and the types of measurements used (Connors-Tadros, 2014; Cordier et al., 2015). Therefore, high-quality assessments that have demonstrated good psychometric properties across different school contexts and cultures are a critical foundation to identify students' specific difficulties as a base to support their learning and development.

3.2 Assessment Approaches

Ideally, screening and monitoring students' social functioning and reading proficiency should support teachers' view of all students' gains and progress as a result of universal and individual instruction. Such assessments are commonly referred to as curriculum-based measurements (CBM) and require various purpose-oriented procedures for conducting, scoring, and analyzing the information they provide. *Screening* refers to a

systematic, universal process of assessing all students in the entire elementary school population for the early identification of individual differences and possible difficulties with social and/or reading skills. *Progress monitoring* refers to strategies that are used to judge students' development and evaluate their response to social skills and/or reading instruction. In accordance with Fuchs and Fuchs (2007) and Gresham (2007), screening and monitoring students' social and academical responses to either universal or specific instructions (RtI: Response to Instruction) should be a continuum of procedures within and across school years. This process provides a systematic method for identifying students who are struggling in social functioning and/or reading by informing instructional decisions and monitoring their progress throughout the school year. Within this thesis, as seen in Paper II and Paper III, the ESBA and the ORF serve as examples of assessments that combine screening, instructional decisions, and progress monitoring.

Assessing students within an educational context commonly require *formal* or *informal* approaches, use of *direct* or *indirect* methods, and either *summative* or *formative* assessments. Formal approaches follow predetermined procedures defined in manuals or administration protocols, while informal approaches are easy-to-use, teacher-constructed assessments commonly designed to judge student progress while informing instructional decisions (Thorndike & Thorndike-Christ, 2014). Direct methods are self-reports, interviews or tests (e.g., reading tests) in which the student responds directly teacher-administered or computer-based tasks. Indirect methods are typically teacher ratings of the students' skills or achievements based on their perceptions and judgements (Crowe, Beauchamp, Catroppa, & Anderson, 2011). These provide inexpensive and efficient methods, but have drawbacks, including a reliance on the rater's objectiveness and error variance that includes changes in behavior over time and settings (Martin, Hooper, Snow, & Knoff, 1986). Notably, assessing students' reading skills with, for example the ORF measure, which is a direct assessment method, is very different from assessing social skills with, for example, the ESBA teacher-rating scale, which is an indirect method and more prone to measurement error.

Summative assessments provide static information regarding students' learning at a given point in time. These might include national and other tests at the end of instruction or at the end of the school year to evaluate whether the student is above or below a cut-off score. Formative assessments provide dynamic information about a student's progress during specific time periods (e.g., 6 weeks to a full school year) to guide instructional decisions for further learning (Black & William, 2010; Sattler, 1992). Most assessments

used in schools are summative and static rather than formative and dynamic (Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013; Parisi, Ihlo, & Glover, 2014). The ESBA and the ORF measure are both examples of formative measures in which the students are assessed within the school context on the skills they were taught.

3.3 Assessment Practices

Norwegian students are administered compulsory national tests of educational subjects throughout their school years, and they are expected to be assessed by their teachers on social well-being and academic outcomes (Tveit, 2014; Seland & Hovdhaugen, 2017; OECD, 2015; The Norwegian Ministry of Education, 2017). Although there are traditions for assessing students' achievements, it is unclear whether this practice is utilized to improve students' learning and development, how and when those assessments are being conducted, the extent to which the students benefit from them, and/or the extent to which teachers use the assessment results. These challenges in the practice of educational assessments have been publicly debated for years (OECD, 2007; The Norwegian Ministry of Education, 2011). One challenge relates to the selection of assessment tools and procedures in terms of resource allocation in relation to the student outcomes of being assessed (e.g., OECD, 2007, 2015; Monsen, 2013). Another challenge questions the quality of assessments and the need to improve assessment procedures and practices (Paper I).

Teachers use their experiences and knowledge to understand and analyze the information derived from observations of their students. However, when this knowledge is based on common sense rather than knowledge of the measured construct, inaccurate information may result (Thorndike & Thorndike-Christ, 2014). Common sense or experienced-based knowledge is based on information we trust as useful; consequently, the chosen information will vary depending on the assessor's (teacher's) experiences, arguments and goals for using the assessment. Moreover, in accordance with Popper (1975), while teacher-based observations of students' achievements are biased perceptions and do not have legitimacy, they do explain the observations. Teachers' experience-based knowledge and subjective beliefs may be relevant and true for a specific context. Consequently, when assessment data are not derived from valid assessments or standardized measures, the findings may be inconsistent and difficult to use.

Although most teachers may have extensive experience assessing students' proficiency, inaccurate or erroneous inferences can prevent students from receiving needed support. Studies have shown that students' difficulties are more likely to be accurately

identified with valid assessments than with informal and teacher-constructed ratings (Antoniazzi, Snow, & Dickson-Swift, 2010). However, teacher-constructed assessments are commonly used (see e.g., Paper I), and which informal and subjective classroom observations are based on teacher's daily observations.

The educational assessments that are the focus of this thesis serve as indicators of social functioning and reading proficiency for identifying students' difficulties, informing resource allocation and, ideally, instructional decisions (Crone et al., 2016). For instance, both the ESBA and ORF measures, described in Paper II and Paper III, respectively, may serve as valuable sources of assessment information for teachers to promote students' achievements. By paying attention to a student's specific social behavior in the educational context of a classroom, a teacher may obtain a better understanding of how the student's social functioning impacts his or her social and academic learning and development. Additionally, by listening to a student's oral reading of connected texts, a teacher can gain valuable insights into the student's decoding, accuracy, fluency, and prosody. Basically, the role of assessments is to provide valid information and outcomes regarding the assessed construct and to determine how interventions should proceed. This is an important aspect of the potential need for changes in the assessment practices of Norwegian schools addressed in this thesis.

In summary, educational assessments may be useful in practice when the relevance or social validity of the information they provide can be used to (a) make instructional decisions that fit each student's needs, and (b) measure what a student has actually learned. Since the use of assessments has implications for decision making and practice, areas toward which both students and teachers direct considerable resources, it is fundamental to evaluate how well those assessments are working and when, where and why they can be applied. It is also important to examine the quality and validity of the assessments to ensure that they measure what they are intended to measure and that they warrant the time spent implementing them. The findings from such examinations may lead to changes in educational assessment practices. Therefore, a ToC evaluation model that can guide these changes is described in the next section.

3.4 Theory of Change Model for Improving Assessment Practice

Explicit theories are helpful for identifying what to evaluate and structuring the processes of possible changes. The ToC evaluation model was founded to guide research in social science and politics to improve evidence for practice recommendations (Patton, 2002;

Weiss, 1972, 1995). Given that the overall objectives of this thesis pertain to the quality of educational assessments in Norwegian elementary schools, the ToC is used to synthesize the potential contributions of the three presented studies to assessment practices, policies and principles. Based on the findings of a gap between existing and probable future assessment practices, these contributions may lead to possible needs for systemic changes at different levels of educational systems (i.e., classroom teachers, special education, school leadership and decision-makers) in terms of assessment practices and school competence (Paper I). Moreover, these findings may also expand the pool of high-quality assessments that can be used in schools to improve student achievement and well-being. This is discussed in Paper I and elaborated through the Norwegian adaptations and validity studies of the ESBA (Paper II) and the ORF measure (Paper III).

The constructs of social functioning and reading proficiency and the foundations of the assessments are key features upon which the evaluation of the quality of an assessment rests. In turn, the quality evaluation included in the present studies may guide and improve screening, instruction and progress monitoring practices and will likely yield useful insights to prevent students from struggling in the areas of social functioning and reading proficiency. Improving assessment practices in schools requires information that may be derived answers to questions dealing with social validity, such as: Should assessments be changed? Why should they be changed? What should be changed? How should they be changed? How will we know they are changed? (Gresham, 2007). It is important not only to judge assessment quality *per se* but to determine how and why the assessments guide instructions that result in changes in student achievement. Additionally, a student's general maturity and development over time, along with other factors that may influence those changes, must be taken into consideration.

Basically, the ToC evaluation model provides a process in which several steps of related short-term and long-term outcomes create a pathway for changes based on the assumptions of a need for improvements (Weiss, 1995). That is, ToC helps to understand how changes can be measured over time to promote students' achievements. The evaluation model is in line with the overarching evaluation theory by Shadish, Cook, and Leviton (1991). They claim that any changes in practice should be based on an overall evaluation of the constructs' (a) quality of structure and functions, (b) evidence, (c) worth, (d) usability, and (e) methods and procedurals. The EFPA review model (Appendix A) presented in Paper I is an example of how educational assessments can be evaluated in line with this theory.

Ideally, the ToC model structures the understanding of how and why the assessment practices can be changed. Thus, a ToC evaluation involves analyzing and describing the steps taken in this process. The ToC evaluation model developed within this thesis, shown in Figure 2, illustrates how four steps in this process can be used to close the identified gap between the existing and possible future assessment practices in Norwegian elementary schools. It is, however, beyond the scope of this thesis to fully describe this process because the use and validity of educational assessments of social functioning and reading is the main focus.

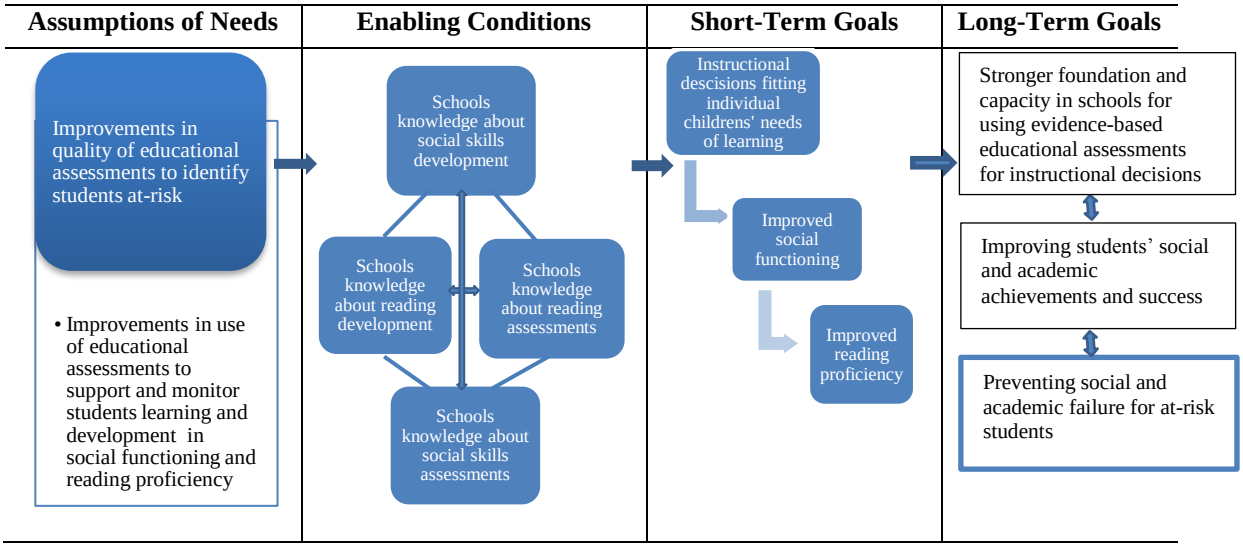


Figure 2. Theory of Change Evaluation Model (modified from Weiss, 1995)

The first column in Figure 2 illustrates, as earlier mentioned, the assumptions of needs that are based on the outcomes of Study 1. Improving the quality of assessments that aim to assess students' benefits from instruction and identify those at-risk, is a highly important outcome (Paper I). The enabling conditions, shown in the second column in Figure 2, are those that may improve a school's knowledge regarding both students' social functioning and reading development and which assessments are appropriate for this purpose. A well-described ToC model aims to inform practice decisions that can develop, improve, and change assessment practices by improving the assessments' material quality, psychometric properties, content, and use (Papers II and III). The figure's third column, illustrates that the short-term goals of instructional decisions that promote students' social functioning and reading are the outcomes of the enabled conditions shown in the previous column. Finally, as shown in the figure's fourth column, the long-term goals are to strengthen the foundation of evidence-based educational assessments and build the

school's capacity to use them to promote student success and prevent at-risk students from failing.

In summary, the ToC model applied in this thesis both addresses knowledge regarding the foundations of valid educational assessments to improve their performance and quality and provides an understanding of how and why changes in practices are needed (Shadish et al. 1991; Stein & Valters, 2012; Weiss, 1995, 1997). From this perspective, ToC emphasizes the meaning of evidence-based practice that is grounded in knowledge obtained from high-quality research. Research quality in turn is based on frameworks of validity, as described in Chapter 4. The assessment instruments are operationalized according to the conceptualized preconditions derived from the theory and the projected needs for change in the delivery of educational assessment practices. The ESBA and the ORF measures demonstrate qualities (Papers II and III) that might decrease the gap and may be useful for screening and monitoring student responses to instruction in social functioning and reading proficiency.

4. Foundation of Assessment Validity

In this chapter I outline validity, the foundation upon which the evidence for the quality of educational assessments is based. The foundation of assessment describes the elements that are vital when developing high-quality assessments and explains the importance of evaluating assessment quality. This chapter presents theoretical perspectives on validity, which focus on test theory and psychometric properties. Both classical and modern test theory provides criteria for high-quality assessments and psychometric properties in terms of reliability, validity, and norms (Allen & Yen, 2001). Mainly, these theories provide several procedures for analyzing variables and examining accuracy, sensitivity, specificity, item responses, factor confirmation, and equality (Crocker & Algina, 2008). These theories were used to frame the thesis's validation studies and include several principles and criteria for evaluating the quality of the assessments (See the EFPA form, Appendix A and descriptions Paper I). As a result, the principles and criteria used to ensure quality indicate the empirical questions to be examined and statistically analyzed. For the purpose of the studies presented in this thesis, the foundations of assessment validity lean on the works of Cronbach and Meehl (1955), Cronbach (1971), Messick (1993, 1995), and Kane (2006, 2013). An overview of several aspects of reliability and validity, including facets of unified validity (Messick, 1993, 1995), is provided in this chapter. Then, a domain model is used as a framework for understanding how different levels of variables relate in the assessment validation process (Benson & Hagtvet, 1996). Finally, norms and predictive accuracy is briefly explained.

4.1 Evidence-based assessments

"Evidence-based" refers to the strength of the empirical findings of studies and their degree of validity (O'Reilly, Weeks, Sabatini, Halderman, & Steinberg, 2014; Sabatini, O'Reilly, & Deane, 2013). However, Cartwright (2011) provides a critical view of the term *evidence-based* as it relates to assessments. She argues that data can only have evidence for a population in the context in which the data are collected; the data from one population cannot predict the results for populations in other contexts. Given this argument, assessments can only provide evidence for those students who are measured in the exact same environmental context and only when the fidelity of the assessment procedures is assured. Because similarity in contexts is a significant factor for an evidence-based assessment, Cartwright questions the validity and any generalization of evidence across different contexts. If Cartwright's assertions are correct, multiple assessments are needed

because one assessment does not fit all contexts. This defines a core problem that this thesis tackles. In the three presented studies we clarify the importance of having valid assessments that accurately measure what they are intended to measure and contribute to underlying need of such assessments. The focus is on evidence-based assessments that aim for equality for all students and assess their needs and progress early and accurately. For instance, the adaptations of the ESBA scale and the ORF measure to a Norwegian culture and school context have taken these concerns into consideration (Paper II and Paper III).

4.2 Reliability

Reliability is essential for determining the trustworthiness of an assessment and serves as an important precondition for validity. In other words, an assessment's scores are not valid unless it is reliable. Reliability refers to the degree to which the assessment produces stable and consistent results. In addition, the reliability of an assessment includes the accuracy of the assessment measure, what it measures, and how precise the resulting scores are (Thorndike & Thorndike-Christ, 2014). Thus, reliability provides a measure of the extent to which the assessment produces random measurement errors and is expressed either as a *standard error of measurement* or as a *reliability coefficient*.

To assess the extent to which an assessment is reliable, several areas of reliability are evaluated. These are *inter-rater reliability* (i.e., the degree to which the results of different assessors agree for the same data/observation), *internal consistency reliability* (i.e., the degree to which different test items that probe the same construct produce similar results), and *test-retest reliability* (i.e., the degree to which the items are consistent over time when tested on the same population).

4.3 Validity and Validation

Validity is a crucial psychometric concept. It refers to the degree to which test scores provide information that is relevant to the inferences drawn from them (Thorndike & Thorndike-Christ, 2014, p. 76). In accordance with Messick (1993, p.1), validity is an ongoing evaluation of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* derived from test scores or other modes of assessment. Therefore, an assessment "does not have validity in an absolute sense" (Thorndike & Thorndike-Christ, 2014, p. 111), but the outcomes derived from analyses of the assessment process can be *valid* (true) or not (Messick, 1993). Through the validation process, evidence is collected to support the inferences and use of the scores derived from an assessment (Cronbach & Meehl, 1955, Cronbach, 1971). This is

in line with Benson & Hagtvet (1996), who state that validation is a matter of degree and not an all-or-nothing property. Moreover, Thorndike and Thorndike-Christ (2014) claim that the inferences drawn from the assessment outcomes are both *interpretive* inferences (i.e., how the score's mean is interpreted) and *action* inferences (i.e., how appropriate and useful the scores or outcomes are for the actions of instructional decisions). Essentially, validity is the property of the scores produced by the assessment rather than of the assessment itself (Kane, 2006).

The results produced by assessment constructs that are valid and reliable provide the evidence needed to support the use of the assessment. Whether the inferences derived from an assessment's scores and interpretations of its outcomes are true depends on how comparable the inferences are across different groups. Validation is an empirical evaluation of the meaning or interpretation of assessment scores and their consequences in terms of utility across people and contexts (Messick, 1995). This said, validity is a non-static property, while validation is an ongoing process. Therefore, assessments that are developed and found valid in one context are not necessarily valid in another context due to possible adaptations to fit the context (Study 2 and Study 3). The assessments reported in this thesis meet validity standards concerning *construct validity*, *content validity,* and *criterion-related validity.* Notably, the evaluation of the validity of the assessments included in the review study (Study 1) is based on the validity reported in documentation obtained through the literature search.

4.3.1 Construct validity

Construct validity refers to whether the items measure the theoretical construct for which they are designed. The constructs of the ESBA and the ORF assessments and their score interpretations reported in Papers II and III are based on theories of how students learn and develop fluency in social functioning and reading (see Chapter 2). The validation of these constructs aims to confirm that they reflect the intended theoretical constructs of the assessments (Cronbach & Meehl, 1955). Hence, the interpretation of the test scores should have evidence that is meaningful in practice and trustworthiness. Construct validity was examined with exploratory factor analysis (EFA) at pre-test and confirmatory analysis (CFA) at post-test for the ESBA measure (Study 2), and with a longitudinal design within a structural equation model for the ORF measure (Study 3).

4.3.2 Content validity

Content validity refers to the face validity and logical validity of the assessment. Regarding the ESBA and the ORF assessments, content validity is whether the items' representativeness and relevance provide an accurate assessment and covers the broad range of variation within students' social functioning and reading proficiency. Content validity is typically based on the qualitative judgement of an assessment by experts (i.e., teachers) of an assessment and/or an interpretation of how well scores or performances in the domain of interest (e.g., social functioning and reading proficiency) serve as an overall estimate and represent the content of the items within a larger domain. In accordance with Messick (1993), content validity alone cannot be used to qualify validity. Additionally, Kane (2006) cautions that evidence based on subjective judgments by test developers tends to confirm their proposed score inferences.

4.3.3 Criterion-related validity

Criterion-related validity refers to predictive and concurrent validity. That is whether the measurements (e.g., the ESBA and the ORF) correlate to other relevant valid assessments used for the same purpose to predict future or current performance. In Study 2, predictive validity was examined by correlating the ESBA scores at pre-test with the SSRS-T scores at post-test. Furthermore, the concurrent validity of the ESBA scores was examined by correlating them with the SSRS-T scores from the same assessment time-point. The predictive validity was examined in Study 3 by correlating the ORF scores at the first assessment time-point with the Norwegian national tests of reading proficiency (NTRP) at the third time-point.

Criterion-related validity also focuses on how norms and cut-off points (specificity and sensitivity) are developed and how the norming sample represents the population that the assessment was designed for in terms of such characteristics as age, socio-economic background, and gender. In terms of predictive accuracy, norms and cut-off scores are essential elements of educational assessments. They are commonly used to determine a student's scores in relation to a population of other students. Norms consider individual similarities and differences in such characteristics as socio-economic, socio-cultural, gender, and language background. The norms are examined to determine whether they are based on a representative sample of target students.

Multiple procedures (specificity, sensitivity, positive and negative predictive, prevalence, and accuracy) are typically used to increase an assessment's predictive

accuracy of scores for students at-risk of failure. Such procedures include analyses of methods to minimize the probability of misclassification and to establish observed scale cut-off scores to make appropriate inferences. To evaluate an assessment's screening ability and identify the "right" students, it is important to determine the test's sensitivity for yielding "true positives" and the specificity for yielding "true negatives". Thus, it is important to establish the cut-off point that best determines whether a student is considered to have difficulties that may require further attention.

Norms and cut-off scores for the measurement of social functioning using the ESBA measure were not established in Study 2. This is because definitions of appropriate social behavior may differ from one school context to another, depending on the variety of students' socio-cultural characteristics. Therefore, it is hypothesized that any universal norms and cut-off points defined for the ESBA measure will probably be incomplete and unfair. However, the measure does capture the variability in students' social behavior, which is the greatest source of teachers' concern. The Norwegian adapted ORF measure includes tables of calculated percentile ranks for Grades 2 to 5 based on the sample (n = 2,228) included in Study 3 (see Appendix B). However, due to the purpose of Study 3, appropriate cut-off scores for the ORF measure have not yet been determined for the Norwegian adaptation.

4.3.4 The unified concept of validity

Although each of the three validity concepts (construct, content and criterion-related validty) is useful, Messick (1993, 1995) argues that they should not be considered in isolation since they are complimentary and should therefore be viewed as a unified concept. The unified concept of validity refers to the appropriateness, meaningfulness, and usefulness of inferences based on scores (Messick, 1993, 1995). Hence, construct validity includes aspects of content and criterion-related validity and is therefore seen to represent validity as a whole. Moreover, construct validity comprises several aspects that unify the validity of a score, including relevance, utility, and social consequences, which are issues of concern in all three forms of validity mentioned above. In accordance with Messick (1993, pp. 5-6), these aspects include content, substantive, structural, generalizability, external, and consequential aspects of construct validity, which function as general validity criteria for educational assessments. The unified concept of validity provides a broad view of issues that are not only essential for drawing and using score inferences but also focus on the correlation of test scores with specific criteria in particular contexts of populations

(Kane, 2006). Furthermore, the concept of validity emphasizes the general role of assumptions in score inferences and the importance of examining these assumptions and inferences (Kane, 2006).

The types of validation mentioned above bring a consideration of both the source of justification and the function of the testing into the validity framework. The source of justification is either the evidence, for meaning implications, or the consequences, for value implications. The function of the testing is either the test interpretation or the test use. The unified concept considers both the meaning and value of the assessment's interpretation and use. Based on evidence and rationales, to what degree, if at all, should the test scores be interpreted and used in the manner proposed? (Messick, 1993, p. 13). This overall question of validity and the role of assessments in terms of both the meaning and the use of scores from educational assessments of social functioning and reading proficiency is an important question throughout the three studies in this thesis.

4.3.5 Domain model of construct validity

As mentioned earlier, in validation studies it is important not only to define or develop an assessment of a given construct (e.g., the ESBA in Study 2 or the ORF in Study 3), but to also determine whether the measure of the given construct relates in expected ways to a different measure of other constructs (e.g., the SSRS-T in Study 2 or the NTRPs in Study 3). The assessment construct is based on theories or factors derived from observations of specific sets of items thought to represent the construct. As such, the constructs are abstracts that describe the theories or observed items on which the assessments are based. In accordance with Benson and Hagtvet (1996, p. 84), it is essential to have both a strong theory of the given construct and an understanding of this theory in terms of the way the construct of interest is influenced by and influences other constructs.

The measurement domain model shown in Figure 3 illustrates how different levels of variables within the theoretical ($C_{1, 2, 3}$), empirical ($E_{1, 2, 3}$) and measurement ($M_{1, 2, 3}$) domains relate in the assessment validation process (see Benson & Hagtvet, 1996). In addition to the domain of constructs at the theoretical levels, it is important to operationalize the domain of the observable set of specific items from each of the constructs' empirical domain and examine the relationship between them. It is, as described in Chapter 5.4.4, important to check our data against the theory which the constructs are based upon. The domain model is a way to structure the examination of model-fit. This said, for construct validation, it is important not only to develop assessment

scales that clarify the observable items but to carefully consider both the theoretical factor domain and the empirical domain (Nunnally, 1967 in Benson & Hagtvet, 1996).

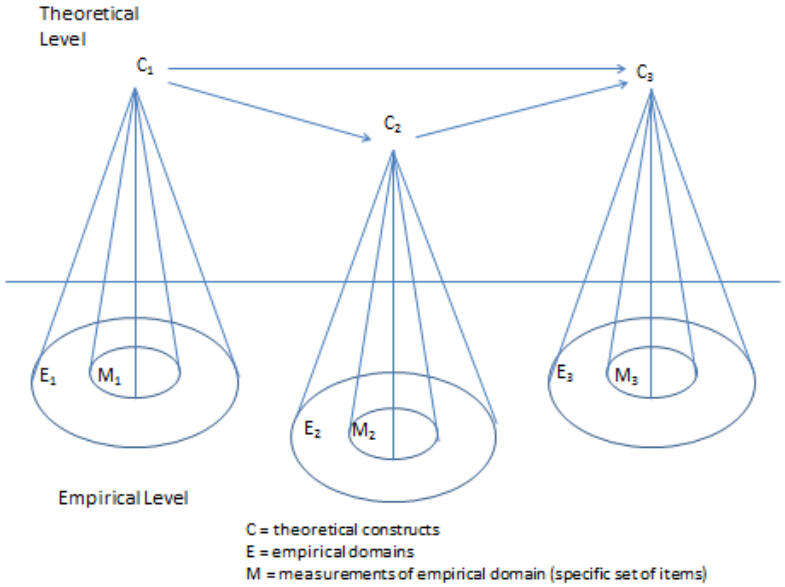


Figure 3. Measurement domain model (Benson & Hagtvet, 1996)

In accordance with Messick (1993) there are two main threats to construct validity which an assessment construct may suffer from: (a) underrepresentation and (b) overrepresentation or construct-irrelevant variance. It is of importance to consider the extent to which the construct of interest is under- or overrepresented when establishing and using an assessment. Construct underrepresentation may occur because the assessment "is too narrow and fails to include important dimensions or facets of the construct" (p. 9). In construct overrepresentation, the assessment "is too broad and contains excess reliable variance associated with other distinct constructs as well as method variance such as response sets or guessing propensities that affects responses in a manner irrelevant to the interpreted construct" (p. 9). These threats to construct validity refer to the adequacy with which the construct domain is sampled.

5. Methodological Considerations

In the following I first address the methodological overview of the research designs and the study procedures used to conduct the three studies presented in Papers I - III. Next, the variables and measurements included in the studies are briefly described. Then, the methods of analysis for each of the three studies and an overall methodological perspective on quality evaluation in terms of psychometric properties are reviewed. The methods for dealing with missing data are presented. Finally, the ethical perspectives of the studies are brought into focus.

Several aspects of construct validity and criterion-related validity are of particular importance in relation to evaluation of the quality of educational assessments investigated in this thesis. To address the various research questions, multiple methodological approaches to quantitative design, samples, procedures, measures, and statistical methods of analysis were used in the three studies. In the following sections, the methodological considerations of these studies are outlined in greater depth than in each of the separate papers.

5.1 Designs

All three studies are based on multiple quantitative, observational research designs. In Study 1, we used a systematic review design (i.e., survey, systematic literature review, and evaluation review) to obtain an overview of the educational assessments used in Norwegian elementary schools and to evaluate their quality. In Study 2, we used a test-retest design with two time-points 8 weeks apart to validate the ESBA measure. For Study 3, we applied a longitudinal design to validate the ORF measure and examine students' growth in oral reading fluency. In addition, a predictive and concurrent correlation design was used to examine the criterion-related validity of both the ESBA (Study 2) and the ORF (Study 3).

5.2 Participants, Samples and Selection Procedures

The data, which were collected solely for the present project, were obtained from multiple respondents (i.e., students, teachers, principals, and parents) in three samples from Norwegian elementary schools. Due to the need for the early identification of difficulties students may face in social functioning and reading, the sample was limited to elementary schools (i.e., Grades 1 to 6) in all three studies. The studies were approved by the Norwegian Center for Research Data (http://www.nsd.uib.no/nsd/english/index.html),

which is the data protection official for educational research in Norway (Appendices C and D).

The purpose of the Study 1 survey was to collect data from approximately 10% of Norwegian elementary schools. Based on previous research on response rates to online surveys, a lower response rate than for other data collection methods was expected (e.g., Baruch & Holtom, 2008; Cook, Heath, & Thompson, 2000; Nulty, 2008). Therefore, 15% of Norwegian elementary schools were randomly selected and invited to participate in the electronic survey regarding the use of social functioning and reading assessments (see Paper I for how the random sample was obtained). As expected, we obtained a low response rate; 57% of the invited schools (n = 234), which represented 10% of the total number of Norwegian elementary schools, completed the survey in the 2014-15 school year.

For Study 2, a random sample of 100 of the 234 schools that participated in Study 1 was selected to provide data on students' social functioning. A total of 151 classroom teachers in 31 schools (Grades 1 to 6) agreed to participate. Based on parents' written informed consent, we randomly drew a total of 793 students from the class lists that the schools provided. Each teacher assessed approximately 5 to 7 of their students using the ESBA and the SSRS-T at two time-points during spring 2015. In addition, 524 parents provided demographic information about the students (see Paper II, Table 1).

In Study 3, a different sample of schools was used. Through a strategic selection process, 21 schools in rural, urban and suburban communities across all Norwegian regions, and a total of 2,228 students in Grades 2 through 5 were included. A previous pilot study in the 2010-11 school year resulted in the use of the ORF-Norwegian adaptation in elementary schools implementing a schoolwide, multi-tier RtI model (Fuchs & Fuchs, 2007; Gresham, 2007) and in trained reading teachers who could administer the ORF measure (Meek-Hansen & Arnesen, 2015). An invitation to participate in Study 3 was sent to those schools. Interested schools were then invited to an information meeting with the researchers. Twenty-one schools agreed to (a) collect ORF data for students whose parents had provided informed written consent and (b) submit data for a given plan. Data were collected by trained reading teachers at three time-points during the 2012-13 school year. The number of participants in each of the three studies is shown in Table 1.

Table 1. *Number of participants in Studies 1 to 3*

	Study 1	Study 2	Study 3	
	Grades 1 to 6	Grades 1 to 6	Grades 2 to 5	Total
Schools	234	31*	21	255
Teachers		151		151
Students (Girls/Boys)		793 (403/390)	2228 (1069/1159)	3021 (1472/1549)
Parents		524		524

Note. * These schools are from the same pool as Study 1.

Because students were involved in Studies 2 and 3, written informed consent was obtained from the parents of all participating students. The consent was translated into the appropriate languages (e.g., English, Lettish, Polish) required from the participating schools to ensure that all parents would understand the purpose and expectations of the study.

Socio-demographic background information for the participating schools was obtained from Statistic Norway's public databases and from the schools. In addition, information (i.e., gender, age, siblings, SES, etc.) on participating students and teachers was obtained from parents and teachers. Table 2 presents an overview of the relevant demographics for the three studies.

Table 2. *Overview of demographic information collected for Studies 1 to 3*

	Study 1	Study 2	Study 3
Teacher gender	X	X	
Students gender		X	X
Student age/grade	X	X	X
Bilingual		X	X
Special education/IEP		X	
Family income		X	
Siblings		X	
Parents education		X	
School size	X		
SES in the area	X	X	X

5.3 Methods and Measures

The constructs of the ESBA and the ORF are based on theories, as described in the previous chapters. Because the two measures were developed within an American school context and the language constructs of one culture do not necessarily fit another, Norwegian cultural and linguistic adjustments were performed prior to the validation studies (Arnesen et al., 2013; Meek-Hansen & Arnesen, 2015). The ORF measure and its relationship with national tests and compulsory assessments of reading proficiency (NTRP) and the ESBA scale and its relationship to the SSRS-T were adapted and examined. General information regarding the methods and measures used to collect the data reported in the three papers are described in the following sections. More specific information about the measures and their psychometrics are reported in Papers I through III, and in the next sections.

5.3.1 Evaluation of assessment quality (Study 1)

For Study 1, we used survey, literature review findings and the review model for the description and evaluation of psychological and educational tests developed by the European Federation of Psychologists' Association (EFPA).

Survey

A survey is a specific type of field study that involves collecting data, through the use of a questionnaire, from a sample of elements drawn from a well-defined population (Visser, Krosnick, & Lavrakas, 2000, p. 223). For the survey in Study 1, an electronic self-administered questionnaire was developed (Appendix E). This provided a way to collect information from the 234 elementary schools across the country regarding the use of educational assessments for social functioning and reading. The questionnaire was piloted with a panel of teachers (to determine face validity) and adjusted before it was sent to the schools. Completion time for the questionnaire was projected to be approximately 5 to 8 minutes. Each school was asked to list the assessments they used, if any. For any assessments used, the school was requested to indicate (a) how often the students were assessed (i.e., > 3 times per year; 3, 2, or 1 time(s) per year; < 1 time per year), and (b) whether the information provided by the assessments was used for instructional decisions (i.e., yes, no, don’t know). The survey provided an overview of the relevant educational assessments that the schools reported using and that were included in the evaluation of assessment quality in Study 1.

Systematic review of literature

The rationale for including a systematic literature review in Study 1 was to collect empirical evidence, gathered in the survey, that identified state-of-the-art assessments. In accordance with the Cochrane Collaboration (http://handbook.cochrane.org/), a systematic literature review should have clearly formulated questions that use systematic and explicit methods to identify, select, and critically appraise relevant research and collect and analyze data from the studies that are included in the review. The pre-defined procedures that guided the review in Study 1 used the general key characteristics proposed by Gough, Oliver, and Thomas (2012), and Green et al. (2011). Specifically, these procedures were outlined in a protocol developed for the study in which the systematic literature review was created (Moher et al., 2015). It included (a) a clearly stated set of objectives with pre-defined eligibility criteria for study participation; (b) an explicit, reproducible methodology; (c) a systematic search that attempted to identify all studies that would meet the eligibility criteria; (d) an assessment of the validity of the findings of the included studies (e.g., through the assessment of risk of bias); and (e) a systematic presentation and synthesis of the characteristics and findings of the included studies (see Appendix F and Paper I). In addition, the eligibility criteria were pre-specified (see Paper I, Figure 1) to answer the study's specific research questions.

The EFPA evaluation review model

One of the main goals of Paper I was to evaluate the quality of the assessments that were reported as being used in the survey and identified through the systematic literature review. As previously mentioned, we followed the procedures of the EFPA review model (Evers, Hagemeister, & Hostmaelingen, 2013; Evers, Muñiz, et al., 2013). The EFPA uses a test review form with international standards and notes for reviewers that consists of two main sections for examining the quality: One is a description of the assessment, and the other is an evaluation of the assessment's quality (see Appendix A).

The description section provides step-by-step details of all the features of the evaluated assessments: (a) a general, non-evaluative description including factual information for identifying the assessment: name of assessment, authors, publishers, and date of publication and adaptation; (b) a classification based on the developer's description, including content domains, intended areas of use, intended population, number of scales and description of variables, item format, intended usage and administration modes, and required administration time; (c) measurement and scoring as described by the

developer, including scoring system and procedures, scales used, and score transformation; (d) a description of generated reports, including computer reports, media, complexity, structure, and context sensitivity; and (e) conditions and costs.

The evaluation section of the test review form includes steps for rating the quality of the following domains: (a) the explanation of the rationale, presentation and information provided, such as the theoretical foundations of the constructs, test-development procedure, content validity, relevant research, comprehensiveness and clarity of documentation, and the provided procedural instructions; (b) test materials, such as paper-and-pencil, computer-based, and web-based, in terms of ease of understanding tasks/instructions, item formulation, design, and clarity of graphical content; and (c) the psychometrics of the assessments supplied by the publishers, authors, and literature review.

The adequacy of the information related to validity, reliability and norms is scored using a rating system that provides descriptions with anchor points. The rating system includes a 4-point scale (i.e., 1 = Inadequate; 2 = Adequate; 3 = Good, 4 = Excellent). In addition, it consists of ratings of "n/a" if an attribute is not applicable, or "0" if the attribute cannot be rated due to insufficient information. Any assessment with one or more scores of "0" or "1" is considered potentially unsafe to use and does not meet the minimum standard to be recommended. In addition to the abovementioned domains, the evaluation of the instruments is based on the overall quality score of (a) norms, in terms of norm-referenced, domain-referenced, or criterion-referenced interpretations including sample groups, sample sizes, local norms, and age of the norms; (b) reliability, consisting of internal consistency, test-retest or temporal stability, parallel or alternate forms, and inter-rater reliability; (c) validity, including construct validity, content validity, and criterion-related validity; (d) quality of computer generated reports; and (e) final evaluation summarizing recommendations, appropriateness and other critical information.

5.3.2 Social functioning (Study 2)

The social functioning of students in Grades 1 to 6 (ages 6 to 12 years) was rated by their teachers using the Norwegian versions of the ESBA 12-items scale (Pennefather & Smolkowski, 2015) and 57 items from the teacher version of the SSRS (Gresham & Elliot, 1990). Due to the validation aim of the ESBA, the Norwegian version of the SSRS-T, which was previously adapted and validated (Ogden, 2003), was chosen as the external outcome measure for the ESBA study. At the time of the study, the SSRS-T was the only relevant instrument that had been validated in a Norwegian context. Therefore, it was used

in Study 2.The ESBA scale and the SSRS-T measure administered to the participating teachers are shown in Appendix G.

The ESBA

The original ESBA scale, which was developed for American students in kindergarten through Grade 3, has been translated into Norwegian. Each of the 12 items are positively stated and cover aspects of students' social functioning in school that may impact their social and academic learning, relationships with others, motivation and engagement in school. The scale is constructed to assess the social behaviors that teachers associate with students' success. The screener uses a 3-point scale (i.e., 3 = often/has mastered the skill; 2 = now and then/needs improvement; 1 = seldom/never/ skill needs to be taught). The items used to monitor progress are the same ones used for the universal screening but use a more finely meshed, incremental 6-point scale to rate the students' progress after social skill interventions based on information derived from the 3-point screening scale. During the process of piloting and adapting the ESBA for Norwegian schools (Arnesen et al., 2013), the face validity of each item was evaluated by an expert panel of teachers. The panel found the items highly relevant for use in all elementary grades. The ESBA screener was applied to students in Grades 1 through 6.

The SSRS

The SSRS is a package of rated items for kindergarten, elementary and middle school students and is available in three versions: 1) the SSRS-T for teachers, 2) the SSRS-P for parents, and 3) the SSRS-SEF for students in Grade 3 and up. For the purpose of Study 2, the SSRS-T was used. It includes three subscales: (a) social skills, (b) problem behavior, and (c) academic skills. The social skills scale comprises 30 items in three social skills domains (i.e., assertion, cooperation, and self-control). The problem behavior scale consists of 18 items within three sub-domains (i.e., externalizing behavior, internalizing behavior, and hyperactivity). Both the social skills scale and the problem behavior scale ask teachers to rate how often the students' skills are observed on a 4-point scale (i.e., 4 = very often; 3 = sometimes; 2 = rarely; 1 = never). The academic skills scale comprises 9 items in which the teacher rates each individual student compared with the other students in the classroom using a 5-point scale (1 = lowest 10%; 2 = almost lowest, 20%; 3 = average, 40%; 4 = almost highest, 20%; 5 = highest 10%).

5.3.3 Reading proficiency (Study 3)

The students' reading proficiency in Grades 2 to 5 (ages 7 to 11 years) was assessed using age-specific, passages from the ORF measure (Appendix H) across one school year at three time-points in an individual setting. Additionally, we obtained the students' scores on the NTRP.

The ORF

The ORF measure is a widely used subtest of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in the United States. Numerous studies have found the measure to be a valid and reliable predictor of students' reading development (e.g., Reschly, Busch, Betts, Deno, & Long, 2009). The construct of the ORF measure is based on theories of reading development (see Chapter 2 and Paper III). Nevertheless, despite its prevalence in the United States, this particular measure has never been adapted to a European setting with a semi-transparent orthography (e.g., Norwegian). The design and administration procedures of the ORF measure were adapted from an American school context to a Norwegian school context. However, the 36 passages used for screening and the 60 passages used to monitor progress were originally Norwegian written passages calibrated for students in Grades 2 to 5. Student scores based on the number of correct words read aloud per minute on 9 passages (3 passages at each of the 3 test-points) for each of the four grades across the year were obtained. The overall purpose was to (a) identify students who may need additional reading intervention, and (b) monitor progress toward instructional goals. Screening assessments are designed to be used at three time-points during each school year (fall, winter, spring). The passages for progress monitoring are used to assess students' responses to interventions based on the screening data. For the purpose of Study 3, the screening passages were used. The measure provides percentile ranks at each measurement time-point. Usually, scores below the 20th percentile indicate high risk, scores between the 20th and the 40th percentiles indicate moderate risk, and scores above the 40th percentile indicate low or no risk (see Appendix B). Cut-off scores were not calculated for the ORF screener within this study.

The National Tests of Reading Proficiency

The National Tests of Reading Proficiency (NTRP) are compulsory tests and assessments. NTRP data were collected for the participating students and transferred by the schools' data manager to an Excel spreadsheet designed for the study. In Norway, national tests and assessments are group-administered once per year: 2nd and 3rd graders are

assessed within a two-week period in the spring and 5th graders are assessed at one occasion in the fall. Because the 4th graders at the time of the study, were not administered a national reading test, the 5th graders' NTRP scores (that is, the scores for the 4th grade group, who started Grade 5 in the fall) were used for the analyses. See Appendix B in Paper III, for more specific information about the NTRP.

5.4 Statistical Methods of Analysis

Because the research questions examined in this thesis considered the use of educational assessments and their psychometric properties, multi-statistical methods of analysis were required. A description of the methods that were chosen to derive the specific empirical results that were relevant to the research questions is presented in this section.

5.4.1 Descriptive orientation

Descriptive statistics requires a process of mapping out *how* something is occurring and *what* is occurring rather than *why* (Rosenthal & Rosnow, 2009). The survey data, the results of the systematic literature review, and the EFPA evaluation of the assessments used are described and summarized in Paper I. These data were of a descriptive nature and reviewed information obtained through the systematic literature search that analyzed the original measurement data. These descriptive data allowed us to present the findings in easily accessible tables and figures that include sample sizes, the number of schools using assessments for social functioning and reading, the percentage use of each reported assessment, a flow chart of the systematic literature search, descriptions and characteristics of the assessments, and an overview of material quality and documented psychometric properties (see Paper I). Demographic information and descriptive statistics, including sample sizes, reliability coefficients, means, standard deviations, minimum and maximum scores, skewness, and kurtosis, are presented for Study 2 (see Paper II, Tables 1 to 6) and Study 3 (see Tables 1 to 8 and Appendices A to C in Paper III).

5.4.2 Inter-rater agreements

Inter-rater agreement provides a way to score the level of agreement and disagreement between two or more assessors' ratings of the same individuals on the same measures or of the same measures for the same evaluation procedures (e.g., Study 1). This is critical for reporting reliability because it relates to accuracy and variability (Crocker & Algina, 2008). For the purpose of Study 1, we conducted a *variance component analysis* of the overall indicator ratings for the assessments' material qualities and psychometric

qualities, as recommended by the EFPA. This allowed us to better understand the sources driving differences in ratings and disagreements among the assessors (see Paper I).

5.4.3Multilevel analysis – intra-class correlation

As a result of the participant selection process for Study 2, the students were nested in classrooms and schools, which might violate the independence of the data. Therefore, we obtained intra-class correlations (ICC) using a three-level model to examine whether this dependence impacted the results. In this case, the ICC was used to measure the extent to which members of the same category (i.e., students in the same classroom in the schools) are more or less similar to each other (Cohen, Cohen, West, & Aiken, 2003). The ICC obtained in Study 2 measured whether students in the same classroom and at the same school were rated differently on the ESBA scale by their classroom teachers compared with students in the other participating schools. The ICC was calculated at both the classroom level and the school level after both pre-tests and post-tests (see Paper II, pp., 10-11).

5.4.4 Common factor model – structural equation modeling

Factor and item analysis allow an evaluation of the stability, internal consistency and equivalence of scale items and the correlation between them (Brown, 2015; Kline, 2015). Structural equation modeling (SEM) was used in Studies 2 and 3 to estimate the relationship between the observed variables (manifest variables) and the underlying constructs (latent variables). The latent variables typically used as psychometric measurement factors in SEM represent the shared variance of the observed variables. In other words, several observed variables have "something" in common that "moves together" in a common unobservable latent variable. However, a prerequisite for this is that the indicators of the constructs correlate sufficiently to provide evidence that something is shared (Little, 2013). This is called the common factor model and is used to describe variability among correlated observed variables or indicators in a lower number of latent factors. In contrast to observed measured variables, which typically have some measurement errors, unobserved latent variables do not have measurement errors since those are accounted for in the model (Kline, 2015).

For the purpose of Study 2, in which the ESBA scale was adapted and validated for a new cultural context (from American to Norwegian), a common factor model was used. An exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA) within an SEM framework based on the common factor model were conducted (see Paper II). For the

purpose of Study 3, second-order latent growth curve modeling within an SEM framework was used to examine (a) the longitudinal measurement invariance of the ORF measure, (b) the growth in oral reading fluency within and across Grades 2 to 5, (c) the relative stability of the ORF measure, and (d) the relationship between the ORF measure and national tests of reading proficiency (see Paper III).

EFA and CFA

In psychometric evaluations of multiple-item assessments that include construct validation, the common factor model with both an EFA and a CFA is commonly used (Brown, 2015). In the common factor model, "each indicator in a set of observed variables is a linear function of one or more common factors and one unique factor", and the variance of the indicator is split into a common and a unique variance (Brown, 2015, p. 11). The common variance is based on an estimate of the variance shared with other indicators in a scale and is accounted for by a factor. The unique variance is a combination of the reliable variance and the random error variance in a specific indicator. However, the EFA and the CFA differ fundamentally. The EFA is generally not a part of the SEM but is an exploratory technique that uses procedures to derive factors. In contrast with the CFA, the EFA does not require specific hypotheses regarding which indicators correspond to one or several factors. Given this, the CFA requires a strong empirical or theoretical foundation to specify and evaluate the factor or factors in the model.

In the EFA, the indicators are allowed to load on one or more factors that are unrestricted and not predefined (Kline, 2015). Brown (2015, p. 12) clarified the relationship between the EFA and CFA by saying, "EFA is typically used earlier in the process of scale development and construct validation, whereas CFA is used in later phases after the underlying structure has been established on prior empirical (EFA) and theoretical grounds" (Brown, 2015, p. 12). Based on Brown's clarification, an EFA of the ESBA pre-test items was used to identify the factors underlying the Norwegian version and thus examined the construct validity of how well the items measure what they are theoretically meant to measure. The CFA of the post-test ESBA items was used to validate the factor structure established at the pre-test using the EFA (Paper II).

Longitudinal growth factor structural equation modeling

When examining developmental processes (e.g., student reading proficiency in Study 3), longitudinal latent growth curve models offer a particularly useful way to predict and explain individual differences in growth curves and changes over time (Duncan,

Duncan, & Stoolmiller, 1994; Little, 2013; Rogosa, Brandt, & Zimowski, 1982; Stoolmiller, 1995). As for the latent variables explained above, latent growth analysis provides accurate estimates that are controlled for measurement error *if* the same measures are used for the same group over time (Kline, 2015; Li, Duncan, Duncan, & Acock, 2001). Thus, the use of latent variables to measure stability over time allows the analysis of the common variance without measurement "disturbances" (Little, 2013; Kline, 2015). This results in an ability to not only estimate the relationship between latent variables but to make those estimates without measurement errors.

Growth curve models are useful for examining both the rate and the shape of the changes that characterize specific groups (e.g., students across grades). Moreover, these models are characterized by the translation of sets of intercepts and slopes for the whole sample of individuals into the mean intercept and slope and their distributions (Little, 2013). In accordance with Duncan, Duncan, and Stoolmiller (1994), a developmental model for a sample of individuals should reflect individual differences in the slopes and intercepts of straight lines if the trajectories are well described by a collection of those lines (e.g., Figure 4, Paper III). Although latent growth models have the same strengths as other models within SEM (i.e., they test the adequacy of hypothesized growth, incorporate covariates, and correct for measurement errors in observed variables), several requirements should be taken into consideration: large samples; multi-normally distributed variables; changes that are systematically related to the time intervals (i.e., the individuals should be observed at approximately the same time, and the number and intervals of measure occasions should be the same for the whole sample) (Duncan et al., 1994). As mentioned earlier, for the purpose of Study 3, a second-order latent growth curve model was established in which the second-order factors (the intercept and the slope) were extracted from the covariation among the first-order constructs (the three ORF factors at different time-points), and the first-order factors were extracted from the observed variables (unique ORF passages) at each of the three measurement occasions (Paper III).

Estimation methods

Maximum likelihood (ML) is the most commonly used estimation method within SEM for the analysis of continuous variables (Brown, 2015; Kline, 2015; Little, 2013). The estimates in ML simply maximize the likelihood that the observed data were drawn from the population and that the distribution is based on continuous variables and assumed to be normal. Thus, ordinary ML can not be used to estimate models with categorical

variables or non-normal data. In Study 2, a normal distribution of the ESBA scores was not assumed, because the ESBA measure has only three ordinal values. The measure captures the variability in behaviors that teachers are most concerned about (i.e., mastered; needs improvement; needs to be taught). This is why ordered categorical analysis procedures are used for the EFA and the CFA; they account for the potentially non-normal data that result from such a measure. Other analytical procedures, such as correlations, are fairly robust to non-normal distributions, and their consistency leads to confidence in their results. Hence, the factor analysis ML estimation with robust standard errors (MLR), which produces unbiased estimates for models with categorical outcomes (Brown, 2015), was used with Mplus (Muthén & Muthén, 1998-2012). In Study 3, a full-information ML estimation was used with the lavaan package (Rosseel, 2012) in the statistical software environment R to handle missing data and make use of all available information for each individual.

Model fit

It is important not only to trust the data and determine how these fit the hypothesized model but to check the data against the theory upon which the constructs are based. The domain model by Benson and Hagtvet (1996) presented in Chapter 4 is one way to structure a model to compare the relationship between the different estimated constructs or factors at a theoretical level and between the factors and the different observed variables at the measurement and empirical levels (see Figure 3). Different models within SEM require different indices or measures of fit based on a statistical and/or a modeling approach for which a range from relative fit to absolute fit can be classified (Little, 2013). Although there are no strict rules for determining goodness of fit, a key may be how well the measurement model (e.g., factor loadings and factor correlations) can reproduce the observed relationship among the indicators (Brown, 2015).

The most commonly used measure for evaluating model fit relies on a statistical approach, namely, the chi-square (χ^2) goodness-of-fit test (Brown, 2015). The χ^2 is a significant test for evaluating the model against the data. In other words, it is used to determine whether the observed data are consistent with or different from the model. However, the χ^2 difference test is sensitive to sample sizes. That is, large samples will nearly always be significant and can lead to the rejection of a highly satisfactory model, whereas small samples can lead to the acceptance of models with many misalignments. Moreover, the χ^2 difference test is a test of exact fit. That is, if there is no difference between the estimated model and the observed data, the 0-hypothesis must be accepted.

Therefore, to evaluate the relative degree to which a given model fits the data, an alternative modeling approach has been developed. This approach includes several fit indices that incorporate absolute and relative fit measures.

The root mean square error of approximation (RMSEA) uses a saturated model and provides an absolute fit ≤ .05 and an acceptable cut-off value of approximately .08 (Brown, 2015; Little, 2013). Another measure of absolute fit is the standardized root mean square residual (SRMR), which considers a range of values between .00 and 1.00. The smaller, the value the better the model fit; .00 equals a perfect fit, and values below .08 are generally recommended (Brown, 2015; Little, 2013). The Comparative Fit Index (CFI; Hu & Bentler, 1998) and the Tucker–Lewis Index (TLI; Tucker & Lewis, 1973) compare the models against a baseline model that has uncorrelated observed variables values. Whereas the CFI is normed with values between .00 and 1.00, the TLI is not and can have values outside the CFI range. The recommended CFI and goodness-of-fit values are ≥ .95 (Brown, 2015).

Model fit indices are unavailable from ML methods used with categorical outcomes or when non-normal data is not assumed. Since Study 2 included both categorical and non-normal data, model parameters with robust ML methods and model fits with robust weighted least squares were estimated for both the EFA and the CFA. The EFA used the oblique rotation geomin, which allows correlations between factors (Browne, 2001). Because model fit criteria were unavailable for the EFA with categorical data, criteria for a CFA of categorical data were adopted from Schreiber et al. (2006) using CFI and TLI. However, no criteria value for the SRMR for categorical data has been suggested. In Study 3, the χ^2 difference test and the goodness-of-fit indices RMSEA, SRMS and CFI were used.

5.5 Missing Data

According to Little (2013), missing data are not problematic per se, but a problem could arise depending on the way missing data are treated. Missing data may bias estimations and decrease statistical power. Missing data may be (a) missing completely at random (MCAR) due to unpredictable reasons and unrelated to any variables of interest, in which case bias is non-existent; (b) missing at random (MAR) due to predictable reasons and related to variables of interest, but these can be accounted for by other variables, and bias is recoverable; or (c) missing not at random (MNAR) as a result of systematically missing data (i.e., a subject's levels of a particular variable), and related to variables of interest that

cannot be accounted for by other variables, thus yielding bias (Little, Jorgensen, Lang, & Moore, 2014; Rosenthal & Rosnow, 2009). The ML estimation provides a procedure for dealing with missing data (Brown, 2015).

In the studies presented in this thesis, several mechanisms of missing data had to be accounted for. First, as previously mentioned, the survey had an expected low response rate (57%). The survey included descriptive data only, and missing data were explained by the extent to which they were representative (Paper I). Second, in Studies 2 and 3, data were MAR. In Study 2, however, 2.4% and 11.5% of the participating students' data were missing at pre-test and post-test, respectively. Therefore, students without data at each time-point were excluded from the analysis. However, we did not expect these missing data to bias our findings (see Paper II). To account for the missing data in Study 3, the full information maximum likelihood (FIML) was used (see Paper III). The FIML approach restores power in cases of both MCAR and MAR and directly adjusts the parameter estimates to reflect the values that would have occurred without missing data (Little et al., 2014).

5.6 Ethical Perspectives

Reporting research results and acquired knowledge requires a high degree of ethical discretion and compliance with formal guidelines and legislation. The Norwegian National Committee for Research Ethics in the Social Sciences and the Humanities (NESH, 2016) defined guidelines for research ethics in the Research Ethics Act (2007). To pursue the indisputable demand for humility and fundamental respect for the work of others - and report these in a fair and worthy manner based on ethical-theoretical principles - several ethical considerations are reported as they relate to the three studies in this thesis.
First, evaluating the quality of other researchers' work, as was done in the review study (Study 1), could present an ethical dilemma. This is due to the disclosure of different quality levels or ratings derived from the assessment instruments that might favor some researchers and disfavor others. Another possible ethical dilemma involves the potential attention resulting from new knowledge gained through the study and how decision-makers and school authorities use that information.

Second, since the research project targets students and directly involves information about their social functioning and reading proficiency, it is critical to consider the associated ethical aspects. Hence, Studies 2 and 3 were approved by the Norwegian Social Sciences Data Services (NSD), and privacy is ensured through the Personal Data Act

(Appendices C and D). All personal information was unidentified and treated with confidentiality. Participation was voluntary and required written informed consent from parents and the provision of information to the students about what participation entailed. This included the right of the students, the parents, and the schools to withdraw from the project at any time. In the current studies, the results are reported at the group level, which meets the strict requirement for anonymity.

Third, regarding Studies 2 and 3, it was important to be strategic when planning what assessment information to obtain to predict developmental difficulties and how that information is used. Moreover, within these studies, it was possible to capture unintended findings and vulnerabilities related to students' developmental and learning difficulties that required interventions. However, the main focus when reporting findings is on the benefits of learning for these groups. It is also important to consider whether participation in the study is in the best interest of the students and those demonstrating problems related to social functioning and reading (e.g., minority groups, low-performing or disabled students). If a student needs further diagnosis to determine the best fit for intervention, participation in the study could compromise the student's growth and should be re-considered.

6. Summaries and Discussion of Main Findings

In this chapter, I summarize and discuss the main findings of the present research, specifically in terms of several aspects of the assessments' validity that were explored in the three studies. Each of the three studies presented in Papers I - III contributes unique aspects to this focus area. The overall goals were to provide educational practitioners (i.e., teachers), leadership (i.e., principals, school leaders) and policymakers with (a) knowledge of educational assessments for both social functioning and reading proficiency, (b) new insights regarding the quality of the assessment practices and approaches used in Norwegian elementary schools, (c) evidence that can define and influence needed changes in schools' assessment practices that in turn may impact students' learning and development in social functioning and reading, and (d) additions to the pool of valid educational assessments for early identifying students' difficulties in social functioning and reading. Due to the thesis's overall objectives regarding the use and validity of educational assessments in elementary schools for the early identification of students' social functioning and reading difficulties, the need for changes in practices and policies is discussed. The chapter ends with some perspectives of limitation of the presented studies and the need for further research.

6.1 Assessing Social Functioning and Reading Proficiency (Study 1)

Study 1 sought to investigate the quality of educational assessments for social functioning and reading proficiency that are currently used in schools. A survey of 234 Norwegian elementary schools revealed that 90% used reading assessments at least once a year, while 31% used assessments to measure the students' social functioning.

6.1.1 Common use of assessments without documented evidence

The most frequently used assessments, with the exception of the mandatory national tests, were teacher-made or had no documented psychometric properties. Moreover, we did not find any relationships between the schools' use of assessments and their use of information derived from the assessments to guide instructional decisions. The analysis of the six EFPA evaluation elements found that none of the 3 included assessments of social functioning reported validation studies, while 11 of the 24 included reading assessments reported validation studies.

Consistent with previous findings from systematic reviews of the quality of social skills and reading assessments conducted outside Norway (Cordier et al., 2015; Floyd et

al., 2015; Gotch & French, 2014; OECD, 2015; Swedish Council on Health Technology Assessment [SBU], 2014; Standards & Testing Agency, 2015), the results are challenging considering the importance of early identification of students' specific difficulties to prevent severe difficulties from developing. Although similar findings could be expected in the current study, evidence from one country's school context and culture cannot automatically be transferred to another (Borsa, Damásio, & Bandeira, 2012; van Widenfelt, Treffers, de Beurs, Siebelink, & Koudijs, 2005). Therefore, the replication of studies across countries is important both for increasing the pool of studies in the same field and for gathering evidence that fits within the context of the current study.

6.1.2 Weak theory-based constructs and lack of pschometric evidence

Theories underlying the constructs were partially described. Although the quality dimension of the test materials was good or adequate for approximately half of the assessments, there was no evidence to support the overall quality of the psychometric properties for the majority of the reviewed assessment documentation. Moreover, the findings question the trustworthiness of assessment practices in schools. Given the importance of strong theories that underlie valid constructs in high-quality assessments to ensure that they actually assess what they are intended to (Benson & Hagtvet, 1996; Merrell, 2009; Messick, 1993, 1995; Thorndike & Thorndike-Christ, 2014), the findings are critical and indicate a need for improvement. The weakness in construct theories and the lack of psychometric evidence are worrisome. First, the construct theories that impact the assessments' validity are weak. The theories may not clearly indicate what constructs exist, if any; and if they do, they lack an understanding of the importance of communicating so. If a lack of transparency of the constructs used to assess students' learning and growth is present, it may, to some degree, explain the teachers' use of assessment information for instructional decisions without regard for the validity of the assessment tools. In line with Monsen (2013), it seems that a lack of awareness regarding the use of the assessments and what they actually measure may be present. This can be substantiated by Haug's (2014) argument that knowledge regarding the situation for students receiving special education is not well developed or widespread in the Norwegian education system.

Second, the lack of evidence to support the overall quality of the psychometric properties for the majority of the reviewed assessments is challenging. Supporting evidence is not only critical for selecting and using such assessments but is also needed to

address the lack of awareness among those who distribute assessments and use them without questioning their quality or validity. The lack of evidence does not necessarily imply that the psychometric properties are weak, but without evidence, it is unclear whether the psychometric properties of the assessment are supported. Moreover, challenges arise when information regarding psychometric properties exists but is not available to end-users (i.e., teachers, students, parents, school leadership).

The results may be interpreted in light of an educational culture that lacks the competence to judge assessment quality. Hence, there may not be any general expectations that the quality of assessments will be explicitly stated (OECD, 2015; The Norwegian Ministry of Education, 2015). This is in line with Monsen (2013), who found that Norwegian teachers have low expectations of how reading assessments can be used to improve students' learning and that they are uncertain and ambivalent about using the reading assessments. Moreover, this attitude toward assessments may be due to the absence of a tradition of using such procedures in schools or because the currently available assessments are considered too time-consuming (Elliot, Huai, & Roach, 2007; OECD, 2007).

6.1.3 The gaps in assessment practice and competence

The current review confirms a gap between existing assessment practices and a practice that is likely to be more effective (see Chapter 3). To close this gap, there is a need for improved assessments for both social functioning and reading proficiency. In addition, there is a need for improved assessment competence, which could lead to changes in practices, principles, and policies at different educational levels. In general, the results indicate that the pool of reviewed assessments lacks valid instruments for assessing social functioning and reading proficiency. In particular, there is a need for valid assessment instruments to screen and monitor students' progress in social skills and reading comprehension and guide instructional decisions in Norwegian schools.

Moreover, the results indicate a need to promote students' social well-being, academic learning and growth by providing teachers with evidence-based assessments that are brief and easy to use. In addition to the lack of psychometric evidence, the need for improvements can be explained by the finding that teachers typically assess students' achievements in the areas of social functioning and reading using teacher-made instruments. Finally, the results might challenge current thinking regarding the commercial

test industry in terms of the development of educational assessment instruments that are high-quality, valid, not-for-profit, widely available, and easily accessible.

6.2 Validation of the Elementary Social Behavior Assessment (Study 2)

Building on the results of the review study presented in Paper I, Paper II set out to examine the validity of the adaptation of the ESBA screener for the Norwegian elementary school context. The psychometric properties of the ESBA were examined at two time-points in a sample of 793 students in Grades 1 to 6, rated by 151 teachers in 31 schools. The SEM used to examine the construct validity of the ESBA demonstrated that the data fit the construct's underlying theories.

6.2.1 Academic engagement and peer social relations

In contrast to the study by Pennefather and Smolkowski (2015), which demonstrated a one-factor scale for the twelve items, the initial EFA of the current study suggested one or possibly two factors, and subsequent CFA at post-test confirmed the two-factor model. The two factors, namely, *Academic Engagement* and *Peer Social Relations*, fit nicely with theories on the constructs of social functioning in terms of social behaviors, learning-related social skills and academic engagement that promote students' social and academic competences, as defined in Chapter 2 (Al-Hendawi, 2012; Beauchamp & Anderson, 2010; Cordier et al., 2015; DiPerna, 2006; Gresham et al., 2010).

In addition to the high correlation between all items and the total ESBA scale, the two factors and their subscales correlated highly at both pre- and post-test. Moreover, the estimated score reliabilities for both the full ESBA scale and the two factors' subscales and the test-retest reliability produced strong values. Criterion-related validity, established with the SSRS-T, demonstrated concurrent and predictive correlations at both times for all students. All correlations were consistent after controlling for the students' background variables. Moreover, the ICCs suggested that teachers across schools and classrooms rated students quite similarly.

6.2.2 Consistency in teachers' ratings of students' social skills

Consistent with Pennefather and Smolkowski (2015), the findings revealed that to a great extent, the teachers rated their students as having mastered (highest score) the specific social skills measured by each of the scale's twelve items. Moreover, the high scores were expected because most students typically acquire basic social skills that promote fluent social functions in school, and relatively few struggle in this area (Bru, 2011; Cummings, Kaminski, & Merrell, 2008; Hinshaw, 1992; Skogen & Torvik, 2013).

However, for the three-point scale, it can be hypothesized that the instrument produces ceiling effects, which might devalue the accuracy of the scores and affect the correlation sizes (e.g., Thorndike & Thorndike-Christ, 2014). Thus, to account for the potentially non-normal data that arise from measures like the ESBA, the ordered categorical analysis procedures of the EFA and CFA were used. Other analysis procedures, such as correlations, are fairly robust to non-normal distributions, and because the results are consistent, there is confidence in the results. Moreover, it is likely that a more normal distribution of the data would emerge if a four-point scale had been used, such as the Norwegian validated SSRS-T (Ogden, 2003).

The ESBA's three-point construct is concerned with the extent to which teachers discriminate between the students who can perform or master the requisite social skills, those who struggle or need some support, and those who "can't" or "don't" perform the required social behavior. In accordance with Gresham (2002), these distinctions are important because the resulting instructional decisions can provide the students with specific support to improve their social functioning. Although a teacher may want his or her students to achieve mastery (and thus score high), the scale does not discriminate among those students who demonstrate acceptable through excellent performance (good, very good, or really awesome) in terms of social behavior. That is, teachers are not asked to discern differences in the skills of students who are not raising any concerns. However, teachers might wish to discriminate between those who are at risk and those who are struggling and need help. Therefore, the scale is constructed to capture the variability in social behaviors that teachers are concerned about, but it does not provide a defined cut-off limit. The scores that represent social functioning problems that indicate a need for help should be left up to the teacher within the actual classroom environment. In fact, teachers already know who among their students has achieved mastery for each skill and who has not; therefore, their judgments have been built into the scale. In that sense, any score below an approximate average that represents mastery would indicate some degree of problems.

This said, the ESBA scale, which is derived from research that identifies prosocial behavioral skills in students that teachers consider important (Gresham, 2007; Walker & McConnell, 1995; Walker & Severson, 1992), can meet the needs for improving students' social and academic achievements and promoting success in early elementary school when used for students with social behavioral difficulties (Haug, 2014; OECD, 2015; The Ministry of Education, 2009, 2017; Tyler-Merrick & Church, 2013). Consistent with previous research regarding the importance of social functioning and academic proficiency

to students' success (e.g., Al-Hendawi, 2012; Gresham, 2007; McClelland et al., 2000; Sutherland & Wehby, 2001), the results suggest that the adapted ESBA screener is useful for providing teachers with information that can inform instructional decisions that promote students' learning-related social skills.

6.2.3 The ESBA is a valid screener to guide specific social skills intervention

The two subscales of the ESBA screener, *academic engagement* and *peer social relations,* might allow teachers to specify particular skills in the domain in which students need supplemental support (Paper II). Thus, rather than focus on a cut-off score, the highly valid ESBA screener takes a different approach. That is, teachers respond to struggling students on the overall scale or the two subscales depending on (a) how many students need improvement in the same skill or (b) how many skills a particular student needs to improve. Then, interventions can be initiated immediately, and students' progress on the specific skill can be monitored with the ESBA.

Teacher-constructed assessments, which are the most commonly used social functioning assessments in Norwegian schools, tend to be biased and prone to measurement errors because teachers rate their students subjectively based on observations and common sense. In contrast, formative scales, such as the ESBA and the SSRS-T, allow teachers to use their experience-based knowledge to rate students' skills using valid constructs based on strong theories. Although teachers may have different expectations regarding how students are capable of performing, the ICCs calculated for the ESBA scale in the present study indicate that teachers rate students quite similarly.

6.3 Growth in Oral Reading Fluency (Study 3)

Building on the results presented in Paper I, Paper III aimed to examine the psychometric properties of the adapted ORF for measuring the students' reading decoding, accuracy and fluency in Norwegian, a semi-transparent language. The current ORF study is the first to use ORF passages developed for students in a school context outside the US. The unique Norwegian passages were created to assess students' reading proficiency in terms of decoding, accuracy and oral reading fluency (Good & Kaminski, 2002; Meek-Hansen & Arnesen, 2015) and, in terms of reading difficulties, to meet the needs for improved knowledge regarding students' reading development and reading assessments in Norwegian elementary schools (Paper I).

6.3.1 Longitudinal invariance and a measure of growth in reading

Second-order latent growth curve modeling was used to examine the longitudinal measurement invariance of the adapted ORF measure and the initial status and growth in oral reading fluency for 2,228 students in Grades 2 to 5. First, the findings indicated that all nine ORF passages at each grade level administered at three time-points during one school year indicated longitudinal measurement invariance, but some stood out empirically. This indicates the difficulty of developing passages with different content that still provide an equivalent measure (Cummings, Park, & Bauer Schaper, 2013). The unique concept of the ORF measure, which consists of passages with different content that measure the same construct, avoids retest effects. Moreover, in line with several ORF studies conducted in the US (e.g., Reschly et al., 2009), high relative stability of the ORF measure was found across the four grades.

Second, the initial status and growth in oral reading fluency were identified for the participating Norwegian students in Grades 2 to 5. Moreover, the oral reading fluency growth curve models demonstrated linear growth in Grades 2 and 3 and nonlinear growth in Grades 4 and 5. The initial individual differences varied more than the growth rates, which were positive but were greatest in Grades 3 and 4. Consistent with Baker and colleagues (2008), the findings demonstrated that the adapted ORF measure can help to identify struggling readers at an early stage and monitor their growth over time within the school year and across years. That is, students who initially fall behind can be given efficient early reading intervention within an RtI model that meets their needs for support (Burns, Silberglitt, Christ, Gibbons, & Coolong-Chaffin, 2016; Fuchs & Fuchs, 2007).

6.3.2 The ORF measure - an indicator to identify reading dificulties for interventions

The examination of the criterion-related (concurrent and predictive) validity of the ORF measure in relation to the NTRP revealed moderate to strong correlations in all grades. Notably, whereas the reliabilities found in the current Norwegian study were strong and similar to those of several studies conducted in the US (e.g., Ardoin et al., 2013), the criterion-related validity demonstrated a wider range of correlations between the ORF measure and external reading measures than the US studies did. That is, many of the subtests used in the NTRP might not be as good external measures of criterion validity as was originally thought.

Analyses of the data by gender and ethnicity did not reveal any differences in the current study. Moreover, the findings were in line with recent curriculum-based ORF

studies, mostly conducted in an American school context in English and Spanish (Goffreda, & DiPerna, 2010). That is, the concept of the ORF measure seems to work as well for students in elementary schools in Norway as in the US. The adapted version of the ORF measure was found to be an important developmental indicator of reading proficiency and may be useful in identifying and monitoring students at risk of reading difficulties who could benefit from reading interventions.

In contrast with summative and static high-stakes assessments and teacher-rating scales, the formative, curriculum-based ORF measure is a direct assessment approach that allows assessments of students' responses to intervention at different tiers within a school (e.g., Burns et al., 2016). Given that the valid ORF construct is used to inform instructional decisions, it can potentially help to promote positive change for struggling students and minimize the gap between poor and proficient readers in the long run (Stewart, Benner, Martella & Martella, 2007). The risk of biases is reduced because the scores are based on the number of words a student reads correctly per minute. This aspect of the ORF measures the student's reading proficiency in terms of decoding, accuracy, automaticity and fluency using connected text and does not allow any interpretation. However, inter-rater correlations among assessors would be important to calculate. This is also important for the qualitative aspect of the measure, which reflects the teacher's observation of the student's reading behavior and prosody during the test situation. The present study did not do so.

Finally, consistent with the Simple View of Reading model (Gough & Tunmer, 1986; Hoover & Gough, 1990), as well as previously findings from studies of oral reading fluency (e.g., Biancarosa & Shanley, 2016; Fuchs, et al., 2001; Rose, 2006; Smolkowski et al., 2016), the current study demonstrated that the ORF construct measures what it is intended to. Moreover, students' oral reading fluency, in terms of their abilities to decode words in a connected text correctly with accuracy and automaticity at a fluent speed, allows efficiency and comprehensiveness in reading. The purpose of this study, however, was limited to investigating students' growth in oral reading fluency and examining the psychometric properties of the Norwegian adaptation of the ORF measure and not a student's ability to retell and comprehend the reading passages, as discussed below.

6.4 Limitations of the studies

The main limitations of the presented studies are described in each of the three papers. In addition, some limitations of the present research are noted as follows. First, the conducted research is limited to (a) a review of the quality of the psychometric properties of the

educational assessments of social functioning and reading used in Norwegian elementary schools, and (b) the validity of the adapted screening measures: the ESBA scale and the ORF measure. That is, although the ESBA scale and the ORF measure are designed for both screening and progress monitoring to guide instructional decisions, the thesis did not include intervention studies in which the assessments were used to measure students' responses to the interventions. Further, the use of the Norwegian-adapted ESBA and ORF for monitoring progress was not validated within the studies in this thesis.

Second, the response rate (57%) of the schools participating in the survey may impact the ability to generalize the results to schools not represented in the study. However, the main purpose of the study was to evaluate the quality of the assessments used in schools. It is reasonable to assume that the survey responses provide information regarding the instruments most commonly used in elementary schools to assess students' social functioning and reading proficiency. Moreover, the systematic literature review (Paper I), which was the basis for the EFPA quality evaluation of the used assessments, is constrained by the lack of validation studies; it reviewed only documented information provided us from the publishers and not original data. In accordance with the EFPA review criteria, the few available original studies were old and therefore did not meet the required quality criteria for the validity evaluation. Therefore, we do not know whether these assessments would have been rated with higher psychometric qualities in new studies with new samples and modern test theories/methodology.

Third, the ESBA validation study did not provide data regarding the participating teachers' experiences with the ESBA scale. Moreover, the small number of participating schools may limit generalization to other teachers' ratings of other students in other schools. However, the current study suggests that the participating teachers at different schools rated their students similarly.

Fifth, the limited focus in the present research does not include the role of parents, the student-teacher relation or other significant factors (e.g., SES, mental health, cognition) which may impact a student's learning and development outside and within the context of school. Therefore, we do not know if this limitation could impact the determination of possible variation in students' scores. Finally, another methodological limitation concerns the use of appropriate ORF cut-off scores to identify some-risk and at-risk students for difficulties.

6.5 Conclusion and Further Perspectives

The studies in this thesis contribute to the fields of educational assessment of students' social functioning and reading proficiency and clarify the link between these two areas of learning and development. The three presented studies are the first of their kind in a Norwegian educational context and culture. Thus, they are unique contributions to the field of educational assessment instruments and add to the pool of knowledge of available evidence-based assessments. The ESBA and the ORF screeners may not only help to reduce the identified gap in assessment practices but may also expand the availability of high-quality, easy-to-use measures of social functioning and reading.

As described in this thesis, the foundations of validity emphasize that assessments need theory-based constructs which are evaluated and adjusted to new cultural contexts with high quality (efficacy) to fit the practice (efficiency and meaningful) and to be recommended (useful). Hence, the presented thesis of the validity of educational assessments used in Norwegian elementary schools has outlined the importance of having both social functioning and reading proficiency as a main focus of students' learning and development. To illustrate the importance of having assessments with high-quality psychometric properties, the three studies have not only been independent studies providing evidence to a Norwegian school context but also providing a framework for the model of change explained in Chapter 3. In addition, the evidence derived from the three studies and their underlying construct theories of social functioning and reading proficiency may contribute to the improvement of the quality and use of assessments for preventing failure for students at risk. Ideally, the studies may contribute to changes in practice in regard to selecting and using assessments for screening and monitoring progress to guide instructional decisions.

In summary, the current work examined the importance of high-quality effective screening in early Norwegian elementary school and evaluated the assessments in current use and two screening assessments for social functioning and reading, namely the ESBA and the ORF. Its findings have practical implications; namely, that evidence-based assessment practices that are useful for teachers are not only important from an evaluation perspective but also from a pragmatic point of view. Schools have limited resources in terms of time, decreased budgets, limited skills, varying interests and staff time allocation. It is critical that changes in practices are based on evaluations of the quality of the assessment instruments; however, they must also be cost effective, efficiently organized and able to meet teachers' and students' needs.

To use the empirical findings and knowledge in a way that strengthens practitioners' competence to promote student achievements, it is important that improvements to the assessments to be used are linked to educational practices and policies. There is, of course, no point in developing a screening procedure that includes high-quality assessments unless there are effective interventions that schools can use to support the students at risk of the development difficulties that the assessments identify.

In addition to the above-mentioned implications for practices, the present studies derived further perspectives and suggestions for future research. First, due to the lack of information about the psychometric properties of the used assessments, a criteria list could be established and made available for schools to judge the quality of assessments to be chosen for specific purposes. Second, due to the lacks of high-quality assessments of social functioning and reading comprehension in particular, it will be of importance to develop assessments that are easy to use for screening, progress monitoring, and decision making. Moreover, it is also a need to develop social skill assessments for older students. Third, for the purpose of evaluating cut-off criteria of the ORF measure studies of sensitivity and specificity using ROC-curves will be of importance.

Fourth, the relationship between social functioning and reading proficiency using data from the ESBA and the ORF scores will be of interest to study regarding identifying the co-occurrence of difficulties in both areas. Fifth, because analyses of progress monitoring of both the ESBA and the ORF measure were not undertaken in the present research, studies with larger representative samples are awaiting. Sixth, establishing a framework to promote competences of the quality and understanding of assessments in education for schools' staff and leadership. Finally, establishing and organizing a three-tiered model for assessment and intervention, namely the response to intervention (RtI).

Taken together, the further perspectives and suggestions for future studies will be of importance for providing efficiently, equal and optimal support to students' development of valuable skills that are integral to social and academic success. It is reasonable to question educational equity, where students have equal rights to develop their talents, abilities, engagement, and motivation to learn, when invalid measures are used. Consequently, questioning whether the education system is failing to provide teachers with knowledge about assessment constructs and professional training in using them to help their students develop essential skills and behaviors is timely.

References

Algozzine, B., Wang, C., & Violette, A. (2011). Reexamining the relationship between academic achievement and social behavior. *Journal of Positive Behavior Interventions* 13, 3–16.

Al-Hendawi, M. (2012). Academic engagement of students with emotional and behavioral disorders: Existing research, issues, and future directions. *Emotional and Behavioural Difficulties*, *17*(2), 125-141.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

Antoniazzi, D., Snow, P., & Dickson-Swift, V. (2010). Teacher identification of children at risk for language impairment in the first year of school. *International journal of speech-language pathology*, *12*(3), 244-252.

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.

Arnesen, A., Meek-Hansen, W., Ottem, E., & Frost, J. (2013). Barns vansker med språk, lesing og sosial atferd i læringsmiljøet: En undersøkelse basert på lærervurderinger og leseprøver i grunnskolens 2.-5.trinn. *Psykologi i kommunen (6)*, 41-56

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'Enui, E. J., & Beck, C. T. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review*, *37*(1), 18.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, N.J: Prentice Hall.

Bandura, A. (1986). The Explanatory and Predictive Scope of Self-Efficacy Theory. *Journal of Social and Clinical Psychology*, *4*(3), 359-373. doi: 10.1521/jscp.1986.4.3.359

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations*, *61*(8), 1139-1160. doi:10.1177/0018726708094863

Beauchamp, M. H., & Anderson, V. (2010). SOCIAL: An Integrative Framework for the Development of Social Skills. *Psychological Bulletin*, *136*(1), 39-64. doi: 10.1037/a0017768

Benson, J. & Hagtvet, K. A. (1996). The interplay among design, data analysis and theory in the measurement of coping. In: M. Zeidner & N. S. Endler (Eds), *Handbook of Coping: theory, research, application*, pp. 83-106. New York: Wiley

Beswick, J., Sloat, E., & Willms, J. (2008). Four Educational Myths That Stymie Social Justice. *The Educational Forum, 72*(2), 115-128.

Biancarosa, G., & Shanley, L. (2016). What Is Fluency? In K. D. Cummings & Y. Petscher (Eds.) *The Fluency Construct: Curriculum-Based Measurement Concepts and Applications*, pp. 1-18. New York, NY: Springer New York. doi: 10.1007/978-1-4939-2803-3_6

Bishop, D., & Adams, C. (1990). A Prospective Study of the Relationship between Specific Language Impairment, Phonological Disorders and Reading Retardation. *Journal of Child Psychology and Psychiatry, 31*(7), 1027-1050. doi:10.1111/j.1469-7610.1990.tb00844.x

Bishop, D., & Snowling, M. (2004). Developmental dyslexia and specific language impairment: Same or different? *Psychological Bulletin, 130*(6), 858-886.

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 92*(1), 81-90.

Blair, C., & Raver, C. C. (2015). School Readiness and Self-Regulation: A Developmental Psychobiological Approach. *Annual Review Psychology, 66*, 711-731. doi:10.1146/annurev-psych-010814-015221

Borsa, J. C., Damásio, B. F., & Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psycholgical instruments: Some considerations. *Paidéia (Ribeirão Preto), 22,* 423–432. doi:10.1590/1982-43272253201314

Breznitz, Z. (2006). *Fluency in reading: Synchronization of processes*. Routledge.

Bronfenbrenner, U. (1979). *The Ecology of Human Development*. Harvard University Press.

Bronfenbrenner, U. (1997). Ecological models of human development. *Readings on the development of children, 1993*, 37-43.

Brown, T. 2015. *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guildford.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate behavioral research, 36*(1), 111-150.

Bru, E. (2011). Emosjonelt sårbare og sosialt passive elever. In U. V. Midthassel, E. Bru, S. K. Ertesvåg, & E. Roland (Eds.), *Sosiale og emosjonelle vansker, barnehagens og skolens møte med sårbare barn og unge*, pp. 17-36. Oslo: Universitetsforlaget.

Bruner, J. (1996): *The Culture of Education.* Harvard University Press.

Burns, M. K., Silberglitt, B., Christ, T. J., Gibbons, K. A., & Coolong-Chaffin, M. (2016). Using oral reading fluency to evaluate response to intervention and to identify students not making sufficient progress. In K. D. Cummings & Y. Petscher (Eds.) *The Fluency Construct: Curriculum-Based Measurement Concepts and Applications*, pp. 123-140. New York, NY: Springer New York. doi: 10.1007/978-1-4939-2803-3_6

Capaldi, D., DeGarmo, D., Patterson, G. R., & Forgatch, M. (2002). Contextual risk across the early life span and association with antisocial behavior. In J. B. Reid, G. R. Patterson, & J. E. Snyder (Eds.), *Antisocial behavior in children and adolescents: A developmental analysis and model for intervention*. American Psychological Association.

Caravolas, M., Lervåg, A., Defior, S., Seidlová Málková, G., & Hulme, C. (2013). Different patterns, but equivalent predictors, of growth in reading in consistent and inconsistent orthographies. *Psychological Science, 24*(8), 1398–1407. doi:10.1177/0956797612473122

Caravolas, M., Lervåg, A., Mousikou, P., Efrim, C., Litavsky, M., Onochie-Quintanilla, E., … Hulme, C. (2012). Common patterns of prediction of literacy development in different alphabetic orthographies. *Psychological Science, 23*(6), 678–686. doi:10.1177/0956797611434536

Carroll, J. M., Maughan, B., Goodman, R. & Meltzer, H. (2005). Literacy difficulties and psychiatric disorders: Evidence for comorbidity. *Journal for Child Psychology and Psychiatry*, *46* (5), pp. 524–32

Cartwright, N. (2011). *Evidence, External Validity, and Explanatory Relevance*: Oxford University Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, N.J: Lawrence Erlbaum.

Connors-Tadros, L. (2014). Definitions and approaches to measuring reading proficiency. *CEELO fast fact by CEELO (Center on Enhancing Early Learning Outcomes). Retrieved September*, *12*, 2016.

Cook, C., Heath, F., & Thompson, R. L. (2000). A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys. *Educational and Psychological Measurement, 60*(6), 821-836. doi:10.1177/00131640021970934

Cooper, B. R., Moore, J. E., Powers, C. J., Cleveland, M., & Greenberg, M. T. (2014). Patterns of Early Reading and Social Skills Associated with Academic Success in Elementary School. *Early Education and Development, 25*(8), 1248-1264. doi: 10.1080/10409289.2014.932236

Cordier, R., Speyer, R., Chen, Y., Wilkes-Gillan, S., Brown, T., Bourke-Taylor, H., Doma, K., Leicht, A. (2015). Evaluating the Psychometric Quality of Social Skills Measures: A Systematic Review. *PLoS One, 10*(7). doi: 10.1371/journal.pone.0132299

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological bulletin*, *115*(1), 74.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH, USA: Cengage Learning.

Cronbach, L. J. (1971). Test validation. In R. L.Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 443–507). Washington , DC : American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302

Crone, D. A., Carlson, S. E., Haack, M. K., Kennedy, P. C., Baker, S. K., & Fien, H. (2016). Data-Based Decision-Making Teams in Middle School: Observations and Implications from the Middle School Intervention Project. *Assessment for Effective Intervention, 41*(2), 79-93. doi: 10.1177/1534508415610322

Crowe, L. M., Beauchamp, M. H., Catroppa, C., & Anderson, V. (2011). Social function assessment tools for children and adolescents: A systematic review from 1988 to 2010. *Clinical Psychology Review, 31*(5), 767-785. doi:http://dx.doi.org/10.1016/j.cpr.2011.03.008

Cummings, K. D., Kaminski, R. A., & Merrell, K. W. (2008). Advances in the Assessment of Social Competence: Findings from a Preliminary Investigation of a General Outcome Measure for Social Behavior. *Psychology in the Schools, 45*(10), 930-946. doi: 10.1002/pits.20343

Cummings, K. D., Park, Y., & Bauer Schaper, H. A. (2013). Form effects on DIBELS Next oral reading fluency progress-monitoring passages. *Assessment for Effective Intervention, 38*(2), 91-104.

Cummings, K. D., & Smolkowski, K. (2015). Selecting Students at Risk of Academic Difficulties. *Assessment for Effective Intervention, 41*(1), 55-61. doi:10.1177/1534508415590396

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*(6), 934-945. http://dx.doi.org/10.1037/0012-1649.33.6.934

Dahle, A. E., Knivsberg, A.-M., & Andreassen, A. B. (2011). Coexisting problem behaviour in severe dyslexia. *Journal of Research in Special Educational Needs, 11*(3), 162-170. doi: 10.1111/j.1471-3802.2010.01190.x

Denham, S. A., & Brown, C. (2010). "Plays nice with others": Social–emotional learning and academic success. *Early Education and Development*, *21*(5), 652-680.

DiPerna, J. C. (2006). Academic enablers and student achievement: Implications for assessment and intervention services in the schools. *Psychology in the Schools, 43*(1), 7-17. doi:10.1002/pits.20125

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2008). "School readiness and later achievement": Correction to Duncan et al. (2007). *Developmental Psychology*, *44*(1), 232. http://dx.doi.org/10.1037/0012-1649.44.1.217

Duncan, T. E., Duncan, S. C., & Stoolmiller, M. (1994). Modeling Developmental Processes Using Latent Growth Structural Equation Methodology. *Applied Psychological Measurement, 18*(4), 343-354. doi: 10.1177/014662169401800405

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development, 82*(1), 405-432. doi: 10.1111/j.1467-8624.2010.01564.x

Ehri, L. C. (1997). Learning to read and learning to spell are one and the same, almost. *Learning to spell: Research, theory, and practice across languages*, *13*, 237-268.

Ehri, L. C. (2005). Learning to Read Words: Theory, Findings, and Issues. *Scientific Studies of Reading*, *9*(2), 167-188. doi: 10.1207/s1532799xssr0902_4

Ehri, L. C., Barron, R. W., & Feldman, J. M. (1978). *The recognition of words*: International Reading Association

Ehri, L. C., & Wilce, L. S. (1985). Movement into reading: Is the first stage of printed word learning visual or phonetic?. *Reading Research Quarterly*, 163-179.

Elliott, S. N., Huai, N., & Roach, A. T. (2007). Universal and early screening for educational difficulties: Current and future approaches. *Journal of School Psychology, 45*(2), 137-161. doi: 10.1016/j.jsp.2006.11.002

Evers, A., Hagemeister, C., & Hostmaelingen, A. (2013). EFPA Review Model for the description and evaluation of psychological and educational tests: Tech. Rep. Version 4.2. 6). Brussels: European Federation of Psychology Associations. http://www.efpa.eu/professional-development

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25*(3), 283 -291.

Floyd, R. G., Shands, E. I., Alfonso, V. C., Phillips, J. F., Autry, B. K., Mosteller, J. A., . . . Irby, S. (2015). A Systematic Review and Psychometric Evaluation of Adaptive Behavior Scales and Recommendations for Practice. *Journal of Applied School Psychology, 31*(1), 83-113. doi: 10.1080/15377903.2014.979384

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional children, 53*(3), 199-208.

Fuchs, L. S., & Fuchs, D. (2007). A model for implementing responsiveness to intervention. *Teaching exceptional children, 39*(5), 14-20.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis. *Scientific Studies of Reading, 5*(3), 239-256. doi: 10.1207/S1532799XSSR0503_3

García, J. R., & Cain, K. (2014). Decoding and Reading Comprehension. *Review of Educational Research, 84*(1), 74-111. doi: doi:10.3102/0034654313499616

García-Madruga, J. A., Vila, J. O., Gómez-Veiga, I., Duque, G., & Elosúa, M. R. (2014). Executive processes, reading comprehension and academic achievement in 3th grade primary students. *Learning and Individual Differences, 35*(0), 41-48. doi: http://dx.doi.org/10.1016/j.lindif.2014.07.013

Goffreda, C. T., & DiPerna, J. C. (2010). An empirical review of psychometric evidence for the dynamic indicators of basic early literacy skills. *School Psychology Review,*

39(3), 463-483. Retrieved from https://search.proquest.com/docview/758657815?accountid=14699

Good, R. H., & Kaminski, R. A. (2002). What are DIBELS. *Oral reading fluency passages for first through third grade (Technical Report No. 10).* Eugene, OR: University of Oregon.

Gotch, C. M., & French, B. F. (2014). A Systematic Review of Assessment Literacy Measures. *Educational Measurement: Issues and Practice, 33*(2), 14-18. doi: 10.1111/emip.12030

Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. Los Angeles: SAGE.

Gough, P., & Tunmer, W. (1986). Decoding, Reading, and Reading Disability. *Remedial and Special Education, 7*(1), 6-10.

Green, S., Higgins, J. P. T., Alderson, P., Clarke, M., Mulrow, C. D., & Oxman, A. D. (2011). What is a systematic review? In: J. P. T. Higgins & S. Green (Eds.). *Cochrane Handbook for Systematic Reviews of Interventions* (Version 5.1.0) [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org

Gresham, F. M. (2002). Teaching Social Skills to High-Risk Students. In M. R. Shinn, H. M. Walker, & G. Stoner (Eds.). *Interventions for academic and behavior problems II: Preventive and remedial approaches* (pp. 403-432). Bethesda, MD, US: National Association of School Psychologists, NASP Publications.

Gresham, F. M. (2007). Response to Intervention and Emotional and Behavioral Disorders: Best Practices in Assessment for Intervention. *Assessment for Effective Intervention, 32*(4), 214–222. doi: 10.1177/15345084070320040301

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system.* Circle Pines, MN, US: American Guidance Service.

Gresham, F. M., & Elliott, S. N. (2008). Social skills improvement system (SSIS) rating scales. *Bloomington, MN: Pearson Assessments*.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System-Rating Scales. *Psychological assessment*, *22*(1), 157.

Gresham, F. M., & Reschly, D. J. (1986). Social Skill Deficits and Low Peer Acceptance of Mainstreamed Learning Disabled Children. *Learning Disability Quarterly, 9*(1), 23-32.

Gustafsson, J.-E., Allodi Westling, M., Åkerman, A., Eriksson, C., Eriksson, L., Fischbein, S., Granlund, M., Ljungdahl, S., Ogden, T., & Persson, R. S., (2010). School, learning and mental health: A systematic review. Stockholm: The Royal Swedish Academy of Sciences, The Health Committee.

Hamre, B. K., & Pianta, R. C. (2001). Early Teacher–Child Relationships and the Trajectory of Children's School Outcomes through Eighth Grade. *Child Development, 72*(2), 625-638. doi: 10.1111/1467-8624.00301

Hart, B., & Risley, R. T. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore: Paul H. Brookes.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator, 27*(1), 4-9.

Hart, S. A., Soden, B., Johnson, W., Schatschneider, C., & Taylor, J. (2013). Expanding the environment: gene × school-level SES interaction on reading comprehension. *Journal of Child Psychology and Psychiatry, 54*(10), 1047-1055. doi:10.1111/jcpp.12083

Haug, P. (2014). The practices of dealing with children with special needs in school: a Norwegian perspective. *Emotional and Behavioural Difficulties, 19*(3), 296-310. doi:10.1080/13632752.2014.883788

Heckman, J. J. (2011). Effective child development strategies. *The Pre-K Debates: Current Controversies and Issues, Paul H. Brookes Publishing Co., Baltimore*.

Heckman, J. J. (2013). *Giving kids a fair chance*. Cambridge, Mass: The MIT Press.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics, 19*(4), 451-464. doi:http://dx.doi.org/10.1016/j.labeco.2012.05.014

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological bulletin, 111*(1), 127. doi: 10.1037/0033-2909.111.1.127

Hogan, T. P., Adlof, S. M., & Alonzo, C. (2014). On the importance of listening comprehension. *International Journal of Speech-Language Pathology, 16*(3), 199–207. http://doi.org/10.3109/17549507.2014.904441

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160. doi: 10.1007/bf00401799

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, *3*(4), 424.

Hulme, C., & Snowling, M. J. (2011). Children's Reading Comprehension Difficulties. *Current Directions in Psychological Science*, *20*(3), 139-142. doi:10.1177/0963721411408673

Jennings, J. L., & DiPrete, T. A. (2010). Teacher effects on social and behavioral skills in early elementary school. *Sociology of Education*, *83*(2), 135-159.

Johnston, T. C., & Kirby, J. R. (2006). The contribution of naming speed to the simple view of reading. *Reading and Writing*, *19*(4), 339-361.

Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early Social-Emotional Functioning and Public Health: The Relationship Between Kindergarten Social Competence and Future Wellness. *American journal of public health*, *105*(11), 2283-2290. doi:10.2105/ajph.2015.302630

Kane, M. (2006). Content-related validity evidence in test development. *Handbook of test development*, 131-153.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448.

Kautz, T., Heckman, J., Diris, R., Weel, B., & Borghans, L. (2014). *Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success*: Paris: OECD Publishing.

Kershaw, S., & Schatschneider, C. (2012). A Latent Variable Approach to the Simple View of Reading. *Reading and Writing: An Interdisciplinary Journal, 25*(2), 433-464. doi:10.1007/s11145-010-9278-3

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, *6*(2), 293-323.

Lane K. L., Barton-Arwood, S., Nelson, R., & Wehby, J. (2008). Academic performance of students with emotional and behavioral disorders served in a self-contained setting. *Journal of Behavioral Education 77,* 43–62.

Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2017). Unpicking the Developmental Relationship Between Oral Language Skills and Reading Comprehension: It's Simple, But Complex. *Child Development.* doi: 10.1111/cdev.12861

Li, F., Duncan, T. E., Duncan, S. C., & Acock, A. (2001). Latent growth modeling of longitudinal data: A finite growth mixture modeling approach. *Structural Equation Modeling*, *8*(4), 493-530.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G.(2014). On the joys of missing data. *Journal of Pediatric Psychology, 39*, 151-162.

Loeber, R., & Farrington, D. P. (2001). *Child delinquents: development, intervention, and service needs*. Thousand Oaks, Calif: Sage Publications.

Martin, R., Hooper, S., Snow, J., & Knoff, H. (1986). The assessment of child and adolescent personality.

McClelland, M. M., Morrison, F. J., & Holmes, D. L. (2000). Children at Risk for Early Academic Problems: The Role of Learning-Related Social Skills. *Early Childhood Research Quarterly, 15*(3), 307-329. doi:10.1016/S0885-2006(00)00069-7

McEvoy, A., & Welker, R. (2000). Antisocial behavior, academic failure, and school climate: A critical review. *Journal of Emotional and Behavioral disorders*, *8*(3), 130-140.

McIntosh, K., Filter, K. J., Bennett, J. L., Ryan, C., & Sugai, G. (2010). Principles of sustainable prevention: Designing scale-up of school-wide positive behavior support to promote durable systems. *Psychology in the Schools*, *47*(1), 5-21.

McIntosh, K., Horner, R., Chard, D., Boland, J., & Good, R. (2006). The Use of Reading and Behavior Screening Measures to Predict Nonresponse to School-Wide Positive Behavior Support: A Longitudinal Analysis. *School Psychology Review*, *35*(2), 275-291.

McIntosh, K., Horner, R. H., Chard, D. J., Dickey, C. R., & Braun, D. H. (2008). Reading skills and function of problem behavior in typical school settings. *The Journal of Special Education*, *42*(3), 131-147.

McKinney, J. D., Mason, J., Perkerson, K., Clifford, M., & Williams, J. (1975). Relationship between classroom behavior and academic achievement. *Journal of Educational Psychology, 67*(2), 198-203. doi: 10.1037/h0077012

Meek-Hansen, W. & Arnesen, A. (2015). *Adaption and Development of the Oral Reading Fluency Measure to Norwegian Grades 2 to 5 within a Response to Intervention Model.* Technical Report. Unpublished manuscript. Oslo: The Norwegian Center for Child Behavioral Development.

Merrell, K. W. (2009). *Behavioral, Social, and Emotional Assessment of Children and Adolscents* (3 ed.). New York, USA: Routledge Taylor & Francis Group.

Messick, S. (1993). Foundations of Validity: Meaning And Consequences In Psychological Assessment. *ETS Research Report Series, 1993*(2), i-18. doi: 10.1002/j.2333-8504.1993.tb01562.x

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Asessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8. doi: 10.1111/j.1745-3992.1995.tb00881.x

Miles, S. B., & Stipek, D. (2006). Contemporaneous and Longitudinal Associations between Social Behavior and Literacy Achievement in a Sample of Low-Income Elementary School Children. *Child Development, 77*(1), 103-117. doi:10.1111/j.1467-8624.2006.00859.x

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1. doi:10.1186/2046-4053-4-1

Monsen, M. (2013). *Store forventninger? Læreroppfatninger om eksterne leseprøver* (Doctoral thesis, Faculty of Educational Sciences, University of Oslo, Norway). https://www.duo.uio.no/bitstream/handle/10852/41446/Monsen.avh.pdf?sequence=1&isAllowed=y

Morken, F., & Helland, T. (2013). Writing in dyslexia: product and process. *Dyslexia*, *19*(3), 131-148.

Muthén, L. K. & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nulty, D. D. (2008). The Adequacy of Response Rates to Online and Paper Surveys: What Can Be Done? *Assessment & Evaluation in Higher Education, 33*(3), 301-314. doi:10.1080/02602930701293231

O' Reilly, T., & Sabatini, J. (2013). Reading for Understanding: How Performance Moderators and Scenarios Impact Assessment Design. *ETS Research Report Series, 2013*(2), i-47. doi:10.1002/j.2333-8504.2013.tb02338.x

O'Reilly, T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing Reading Comprehension Assessments for Reading Interventions: How a Theoretically Motivated Assessment Can Serve as an Outcome Measure. *Educational Psychology Review, 26*(3), 403-424. doi: 10.1007/s10648-014-9269-z

Organisation for Economic Co-operation and Development [OECD]. (2007). *PISA (Programme for International Student Assessment) 2006 Science Competencies for Tomorrow's World* (Vol. 1 & 2): OECD

Organisation for Economic Co-operation and Development [OECD]. (2013). *OECD economic surveys*: France 2013. Paris, France: Author.

Organisation for Economic Co-operation and Development [OECD]. (2015). *Skills for Social Progress: The Power of Social and Emotional Skills*. OECD Skills Studies. Paris: OECD Publishing.

Ogden, T. (2003). The Validity of Teacher Ratings of Adolescents' Social Skills. Scandinavian *Journal of Educational Research 47,* 63–76. doi:10.1080/00313830308605

Ogden, T., & Hagen, K. A. (2013). *Adolescent mental health: Prevention and intervention*. Routledge.

Olson, R. K., Keenan, J. M., Byrne, B., & Samuelsson, S. (2014). Why Do Children Differ in Their Development of Reading and Related Skills? *Scientific Studies of Reading, 18*(1), 38-54. doi:10.1080/10888438.2013.800521

Parisi, D. M., Ihlo, T., & Glover, T. A. (2014). *Screening within a multitiered early prevention model: Using assessment to inform instruction and promote students' response to intervention*: American Psychological Association.

Parker, D. C., Zaslofsky, A. F., Burns, M. K., Kanive, R., Hodgson, J., Scholin, S. E., & Klingbeil, D. A. (2015). A brief report of the diagnostic accuracy of oral reading fluency and reading inventory levels for reading failure risk among second-and third-grade students. *Reading & Writing Quarterly, 31*(1), 56-67.

Patterson, G. R. (1982). *A Social learning approach: Coercive family process* (Vol. 3). Castalia Publishing Company.

Patterson, G. R. (2016). Coercion theory: The study of change. *The Oxford handbook of coercive relationship dynamics*, 7-22.

Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (1989). *A developmental perspective on antisocial behavior* (Vol. 44): American Psychological Association.

Patterson, G. R., Reid, J., & Dishion, T. (1992). *Antisocial boys: A social interactional approach*. Oregon: Castalia Publishing Company.

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, Calif: Sage Publications.

Pennefather, J., & Smolkowski, K. (2015). Validation of the Elementary Social Behavior Assessment: A Measure of Student Prosocial School Behaviors. *Assessment for Effective Intervention, 40*, 143–154. doi:10.1177/1534508414557562

Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology, 67*(4), 461-469. http://dx.doi.org/10.1037/h0077013

Petscher, Y. (2010). A meta-analysis of the relationship between student attitudes towards reading and achievement in reading. *Journal of Research in Reading*, *33*(4), 335-355.

Pikulski, J. J., & Chard, D. J. (2005). Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher*, *58*(6), 510-519.

Popper, K. R. (1975). The Problem of the Empirical Basis. *The Logic of Scientific Discovery* (pp. 93-111): Hutchinson.

Quirk, M., Dowdy, E., Goldstein, A., & Carnazzo, K. (2017). School Readiness as a Longitudinal Predictor of Social-Emotional and Reading Performance Across the Elementary Grades. *Assessment for Effective Intervention, 0*(0), 1534508417719680. doi: doi:10.1177/1534508417719680

Rack, J. P., Hulme, C., & Snowling, M. J. (1993). Learning to Read: A Theoretical Synthesis. *Advances in Child Development and Behavior*, *24*(C), 99-132. doi: 10.1016/S0065-2407(08)60301-8

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-Based Measurement Oral Reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427-469. doi:http://dx.doi.org/10.1016/j.jsp.2009.07.001

Reitsma, P. (1983). Printed Word Learning in Beginning Readers. *Journal of Experimental Child Psychology*, *36*(2), 321-339. doi:10.1016/0022-0965(83)90036-X

Rhoades, B. L., Warren, H. K., Domitrovich, C. E., & Greenberg, M. T. (2011). Examining the link between preschool social–emotional competence and first grade academic achievement: The role of attention skills. *Early Childhood Research Quarterly*, *26*(2), 182-191.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological bulletin*, *92*(3), 726.

Rose, J. (2006) *Independent Review of the Teaching of Early Reading: Final Report*. London: Department for Education and Skills Publications

Rosenthal, R., & Rosnow, R. L. (2009). *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books*. Oxford University Press.

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36.

Russell, G., Ryder, D., Norwich, B., & Ford, T. (2015). Behavioural Difficulties That Co-occur With Specific Word Reading Difficulties: A UK Population-Based Cohort Study. *Dyslexia, 21*(2), 123-141.

Rutter, M. (1983). School effects on pupil progress: Research findings and policy implications. *Child development*, 1-29.

Rutter, M. (1989). Pathways from Childhood to Adult Life. *Journal of Child Psychology and Psychiatry, 30*(1), 23-51. doi:10.1111/j.1469-7610.1989.tb00768.x

Sabatini, J., O'Reilly, T., & Deane, P. (2013). Preliminary Reading Literacy Assessment Framework: Foundation And Rationale For Assessment And System Design. *ETS Research Report Series, 2013*(2), i-50. doi: 10.1002/j.2333-8504.2013.tb02337.x

Sattler, J. M. (1992). *Assessment of children* (Rev. and updated 3rd ed.). San Diego: J.M. Sattler.

Schaffer, H. R. (1999). *Social development*. Oxford: Blackwell.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research, 99*, 323–337. doi:10.3200/JOER.99.6.323-338

Seland, I., & Hovdhaugen, E. (2017). National Tests in Norway: An Undeclared Standard in Education? Practical and Political Implications of Norm-Referenced Standards. In S. Blömeke & J. E. Gustafsson (Eds.), *Standard Setting in Education* (pp. 161-179). Springer International Publishing.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of Program Evaluation: Theories of Practice*. Newbury Park, Calif: Sage Publications.

Skogen, J. C. & Torvik, F. A. (2013). *Atferdsforstyrrelser blant barn og unge i Norge: Beregnet forekomst og bruk av hjelpetiltak.* Rapport 2013:4. Oslo: Nasjonalt folkehelseinstitutt.

Smolkowski, K. , Cummings, K. D., & Strycker, L. (2016). An Introduction to the Statistical Evaluation of Fluency Measures with Signal Detection Theory. In K. D. Cummings & Y. Petscher (Eds.), *The Fluency Construct: Curriculum-Based Measurement Concepts and Applications.* (2016). New York, NY: Springer New York: New York, NY. doi: 10.1007/978-1-4939-2803-3_6

Snow, C. E., Burns, M. S., & Griffin, P. (1998). Preventing reading failure in young children. *Preventing reading failure in young children*.

Soden, B., Christopher, M. E., Hulslander, J., Olson, R. K., Cutting, L., Keenan, J. M., . . . Petrill, S. A. (2015). Longitudinal stability in reading comprehension is largely heritable from grades 1 to 6. *PloS one, 10*(1), e0113807. doi:10.1371/journal.pone.0113807

Spence, S. H. (1995). *Social skills training: Enhancing social competence and children and adolescents.* Windsor, UK: The NFER-NELSON Publishing Company Ltd.

Spence, S. H. (2003). Social Skills Training with Children and Young People: Theory, Evidence and Practice. *Child and Adolescent Mental Health, 8*(2), 84-96. doi:10.1111/1475-3588.00051

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading research quarterly*, 32-71.

Statistics Norway. (2016). Elever i grunnskolen. Spesialundervisning. Retrieved from http://www.ssb.no/utgrs/#

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using Curriculum-Based Measurement to Improve Student Achievement: Review of Research. *Psychology in the Schools, 42*(8), 795-819. doi: 10.1002/pits.20113

Stein, D., & Valters, C. (2012) *Understanding 'Theory of Change' in international development: A review of existing knowledge'*. JSRP and The Asia Foundation, London, JSRP Paper 1.

Stewart, R. M., Benner, G. J., Martella, R. C., & Marchand-Martella, N. E. (2007). Three-Tier Models of Reading and Behavior: A Research Review. *Journal of Positive Behavior Interventions, 9*(4), 239-253. doi:10.1177/10983007070090040601

Stipek, D. J. (2001). Pathways to constructive lives: The importance of early school success. In A. C. Bohart & D. J. Stipek (Eds.), *Constructive & destructive behavior: Implications for family, school, & society* (pp. 291-315). Washington, DC, US: American Psychological Association.

Stoolmiller, M. (1995). Using latent growth curve models to study developmental processes. In J. M. Gottman & G. Sackett (Eds.), *The analysis of change* (pp. 105-138). Hillsdale, NJ: Erlbaum.

Storch, S. A., & Whitehurst, G. J., (2002). Oral Language and Code-Related Precursors to Reading: Evidence From a Longitudinal Structural Model. *Developmental Psychology, 38*(6), 934-947. doi:10.1037/0012-1649.38.6.934

Sutherland, K. S., & Wehby, J. H. (2001). Exploring the relationship between increased opportunities to respond to academic requests and the academic and behavioral outcomes of students with EBD: A review. *Remedial and Special Education, 22*(2), 113-121.

Sutherland, K. S., McLeod, B. D., Conroy, M. A., & Cox, J. R. (2013). Measuring Implementation of Evidence-Based Programs Targeting Young Children at Risk for Emotional/Behavioral Disorders: Conceptual Issues and Recommendations. *Journal of Early Intervention, 35*(2), 129-149. doi: 10.1177/1053815113515025

Swedish Council on Health Technology Assessment [SBU]. (2014). *Dyslexi hos barn och ungdomar - tester och innsatser. En systematisk litteraurøversik. [Dyslexia in Children and Adolescence - Tests and Efforts: A Systematic Review]*. Stockholm: Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment]

Sørlie, M. A., Hagen, K. A., & Ogden, T. (2008). Social competence and antisocial behavior: Continuity and distinctiveness across early adolescence. *Journal of research on adolescence, 18*(1), 121-144. doi:10.1111/j.1532-7795.2008.00553.x

Terras, M. M., Thompson, L. C., & Minnis, H. (2009). Dyslexia and psycho-social functioning: an exploratory study of the role of self-esteem and understanding. *Dyslexia, 15*(4), 304-327. doi: 10.1002/dys.386

The Norwegian Ministry of Education. (2009). *Rett til læring. NOU 2009:18*. Oslo: Departementenes servicesenter.

The Norwegian Ministry of Education. (2011). Meld. St. 18 (2010-11). *Læring og fellesskap. Tidlig innsats og gode læringsmiljøer for barn, unge og voksne med særlige behov.* In Kunnskapsdepartementet (Ed.), (Vol. 18). Oslo.

The Norwegian Ministry of Education. (2015). *The School of the Future Renewal of subjects and competences. NOU 2015:8* (Official Norwegian Reports). Oslo: The Norwegian Ministry of Education and Research.

The Norwegian Ministry of Education. (2017). Meld. St. 21 (2016-2017). *Lærelyst – tidlig innsats og kvalitet i skolen*.

The Norwegian National Committee for Research Ethics in the Social Sciences and the Humanities. (2016). https://www.etikkom.no/globalassets/documents/english-publications/60127_fek_guidelines_nesh_digital_corr.pdf

The Research Ethics Act (2007). https://www.etikkom.no/en/library/practical-information/legal-statutes-and-guidelines/act-on-ethics-and-integrity-in-research/

Thorndike, R. M., & Thorndike-Christ, T. (2014). *Measurement and evaluation in psychology and education* (8th ed. ed.). Harlow: Pearson.

Trzesniewski, K. H., Moffitt, T. E., Caspi, A., Taylor, A., & Maughan, B. (2006). Revisiting the Association Between Reading Achievement and Antisocial Behavior: New Evidence of an Environmental Explanation From a Twin Study. *Child Development, 77*(1), 72-88. doi: 10.1111/j.1467-8624.2006.00857.x

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,*1-10.

Tunmer, W. E., & Chapman, J. W. (2012). The Simple View of Reading Redux: Vocabulary Knowledge and the Independent Components Hypothesis. *Journal of Learning Disabilities, 45*(5), 453-466. doi:10.1177/0022219411432685

Tveit, S. (2014). Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice, 21*(2), 221-237. doi:10.1080/0969594X.2013.830079

Tyler-Merrick, G., & Church, J. (2013). The Importance of Effective Behaviour Screening in the Early Years. *Emotional & Behavioural Difficulties, 18*(1), 77-87. doi:10.1080/13632752.2012.697747

Undheim, A. M., Wichstrøm, L., & Sund, A. M. (2011). Emotional and behavioral problems among school adolescents with and without reading difficulties as measured by the youth self-report: a one-year follow-up study. *Scandinavian Journal of Educational Research, 55*(3), 291-305.

van Bergen, E., van Zuijen, T., Bishop, D., & de Jong, P. F. (2017). Why Are Home Literacy Environment and Children's Reading Skills Associated? What Parental Skills Reveal. *Reading Research Quarterly, 52*(2), 147-160. doi: 10.1002/rrq.160

van Widenfelt, B. M., Treffers, P. D. A., de Beurs, E., Siebelink, B. M., & Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clinical Child and Family Psychology Review, 8,* 135–147. doi:10.1007/s10567-005-4752-1

Visser, P. S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research. In H. T. R. C. M. Judd (Ed.), *Handbook of research methods in social and personality psychology* (pp. 223-252). New York, NY, US: Cambridge University Press.

Vygotskij, L.S. (red.) (1978). *Mind in Society: Development of Higher Psychological Processes.* Cambridge, Mass.: Harvard University Press.

Walker, H. M., Colvin, G., & Ramsey, E. (1995). *Antisocial Behavior in School: Strategies and Best Practices*. CA, USA: Brooks/Cole Publishing Company

Walker, H. M., Marquez, B., Yeaton, P., Pennefather, J., Forness, S. R., & Vincent, C. G. (2015). Teacher judgment in assessing students' social behavior within a response-to-intervention framework: using what teachers know. *Education and Treatment of Children, 38*(3), 363-382.

Walker, H. M., & McConnell, S. R. (1995). *The Walker-McConnell Scale of Social Competence and School Adjustment, Elementary Version.* San Diego, CA: Singular Publishing Group.

Walker, H. M., Ramsey, E., & Gresham, F. M. (2003). Heading off disruptive behavior: How early intervention can reduce defiant behavior—and win back teaching time. *American Educator*, *26*(4), 6-45.

Walker, H. M., & Severson, H. H. (1992). *Systematic screening for behavior disorders (SSBD).* Sopris West, 1140 Boston Ave., Longmont, CO 80501.

Walker, H. M., & Sprague, J. R. (1999). The Path to School Failure, Delinquency, and Violence. *Intervention in School and Clinic, 35*(2), 67-73. doi:doi:10.1177/105345129903500201

Warnes, E. D., Sheridan, S. M., Geske, J., & Warnes, W. A. (2005). A contextual approach to the assessment of social skills: Identifying meaningful behaviors for social competence. *Psychology in the Schools, 42*(2), 173-187. doi: 10.1002/pits.20052

Weiss, C. H. (1972). *Evaluation Research: Methods for Assessing Program Effectiveness.* Englewood Cliffs N. J.: Prentice Hall.

Weiss, C.H. (1995). *Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families.* In J. Connell, A. Kubisch, L. Schorr and C. Weiss (Eds.) New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts. New York: Aspen Institute (65-92).

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation, 1997(76*), 41-55.

Welsh, M., Parke, R. D., Widaman, K., & O'Neil, R. (2001). Linkages between children's social and academic competence: A longitudinal analysis. *Journal of School Psychology*, *39*(6), 463-482.

Woolley, G. (2011). *Reading Comprehension: Assisting Children with Learning Difficulties*. Dordrecht: Springer Netherlands.

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*(6), 412-422.

Zelazo, P. H. & Müller, U. (2010). Executive function in typical and atypical development. In U. Goswami (Ed.), Handbook of childhood cognitive development (s. 445–469). Malden: Blackwell Publishing

Dissertational Papers

I

II

III

Paper I:

Arnesen, A., Braeken, J., Ogden, T., & Melby-Lervåg, M. (2017). Assessing Students' Social Functioning and Reading Proficiency: A Systematic Review of the Quality of Educational Assessment Instruments used in Norwegian Elementary Schools. *Resubmitted for publication to Scandinavian Journal of Educational Research*

I

Assessing Students' Social Functioning and Reading Proficiency: A Systematic Review of the Quality of Educational Assessment Instruments used in Norwegian Elementary Schools

Anne Arnesen,

Department of Special Needs Education, University of Oslo, Norway

Johan Braeken,

Centre for Educational Measurement (CEMO), University of Oslo, Norway

Terje Ogden,

Norwegian Centre for Child Behavior Development, Oslo, Norway

Monica Melby-Lervåg,

Department of Special Needs Education, University of Oslo, Norway

Author Note

The authors gratefully acknowledge the participating schools' staff for information to the study.

Correspondence concerning this article should be addressed to Anne Arnesen, Department of Special Needs Education, University of Oslo, P.O. Box 1140 Blindern, 0318 Oslo, Norway.

E-mail: anne.arnesen@isp.uio.no

The word count of the manuscript: 6560

Abstract

Social functioning and reading proficiency are critical for success in school and society. Therefore, identifying students with such problems is important. This study has two parts: First, a survey to a random sample of 234 elementary schools about which instruments they use to assess reading proficiency and social functioning. Second, a systematic review of the quality of these instruments used international standards for examining quality of assessment instruments. The survey showed that schools more often assess and have more instruments available for reading than social functioning. The systematic review of the assessment instruments that the schools use revealed that the instrumental quality (e.g. how they looks like) was good or adequate, yet the quality of their psychometric quality is weak or undocumented. The findings demonstrate the need for a more thorough examination of psychometric properties of assessments to ensure their reliable and valid use for decision making at school.

Key words: social functioning, reading proficiency, systematic review, educational assessment, psychometric quality

Assessing Students' Social Functioning and Reading Proficiency: A Systematic Review of the Quality of Educational Assessment Instruments used in Norwegian Elementary Schools

Every day decisions that impact students' social and academic development are based on information derived from a variety of educational assessments conducted in schools. Such decisions can influence the students' curricula, whether a student receives additional support to prevent and ameliorate difficulties or receives a referral to educational psychology services for further diagnostics and special education. Thus, assessments are important for instructional decisions that may have a great impact on students' learning and well-being.

It is reported that 15% to 20% of Norwegian students in Grades 1 to 10 are facing social emotional (i.e. anxious, conduct disorders, depression) and/or academic (i.e. reading , math) difficulties that impact their school attainment [The Norwegian Ministry of Education, 2009, 2017]. Moreover, 20% of the students have special needs for more intensive support than their peers to succeed socially and/or academically (The Norwegian Ministry of Education, 2017). Because social functioning and reading proficiency are strongly related to future life outcomes for students at risk, promoting such skills is crucial (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Gustafsson et al., 2010; OECD, 2015).

Compared with the United States and the United Kingdom, Norway began using systematic assessments in schools relatively recently. Additionally, Norway, like many other European countries, faces a disadvantage in regard to the development of educational assessment instruments because it has a small population that uses its own language. However, as noted above, schools make many important decisions – based on the assessment instruments they use – that may affect students' lives. Additionally, to prevent difficulties in students' social functioning and/ or reading, difficulties should be identified early and targeted interventions should be

implemented (Elliott, Huai, & Roach, 2007). When such an approach is followed, less intensive support is needed (Merrell, 2001).

Assessing and identifying at-risk students at an early stage requires that teachers have access to assessment instruments which not only are easy to use and quick to administer but of high quality. To examine this important issue, we present a study in which the aim was twofold. First, we examined, by survey, a random sample of Norwegian elementary schools to determine what instruments they actually use to assess students' social functioning and reading proficiency. Second, we evaluated the quality of the instruments that the schools reported using.

Social Functioning - Reading Proficiency and Their Relationship

Social functioning in a school setting is often defined as how students behave and interact with others, relying on their social skills (Beauchamp & Anderson, 2010). A number of studies show that social skills are required for the development of good social relations, emotional and academic engagement, and school motivation (Beauchamp & Anderson, 2010; Cordier et al., 2015; Gresham, 2007). Reading proficiency refers to the process of learning to decode words accurately and fluently and to comprehend the meaning of text (Hoover & Gough, 1990). Being a proficient reader and to be able to extract meaning from text is crucial to academic achievement in most theoretical school subjects (García-Madruga, Vila, Gómez-Veiga, Duque, & Elosúa, 2014; The National Assessment Governing Board, 2013). Thus, together, social functioning and reading proficiency are important for a student's well-being and academic performance (OECD, 2015; McIntosh, Reinke, Kelm, & Sadler, 2012). Mastering the skills of social functioning and reading will not only allow students to develop social and academic competence in school (Durlak & Weissberg, 2011; Stewart, Benner, Martella, & Marchand-Martella, 2007), but also prepare them for successful participation in society and the workplace (Heckman, 2000, 2011; NOU, 2015:8).

Research has also demonstrated that reading skills and social skills are highly related (Algozzine, Wang, & Violette, 2011; DeRosier & Lloyd, 2011). For instance, a study of students from low-income homes showed that relatively poor literacy achievement in Grade 1 was significantly correlated with relatively high aggressive behaviour in Grade 3 (r = -.32; p < .01) and Grade 5 (r = -.28; p = < .01) (Miles & Stipek, 2006). The study also demonstrated that prosocial behaviour in Grade 1 was significantly correlated with literacy achievement in Grades 3 and 5 (r = .24; p < .05). Furthermore, results from a longitudinal study indicated that students' academic achievement directly influenced their social functioning from Grades 1 to 2 and from Grades 2 to 3 and that students' social functioning was reciprocal related to academic achievement from Grades 2 to 3 (Welsh, Parke, Widaman, & O'Neill, 2001). Early difficulties in language and reading are risk factors for social behavioral disorders later on (Stewart et al., 2007).

The comorbidity between social behavior disorders and reading difficulties (i.e. dyslexia, poor reading comprehension) is documented in several studies (see e.g. Boada, Willcutt, & Pennington, 2012; Dahle, Knivsberg, & Andreassen, 2011; Terras, Thompson, & Minnis, 2009; Undheim, Wickstrøm, & Sund, 2011). In a study of students' social and literacy abilities in Grades 2-5, teachers reported concerns about more than 50 percent of the students who were struggling in one or both of these domains (Arnesen, Meek-Hansen, Ottem, & Frost, 2013). Thus, students who struggle in one of these domains are more likely to struggle in the other domain (Elliott et al., 2007; Rivera, Al-Otaiba, & Koorland, 2006). In summary, the literature supports the importance of early identifying students who are struggling in one or both of the two domains to promote positive development.

Methods of Assessing Social Functioning and Reading Proficiency

Differences in the constructs of social functioning and reading proficiency require different approaches in terms of assessment methods. Whereas measures of social functioning are

commonly based on informal teachers' ratings and students' self-reports, reading proficiency is measured based on summative formal tests or formative informal assessments and teacher ratings. Findings from systematic reviews show large variations in the methods that schools use to assess students' development in social skills and reading (Cordier et al., 2015; Floyd et al., 2015; Gotch & French, 2014; OECD, 2015; Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment], 2014; Standards & Testing Agency, 2015). Altogether, these reviews show that the educational assessment instruments used in schools vary with respect to a number of dimensions: (a) the level of informal or formal structure (e.g., open notes of teacher ratings versus criterion-based tests); (b) the level of interactivity (e.g., static versus dynamic); (c) whether the assessment is summative or formative (e.g., assessment of learning versus assessment for learning); (d) the assessment structure (e.g., presentation of the items to the test-taker); (e) the response formats (e.g., selected or constructed items); and (f) the item scoring (e.g., hand scoring versus computer-based scoring). Furthermore, the instruments also varied regarding the purpose of the educational assessment, as different types are used for screening, monitoring progress, and diagnosis.

Quality of Educational Assessment Instruments

In the wake of Cronbach's and Meelh's (1955) seminal paper on the validity of psychological assessment instruments, researchers have developed a number of systems and criteria for judging the validity of a measurement. Some criteria seem to be agreed upon and are considered critical to the quality of an instrument, independent of whether the assessment's purpose is summative or formative.

Validity refers to whether an instrument has systematic measurement error. There are different types of validity, but those most commonly used in evaluations of educational assessment instruments (see, e.g., Evers, Hagemeister, & Hostmaelingen, 2013; Evers, Muñiz, et

al., 2013) are *construct validity* (whether the items represent the theoretical constructs that they are designed for), *criterion-related validity* relating to concurrent and predictive validity (whether the assessment instruments correlate to other relevant valid instruments used for the same purpose to predict future or current performance), and *content validity* in terms of face validity and logical validity (whether the items are representative and are an accurate assessment covering the broad range of variation within students' social skills and reading skills). Note that criterion validity also concerns how the cut-off points (specificity and sensitivity) and norms are developed, and how the norming sample represents the population the instrument is designed to assess in regard to age, socio-economic background, gender, language background, and other important characteristics (Thorndike & Thorndike-Christ, 2014).

Furthermore, reliability refers to the extent to which an instrument produces random measurement errors. Reliability is crucial if an assessment is to be useful, as a test that is not reliable can never be a valid instrument (Thorndike & Thorndike-Christ, 2014). Reliability is commonly assessed in terms of *internal consistency reliability* (the degree to which different test items that probe the same construct produce similar results), *test retest reliability* (stability over time), and *inter-rater reliability* (the degree to which different observers or raters agree).

Recent systematic reviews evaluating the quality of educational assessment instruments have demonstrated the lack of studies of psychometric properties for many of the instruments used in schools (see for instance Cordier et al., 2015; Floyd et al., 2015; Gotch & French, 2014; Siddiq, Hatlevik, Olsen, Throndsen, & Scherer, 2016; Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment], 2014). Although several of the reviewed measures demonstrated good psychometric qualities, there were still many that showed weak or lacking evidence. In a review of thirteen measures of social skills, Cordier and colleagues (2015) found excellent reliability scores overall, but none of the measures were found

to report validity. Additionally, Floyd and colleagues (2015) demonstrated that the fourteen behaviour scales they reviewed had mostly adequate or inadequate norming data; a mix of adequate, inadequate or not-reported reliability; and, finally, inadequate validity overall. Furthermore, Gotch & French, (2014) found weak psychometric evidence for the 36 educational literacy measures they reviewed. Siddiq and colleagues (2016) reviewed 38 educational assessment instruments that aim to measure students' literacy in information and communication technology; they found that the documentation and reporting of test quality were lacking overall. In Norway, however, there has been (to our knowledge) only one previous systematic review of educational assessments. It reviewed the quality of eight language assessment instruments used in kindergarten and found that none met the required criteria (Kunnskapsdepartmentet [The Norwegian Ministry of Education], 2011b). Moreover, a Swedish review of assessments of reading and literacy measures found that evidence was lacking overall in more than 50 of the reviewed tests (Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment], 2014).

The Current Study

Social functioning and reading proficiency provide the foundation for both academic performance and social well-being. Having and using valid instruments in schools to identify students who are struggling in one or both of these domains is vital to lead instructions to provide them with adequate support. Despite ongoing discussions of educational policy, principles and practice with regard to the assessment of students' learning in Norwegian schools (Kunnskapsdepartmentet [The Norwegian Ministry of Education], 2011b, 2017), there are no systematic studies of the quality and use of educational assessment instruments for social functioning and reading proficiency. Therefore, we investigated the following research questions:

(a) To what extent do Norwegian elementary schools use educational assessment instruments targeting students' social functioning and reading proficiency, and to what extent do schools use these to lead instructions and interventions?

(b) What characterizes the quality of the educational assessment instruments used to measure students' social functioning and reading proficiency in Norwegian elementary schools in terms of descriptions and documented psychometric properties?

Method

The current study has two parts: First, a survey was conducted in a random sample of approximately 15% of Norwegian elementary schools to provide an overview of all current assessment instruments that are used and the extent to which they are used to make decisions about interventions. Second, based on the survey, we conducted a systematic literature review of documentation of the quality of the assessment instruments (i.e. validity studies published either in the test materials/manuals for the instruments or in research articles/reports) that met the inclusion criteria for the study (see Figure 1). Then, we used the European Federation of Psychologists' Associations (EFPA) "Review Model for the Description and Evaluation of Psychological and Educational Tests" (Evers et al., 2013; Evers, Muñiz, et al., 2013 [http://www.efpa.eu/professional-development]) to examine the instrumental and psychometric qualities of the identified documented instruments.

Part 1: Survey

A random sample of 410 elementary schools across Norway was invited (by email) to complete an electronic questionnaire about the assessment instruments used to measure students' proficiency in social functioning and reading. A total of 234 (57%) of the invited schools completed the questionnaire in the spring of 2015. The schools were located in both urban and rural districts across Norway, covered all regions of the country, and enrolled students from a

variety of socio–economic backgrounds. The schools that did not respond to the survey came from the same random sample of municipalities as those that did respond. Additionally, some of the non-respondent schools replied that they not could find time to complete the survey, while others replied that they had nothing to report other than their use of national compulsory assessments and national tests.

Part 2: Systematic Literature Review and Quality Evaluation of Assessment Instruments

Based on the results of the survey, 4 of the social functioning assessment instruments and 28 of the reading assessment instruments that were reported as being used in the schools (see Figures 2 and 3) were identified for inclusion in the systematic literature review (see Table 1 and Figure 1). Notably, the schools also reported use of several instruments that were intervention material rather than assessments, developed by the teachers for informal classroom use, clinical instruments to be used only by certified educational psychologists, and group-based reports on the school environment. Therefore, as shown in Figure 1, materials reported by the schools were for reasons excluded from the quality evaluation (EFPA review).

[Table 1 near here]

[Figure 1 near here]

The purpose of the systematic literature search was to identify publications of the above mentioned instruments that were to be included in the EFPA review. Typically, this information was reported in the assessment materials (manuals, information materials). However, it was also possible that there are validation studies published as research reports/articles in addition to these materials. Therefore, to supplement the information, we did an additional systematic literature search using the Prisma guidelines (Moher, Liberati, Tetzlaff, Altman, & The PRISMA group, 2009).

Systematic literature search procedures. We set up an extensive search strategy that combined keywords with all relevant synonyms and alternative expressions widely used in the literature for each domain (i.e., social functioning and reading proficiency). Figure 1 shows details of the search and the flow of records of documentation on the assessment instruments and the eligibility criteria used for our study.

In addition to searching for each label or acronym of the instruments listed by the schools (see Figures 2 and 3), we identified keywords to search for documentation of the two types of instruments targeting social functioning and reading. We identified the following keywords based on the terms *Educational assessment instruments, Social functioning* and *Reading proficiency* as defined in the introduction: *Social; Reading; Assessment; Psychometric; Elementary school;* and *At risk.* The search keywords, with accompanying synonyms and alternative expressions for each of the two types of assessment instruments, are listed in Appendices A and B. The OR operator was used between synonyms and the alternative expressions for each keyword, and the AND operator was used between the different keywords. The truncation function * was used to capture different forms of the search words (for instance, assessment vs. assessments or assessing was truncated to assess*; measurement vs. measurements or measuring or measure or measures was truncated to measure*).

The search was conducted in two waves during the period from March 3rd 2016 to June 30th 2016. One search was conducted for the social functioning instruments and one for the reading instruments. To avoid limiting our hits of studies or our documentation of the assessment instruments, we did not restrict the search to any starting point. Furthermore, the literature included materials written in English, Norwegian, Danish, Swedish and Finnish. Developers and researchers in the field of national assessments and tests initiated by the Norwegian Directorate

for Education and Training were contacted by email and asked to share studies or technical reports that we might have missed and that pertained to the instruments.

As shown in Figure 1 and Table 2, 3 of the 4 social functioning instruments and 24 of the 28 reading instruments met the inclusion criteria for the EFPA review. The 3 social functioning instruments appeared in 3 records which included published test materials (manuals, information materials) derived from the publishers and/or authors. The 24 included reading assessments appeared in a total of 57 records in which publications of both test materials (manuals, information materials) and studies were included.

[Table 2 near here]

EFPA review model for the description and evaluation of the assessment instruments. We used the EFPA review model to evaluate the quality of the assessment instruments that the schools reported to use and that met the inclusion criteria described in Figure 1 (Evers et al., 2013; Evers, Muñiz, et al., 2013; PsykTestBarn, 2016). The review model has two parts: One part for the description of the instrument and one part for the evaluation of the instrument. The description consists of the following elements: (a) General description (e.g., instrument name, authors, publisher, date of publication); (b) Classification (e.g., content domains, populations, scales and variables measured, response mode, demands on the test taker, item formats, intended mode of use, administration mode, time required for administering); (c) Measurement and scoring (e.g., scoring procedure, scales used, transformation for standard scores; (d) Computer generated reports (e.g., availability, media, complexity, structure, sensitivity to context, modifiability, transparency, style and tone, intended recipients); and (e) Conditions and costs (e.g., documentation, methods of publication, start-up and recurrent costs, prices for reports, test-related and professional qualifications required for use of the instrument).

The evaluation part of the EFPA review form consists of the following elements to be reviewed: (a) Quality of explanation (e.g., rationale, adequacy of documentation, procedural instructions); (b) Quality of the test materials used (e.g., paper-and-pencil tests, computer bases and web-based tests); (c) Norms (e.g., norm-referenced interpretation, criterion-referenced interpretation); (d) Reliability (e.g., data provided, internal consistency, test – retest, equivalence in terms of parallel or alternative forms, Item Response Theory-based method (IRT), Inter-rater reliability); (e) Validity (e.g., construct validity, criterion-related validity, overall adequacy); (f) Quality of computer-generated reports (e.g., scope or coverage, reliability, relevance or validity, fairness, acceptability, length, overall adequacy); and (g) Final evaluation (e.g., conclusions, recommendations). All reviewed elements in the evaluation part of the form use a rating system with scores of 0 (not possible to rate or insufficient information provided), 1 (inadequate), 2 (adequate), 3 (good), or 4 (excellent). Additionally, 9 (not applicable) is used but not for the reliability and validity elements.

EFPA reviewing procedure. One of the authors completed the EFPA review form for all the included assessment instruments. As a verification check to ensure consistency and quality of the review, half of the documented instruments were randomly selected for reviewing by an additional reviewer. The additional reviews were distributed half-half among two of the other authors who evaluated the assessment instruments independently of each other and of the main reviewer. The National Assessments and Test of Reading Proficiency (NTRP) were chosen as benchmark reviewed by all three reviewers. Initial and follow-up meetings between the three authors responsible for completing the EFPA review form were arranged to discuss the review criteria. An inter-rater variance component analysis to highlight potential review disagreements was used to inform a final meeting that was organized to establish the final consensus evaluation of all assessment instruments shown in Table 7.

The variance component analysis of the overall indicator ratings on the six evaluation elements was conducted to better understand the sources driving differences in ratings and disagreements among the three raters. The average score across overall indicator ratings on the six evaluation elements was 1.20 with a standard deviation of 1.21, with 79% of the ratings being below or equal to 2 (adequate). The biggest source of rating differences was accounted for by the main effect of the evaluated tests (36%) indicating a relatively large variation in quality among the reviewed tests. The second biggest source of rating differences was accounted for by the main effect of the evaluation elements (27%), with higher ratings for the quality of the material (average rating = 2.29) and lowest ratings for both reliability and validity of the tests (average rating = .56 & .34, respectively). The test-by-element interaction accounts for only 7% of the rating variation which implies that tests tend to be rated rather at a homogeneous quality level across the six evaluation elements (i.e., if it is relatively bad, it tends to be relatively bad in every aspect).

Results

The results are reported in two sections: 1) The results of the survey concerning the elementary schools' use of educational assessment instruments for social functioning and reading proficiency, and 2) the EFPA review of the instruments' characteristics, and their instrumental and psychometric quality.

Use of Social Functioning and Reading Proficiency Assessment Instruments

The survey showed that the schools used 21 different social functioning assessment instruments and 36 reading assessment instruments. Figures 2 and 3 show the percentage of Norwegian elementary schools (n=234) that reported their use of the different instruments. Because the use of national compulsory assessments in reading is required in all Norwegian

elementary schools, we did not include them in Figure 3. We did, however, include them in the review of the descriptions and documented psychometric properties (Research Question 2).

[Figure 2 and Figure 3 near here]

Figure 2 shows that the most frequently used assessments of social skills are those described as "Teacher-made", followed by "Olweus' Students' Self-report on Bullying" and "Students' Survey of Self-assessment of Learning and Well-being". Among these, only the informal "Teacher-made" ratings are, as reported by the schools, intended to target social functioning. Regarding the reading assessment instruments (see Figure 3), the most frequently used is "Carlsten", a group-administered reading test without any reported psychometric properties (see Tables 6 and 7). Notably, except for the national compulsory assessments and test, the majority of the reading assessments focused more on decoding skills than comprehension skills.

Table 3 shows the total number of schools that reported using educational assessment instruments for students' social functioning and reading proficiency. Notably, as many as 68.8% of the schools did not use any instruments to assess students' social functioning, but only 11.1% did not use any instruments to assess reading. Additionally, 9.4% of the schools reported that they did not use any assessment instruments for students' social functioning or reading proficiency (except the national compulsory assessments in reading for Grades 1 to 3 and the national reading test for Grade 5).

[Table 3 near here]

The majority of schools reported that they assessed students' social functioning and reading proficiency either two or more than three times per year (see Table 4). Furthermore, the schools reported whether they used the information derived from the assessments when making decisions to further promote students' social and reading skills.

[Table 4 near here]

Table 5 shows that as many as 91.5% of the schools used the results derived from assessments of students' social functioning. However, as seen in Table 3, only 31.2% reported that they used any assessment instrument for social functioning. Additionally, there is a discrepancy between the percentage of schools (88.9%) that reported using reading assessment instruments (see Table 3) and the percentage of schools (98.7%) that used information derived from the results of the assessments to make decisions about reading instruction (see Table 5).

[Table 5 near here]

In summary, the most frequently used measure was "Teacher-made" for social functioning and the "Carlsten" for reading (when the national mandatory tests are excluded). Notably, the "Teacher-made" and "Carlsten" measures had no documented psychometric properties. Additionally, the findings demonstrated that a lower percentage of the schools reported to use assessment instruments (31.2% assessed social function and 88.9% assessed reading proficiency) than to use the results derived from these assessments to promote the development of students' skills (91.5% used information on social functioning and 98.7% used the results from reading assessments). Furthermore, the descriptive data analyses did not find any relations between the schools' use of the assessment instruments and their use of information derived from assessing students when making decisions about interventions to promote either social skills or reading skills.

Evaluation of Instruments' Characteristics, and Instrumental and Psychometric Quality

The documentation of the instruments included in the EFPA review consisted of manuals, articles and master's theses (see Table 2). We were able to use the EFPA review model to assess descriptions of the characteristics and to evaluate the quality and documented psychometric

properties of 3 of the social functioning instruments and 24 of the reading instruments that were reported used in Norwegian elementary schools.

Descriptions of the assessment instruments' characteristics. Table 6 shows the descriptions of characteristics for the assessment instruments of social functioning and reading proficiency. All reviewed instruments contained some descriptions of their purpose and target group. One of the social functioning instruments was published 17 years ago (1999), whereas the evaluated versions of two others were published within the last year. The reading instruments were published between 1980 and 2016. The two types of instruments were either individually or group administered and took three different forms: Teacher ratings, students' self-reports and performance assessments or tests. None of the social functioning instruments, and only 7 of the reading instruments, were defined as screening instruments. Two social functioning instruments (Student survey and Students' social competence) and one reading instrument (National reading tests of proficiency: NTRP) were distributed by the Norwegian Directorate of Education. The Student survey is compulsory for Grade 7, and the NTRP are compulsory for Grades 1 to 3 and Grade 5 in all Norwegian schools. These have to be completed annually.

[Table 6 near here]

Regarding the response mode, the majority reported the use of paper-pencil as an option either similar to or in addition to direct observation or/and computer-based assessment. Three measures reported the use of computer-based assessment as the only response mode, and one used both direct observation and computer-based options. The item formats of the social functioning measures were Likert scales, open questions and oral interviews. The majority of the reading measures reported multiple choice (MC) tasks, a similar item format, or MC in addition to Likert scales, open questions and/or dictation; and 3 of the MC were administered with a time-limit. Two reading instruments (LUS and SOL) reported the use of a teacher's observation form

and one (Leselos) used a teacher's check-form. The majority of scorings were raw-scores based on the number of dichotomous responses (right or wrong, yes or no). In addition, two of the social functioning measures and ten of the reading measures reported the use and interpretation of teachers' observational notes. Furthermore, six reading instruments reported the use of cut-off scores, whereas three instruments had norms for the cut-off scores based on raw-scores or z-scores. Additionally, seven reading measures used normed scores based on raw-scores, z-scores, or stanines. Ten of the reading assessment instruments reported the time required to administer them, whereas none of the social functioning assessments reported this. Moreover, seven of the reading instruments required test-related qualifications.

Finally, none of the 3 social functioning instruments reported any validation studies, while ten of the 24 reading instruments did report such studies. Five of these studies were presented in a single publication, typically the manual, whereas five instruments were reported in two or more publications. The sample sizes ranged between 81 (Skaathun, 2013) and 55,703 students (Utdanningsdirektoratet [Norwegian Directorate for Education and Training], 2015). Two measures reported on studies conducted in other countries in addition to those conducted with Norwegian samples (Nielsen et al., 2008; Sutherland & Smith, 1991).

Instrumental and psychometric quality. Information from the EFPA review of the assessment materials and documented psychometric properties is shown in Table 7. The detailed evaluation criteria are described in the EFPA manual (see Evers et al., 2013; Evers, Muñiz, et al., 2013). First, we judged the quality of explanation (e.g., rationale, adequacy of documentation, procedural instructions). We found that one of the social functioning measures (Student's Self-report) was adequately explained, whereas the other two had no information that we could rate. Regarding the explanations contained in the reading instrument materials, one was rated as excellent (Language 6-16), eight were rated as good, seven were rated as adequate, nine had

inadequate explanations, one had no information available to rate, and one had no applicable explanation.

[Table 7 near here]

Second, in rating the quality of the applicable paper-and-pencil test materials, one measure was rated as excellent (Language 6-16), 12 were rated as good, seven were adequate, and four were inadequate. Regarding the applicable computer based or web-based materials, two met the criteria for excellence (KOAS and LOGOS), three were rated as good (Student Survey, kartleggeren.no and SOL), and two instruments had no information available for rating (Aski Raski and LUS). The quality of computer-generated reports (e.g., scope or coverage, reliability, relevance or validity, fairness, acceptability, length, overall adequacy) for the one social functioning measure and the three reading measures was rated as good for the Student Survey and LOGOS, adequate for kartleggeren.no and KOAS, inadequate for Aski Raski, and for LUS and SOL, there was no information that could be rated.

Finally, we reviewed the instruments' documented psychometric properties in terms of norms (e.g., norm-referenced interpretation, criterion-referenced interpretation), reliability (e.g., data provided, internal consistency, test – retest, equivalence in terms of parallel or alternative forms, item response theory-based method (IRT), inter-rater reliability), and validity (e.g., construct validity, content validity, criterion-related validity, overall adequacy). Ten of the 27 instruments had no applicable norms, nine had no information that could be rated, one had good norms (Language 6-16), three were adequate, and four were inadequate. The majority of the instruments (none of the social functioning and 18 of the 24 reading instruments) had no documented information on their reliability. The overall adequacy of the reported reliability of the six reading measures was rated as good for three (Lesesenterets Spelling Test, Word Chain Test, and Language 6-16), adequate for two and inadequate for one. Regarding the overall

adequacy of validity, only five instruments had any documentation available to rate. Two of these had good validity (Word Chain Test and Language 6-16), while one was adequate and two were rated as inadequate.

In summary, even though most of the instrumental quality (e.g. how the paper-pencil, computer- or web-based materials look like) and their explained rationale was good or adequate, there was no evidence to support the overall quality of the psychometric quality for the majority of the reviewed assessment materials. Although most of the instruments may be of practical use for experienced teachers, further development and research are required to ensure their quality for use in practice. Additionally, it is noteworthy that in the instruments provided as qualitative or dynamic observation assessments (e.g., SOL, The Working Test), the authors and developers explain that reliability and validity are irrelevant.

Discussion

Our study reveals important information about the use of educational assessments in Norwegian schools and the quality of the assessment instruments used; this information is relevant to both future research and policy development. First, our study demonstrates that Norwegian elementary schools typically assessed students' reading proficiency more often than they assessed social functioning, and they used a wider variety of reading assessment instruments than of social functioning instruments. This large difference in use may not only reflect very different traditions of assessment in the domains of social and academic achievements but also the recency of interest in measuring social skills as compared to reading skills. The most frequently used measures in the two domains, with the exceptions of national compulsory tests and assessments, were teacher-made social functioning assessments and the "Carlsten" reading test, neither of which has documented psychometric properties. The majority of schools used information derived from assessments of students' skills to make decisions about interventions,

but this seems to be based more on informal teacher ratings than on the educational assessment instruments that the schools report using.

Our evaluation of the quality of the instruments revealed that the vast majority of the reviewed applicable materials had good descriptions of their purpose and content. However, our findings regarding the explanation of the rationale, the presentation and the information provided, and the evidence of psychometric properties demonstrated, with very few exceptions, an overall weakness and lack of applicable quality (see Table 7). This is highly troublesome, and addressing these problems should be a priority in both educational research and policy in the years to come.

Similar to the findings of previous studies (e.g., Cordier et al., 2015; Merell, 2001; Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment], 2014), our findings demonstrate that the documented educational assessment instruments that were reported as being used in schools to assess students' social functioning and reading proficiency varied with respect to the informal or formal structure of their assessment methods (e.g., open notes of teacher ratings versus criterion-based tests), the level of interactivity (e.g., static versus dynamic), whether they were summative or formative (e.g., assessment of learning versus assessment for learning), the assessment structure (e.g., presentation of the items to the test-taker), the response formats (e.g., selected or constructed items), and the item scoring (e.g., hand scoring versus computer-based scoring). However, our view is that assessments and interventions are often intertwined and that schools use what is readily available. Considering the lower percentage of schools reporting that they used educational assessment instruments compared to the percentage of schools reporting the use of information derived from the results of such assessments, the answer to the first research question seems somewhat inconsistent. This may be interpreted as indicating that schools are using information derived from informal assessments of students' social functioning and reading proficiency more than they are using

formal assessment instruments. Moreover, we do not know from these answers whether and how schools actually use the assessments for their real purpose. Our understanding is that schools have extensively used teacher-made assessments and informal classroom observations to assess and design interventions to promote students' skills. Furthermore, this may be because the available educational assessments are time-consuming and not efficient for teachers to use in practice (Elliot et al., 2007; Kunnskapsdepartementet [The Norwegian Ministry of Education], 2011a). However, arbitrary and inaccurate measures may misinterpret students' needs for additional support to prevent or minimize difficulties, and such measures may also initiate interventions that do not meet students' needs.

Although the instrumental quality (e.g., paper-pencil test-materials) and the explanation of the instruments' rationale is rated as good or adequate for about half of the reviewed measurement materials, our findings regarding the quality of the documented psychometric properties contrast this judgement. Thus, our findings support the conclusions of other systematic reviews of assessment measures, namely that the psychometric evidence is weak or lacking (e.g., Cordier et al., 2014; Floyd et al., 2015; Gotch & French, 2014; Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment], 2014). That said, we did not analyse the measurements' original data but rather reviewed the information and the documented materials of the instruments provided to us by the authors and publishers and obtained through the literature search. Knowing that early identification and support of students struggling with social functioning and reading are crucial, it is troubling that the majority of the reported instruments are lacking or have inadequate documentation. Consequently, when using assessment instruments without any documented properties, the related inferences and judgements can be erroneous and lead to negative consequences for students at risk. Without precise information to

select and design accurate support, students are likely to attend school for years without reaching their full potential.

Limitations and implications of this study for research and policy

Two features of this study limit the conclusions we can draw about schools' use of educational assessment instruments that measure social functioning and reading proficiency. First, the response rate (57%) to the electronic survey sent to schools regarding their use of educational assessments might limit the representativeness of the sample. However, it is expected that surveys to organizations (e.g., schools) have a lower response rate than data collected from individuals (Baruch & Holtom, 2008). Also, electronic data collection efforts resulted in response rates as high as or higher than traditional methods of data collection (e.g., mail and phone call interviews). In fact, there is no scientifically proven minimally acceptable response rate. However, a response rate of 60% has been used as a "rule of thumb" (Johnson & Wislar, 2012). Based on the above presented research, the response rate covered 234 elementary schools, is assumed to be an acceptable sample of a representative group of Norwegian elementary schools. Although the response rate was not optimal, the responses offer an interesting picture of the variance in the schools' use of assessment instruments. Additionally, it might be assumed that the non-respondent schools do not use any other assessment instruments beyond those used by the schools that responded. Second, due to the response rate, we do not know if there are other assessment instruments that are used in the schools that did not participate in this study. However, the initial search of the systematic literature review did not obtain any hits of available measurements, except of those not included in the current study.

Moreover, the question of schools' use of educational assessments of social functioning and reading proficiency is limited to the extent to which such assessments are used, what they use and how these are used, more than detailed information whether the assessments are really used

for their real purpose. This said, the study has focused on a general overview more than differences between assessments in each of the two areas and therefore not touched upon whether the assessments are used as intended. Thus, we have no guaranties of how the assessments are used in practice.

Several implications arise from the findings of this study. Assessing students' social functioning and reading proficiency in an educational context is complex. It requires measurements that use differentiated methods and address students' growth in both social and academic learning. Measurements must also provide reliable and valid information about what they intend to measure. To improve the practice of screening and monitoring students' learning and development in the domains of social functioning and reading, there is a need for more differentiated assessment instruments that are easy and efficient for teachers to use. Additionally, social functioning and reading proficiency should be assessed at the same time within and across school years to guide instruction, monitor students' responses to instruction, and monitor students' progress in the two domains to obtain a basis for analysing variations and mutual causal influences. Also, because we do not know whether schools are using assessment instruments for decision makings to lead instructions and how the assessments are really used as they are intended to do, more research in this field is needed. Given the present findings regarding assessments of students' social functioning and reading proficiency, an extended and explicit evaluation of the qualities of educational assessment instruments in terms of constructs, materials and psychometrics should be prioritized in educational policy and practice. This could ensure more appropriate use of teachers' time and efforts to early identify struggling students and provide them with the less intensive support that is effective when implemented early.

References

Algozzine, B., Wang, C., & Violette, A. S. (2011). Reexamining the Relationship Between Academic Achievement and Social Behavior. *Journal of Positive Behavior Interventions, 13*(1), 3-16. doi: 10.1177/1098300709359084

*Allard, B., Rudqvist, M., Sundblad, B., Corneliussen, G. G., Smeland, O. I., & Moen, S. (2006). *Den nye LUS-boken : Leseutviklingsskjema - LUS : en bok om leseutvikling*. Oslo: Cappelen akademisk forlag.

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. doi: 10.1016/j.jsp.2012.09.004

Arnesen, A., Meek-Hansen, W., Ottem, E., & Frost, J. (2013). Barns vansker med språk, lesing og sosial atferd i læringsmiljøet: En undersøkelse basert på lærervurderinger og leseprøver i grunnskolens 2.-5.trinn. *Psykologi i kommunen*(6), 41-56

*Asbjørnsen, A. E. , Obrzut, J.E. , Eikeland, O.-J., & Manger, T. (2010). Can Solving of Wordchains be explained by Phonological Skills alone? *Dyslexia, 16*:24-35

*Aschim, A. K. (2006). *Damms leseunivers 1 Ressursperm*. Oslo: Cappelen Damm

*Ask, I. (2016, August, 30). *Aski Raski*. [Software and WebApplication]. Unpublished instrument. Retrieved from http://www.askiraski.no/index.cfm

Association for Educational Assessment – Europe. (2016, November, 10). *European Framework of Standards for Educational Assessment* 1.0. Retrieved from http://www.aea-europe.net/index.php/professional-development/standards-for-educational-assessment

* Publications included in the evaluation review are marked with an asterisk (*).

Baruch, Y., & Holtom, B. (2008). Survey response rate levels and trends in organizational research. *Human Relations, 61*(8), 1139.

Beauchamp, M. H., & Anderson, V. (2010). SOCIAL: An Integrative Framework for the Development of Social Skills. *Psychological Bulletin, 2010, Vol. 136*(1), 39-64.

Boada, R., Willcutt, E. G., & Pennington, B. F. (2012). Understanding the comorbidity between dyslexia and attention-deficit/hyperactivity disorder. *Topics in Language Disorders, 32*(3), 264-284.

*Carlsten, C. T. (2016). *Carlstenprøvene*. Oslo: Cappelen Damm

Cordier, R., Speyer, R., Chen, Y., Wilkes-Gillan, S., Brown, T., Bourke-Taylor, H., . . . Leicht, A. (2015). Evaluating the Psychometric Quality of Social Skills Measures: A Systematic Review. *PLoS One, 10*(7). doi: 10.1371/journal.pone.0132299

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*, 281-302

Dahle, A. E., Knivsberg, A.-M., & Andreassen, A. B. (2011). Coexisting problem behaviour in severe dyslexia. *Journal of Research in Special Educational Needs, 11*(3), 162-170. doi: 10.1111/j.1471-3802.2010.01190.x

DeRosier, M. E., & Lloyd, S. W. (2011). The Impact of Students' Social Adjustment on Academic Outcomes. *Reading & Writing Quarterly, 27*, 25-22), p.25-47. doi: 10.1080/10573569.2011.532710

*Duna, K. E., & Frost, J. (1999). *Elevens selvrapport: Systematisk kartlegging av elevens subjektive forståelse av egen livssituasjon*. Jaren: PP-tjenestens materiellservice. http://www.aspergerbedriftene.no/materiellservice/butikk/elevens-selvrapport/1-laererveiledning-20-noteringshefter-56-kort-3-esker-m-m/

*Duna, K. E., Frost, J., Godøy, O., & Monsrud, M.-B. (2003). *Kartlegging av barn og unges lese- og skrivevansker med Arbeidsprøven.* Oslo: Bredtvet kompetansesenter

Durlak, J. A., & Weissberg, R. P. (2011). Promoting social and emotional development is an essential part of students' education. *Human Development, 54*(1), 1-3. doi: 10.1159/000324337

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development, 82*(1), 405-432. doi: 10.1111/j.1467-8624.2010.01564.x

Elliott, S. N., Huai, N., & Roach, A. T. (2007). Universal and early screening for educational difficulties: Current and future approaches. *Journal of School Psychology, 45*(2), 137-161. doi: 10.1016/j.jsp.2006.11.002

*Engen, L., & Helgevold, L. (2012). *Leselos: Veiledningshefte*. Retrieved from www.lesesenteret.no Stavanger: Lesesenteret, Universitetet i Stavanger. ISBN: 978-82-7649-071-8

*Evensen, K. B. (2011). *Forebygging av lesevansker: En beskrivelse av IL-basis sitt potensial til å predikere barns leseferdighet i 2. klasse* (Master's Thesis). Available from http://hdl.handle.net/10852/31428

Evers, A., Hagemeister, C., & Hostmaelingen, A. (2013). *EFPA Review Model for the description and evaluation of psychological and educational tests*. Tech. Rep. Version 4.2. 6). Brussels: European Federation of Psychology Associations.

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25*(3), 283 -291.

*Fagbokforlaget. (2016). *Kartleggeren.no*: [Computer software]. Unpublished material. Retrieved from http://kartleggeren.no/

Floyd, R. G., Shands, E. I., Alfonso, V. C., Phillips, J. F., Autry, B. K., Mosteller, J. A., . . . Irby, S. (2015). A Systematic Review and Psychometric Evaluation of Adaptive Behavior Scales and Recommendations for Practice. *Journal of Applied School Psychology, 31*(1), 83-113. doi: 10.1080/15377903.2014.979384

*Frost, J., & Nielsen, J.C. (2000). *IL-basis – et prøvemateriell for å beskrive og vurdere barns leseforutsetninger og tidlige leseutvikling*. Norsk psykologforening

*Gallefoss, B. S. (1996). *ASTON INDEX, en test for observasjon og vudering av lese/ skrive/språkvansker : en analyse av testresultat, med hovedvekt på utføringsdelen (54 elever, pluss 6 voksne)* (Master's Thesis). Oslo: University of Oslo

García-Madruga, J. A., Vila, J. O., Gómez-Veiga, I., Duque, G., & Elosúa, M. R. (2014). Executive processes, reading comprehension and academic achievement in 3th grade primary students. *Learning and Individual Differences, 35*(0), 41-48. doi: http://dx.doi.org/10.1016/j.lindif.2014.07.013

*Gjesdal kommune. (2011). *SOL: Systematisk observasjon av lesing*. [Computer software]. Available from http://www.sol-lesing.no/

Gotch, C. M., & French, B. F. (2014). A Systematic Review of Assessment Literacy Measures. *Educational Measurement: Issues and Practice, 33*(2), 14-18. doi: 10.1111/emip.12030

Gresham, F. (2007). Response to Intervention and Emotional and Behavioral Disorders: Best Practices in Assessment for Intervention. *Assessment for Effective Intervention, 32*(4), 214-222. doi: 10.1177/15345084070320040301

Gustafsson, J.-E., Allodi Westling, M., Åkerman, A., Eriksson, C., Eriksson, L., Fischbein, S., Granlund, M., Gustafsson, P., Ljungdahl, S., & Ogden, T. (2010). School, learning and

mental health: A systematic review. Stockholm: The Royal Swedish Academy of Sciences, The Health Committee.

Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics, 54*(1), 3-56. doi: http://dx.doi.org/10.1006/reec.1999.0225

Heckman, J. J. (2011). The Economics of Inequality: The Value of Early Childhood Education. *American Educator, 35*(1), 31-35.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160. doi: 10.1007/bf00401799

*Høien, T. (2007). *Håndbok til LOGOS : Teoribasert diagnostisering av lesevansker.* Bryne: Logometrica.

*Høien, T. & Tønnesen, G. (2008). *S40: Setningsleseprøven.* Lesesenteret, Universitetet i Stavanger; Logometrica

*Høien, T., & Lundberg, I. (1988). Stages of Word Recognition in Early Reading Development. *Scandinavian Journal of Educational Research, 32*(4), 163-182. doi: 10.1080/0031383880320402

*Høien, T., & Lundberg, I. (1991). *KOAS: Kartlegging av ordavkodingsstrategiene*. Stavanger: Senter for leseforsking

*Høien, T., & Tønnesen, G. (2008). *Instruksjonshefte til Ordkjedetesten*. Bryne: Logometrica

*Johnsen, K. (1980). *Noteringshefte med rettledning til Diagnostisk lese/skriveprøve 1 (1-3.trinn) og 2 (3. – 9.trinn)*. Drammen: PP-tjenestens materiellservice. Available from http://www.materiellservice.no/produktkategori/lese-og-skriveprover/

Johnson, T. P., & Wislar, J. S. (2012). Response rates and nonresponse errors in surveys. (Viewpoint essay). *JAMA, The Journal of the American Medical Association, 307*(17), 1805

*Lyster, S.-A. H., & Tingleff, H. (2002). *Ringeriksmaterialet kartlegging av språklig oppmerksomhet hos barn i alderen 5-7 år. Test/kopieringsoriginaler*. Oslo: Cappelen Damm

*Klinkenberg, J. E., & Skaar, E. (2003). *STAS. Manual*. Brandbu: PP-tjenestens Materiell Service.

Kunnskapsdepartementet [The Norwegian Ministry of Education]. (2011a). Meld. St. 18 (2010-11). *Læring og fellesskap. Tidlig innsats og gode læringsmiljøer for barn, unge og voksne med særlige behov.* In Kunnskapsdepartementet (Ed.), (Vol. 18). Oslo.

Kunnskapsdepartementet [The Norwegian Ministry of Education]. (2011b). *Vurdering av verktøy som brukes til å kartlegge barns språk i norske barnehager.*

Kunnskapsdepartementet [The Norwegian Ministry of Education]. (2017). Meld. St. 21 (2016-2017). *Lærelyst – tidlig innsats og kvalitet i skolen*.

McIntosh, K., Reinke, W. M., Kelm, J. L., & Sadler, C. A. (2012). Gender Differences in Reading Skill and Problem Behavior in Elementary School. *Journal of Positive Behavior Interventions*. doi: 10.1177/1098300712459080

Merrell, K. W. (2001). Assessment of Students' Social Skills: Recent Developments, Best Practices, and New Directions. *A Special Education Journal, 9*(1-2), 3-18. doi: 10.1080/09362835.2001.9666988

Miles, S. B., & Stipek, D. (2006). Contemporaneous and Longitudinal Associations Between Social Behavior and Literacy Achievement in a Sample of Low-Income Elementary School Children. *Child Development, 77*, 103–117. doi:10.1111/j.1467-8624.2006.00859.x

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). *Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6*(6): e1000097. doi:10.1371/journal.pmed1000097

*Nielsen, J. C. (2010). *Beskrivelse og vurdering af elevernes læsning og stavning: Vejledende materialer og diagnostiske prøver med henblik på målfastsættelse og planlægning*. København: Danmarks Pædagogiske Universitet

*Nielsen, J. C., Kreiner, S., Poulsen, A., & Søegård, A. (1995). *Lærerveiledning til setningsleseprøvene SL60 og SL40.* Senter for leseforsking

*Nielsen, J.C., Kreiner, S., Poulsen, A., & Søegård, A. (2008). *Lærerveiledning til leseprøvene OL64, OL120, MiniSL1 og MiniSL2.* Norsk versjon: Monsrud, M.-B., Godøy, O., Heller, A.K., & Thurmann-Moe, A.C. Oslo: Cappelen Akademisk

NOU 2015:8 (Official Norwegian Reports). *The School of the Future Renewal of subjects and competences.* Oslo: The Norwegian Ministry of Education and Research

OECD. (2015). *Skills for Social Progress: The Power of Social and Emotional Skills*. OECD Skills Studies. Paris: OECD Publishing

O'Reilly, T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing Reading Comprehension Assessments for Reading Interventions: How a Theoretically Motivated Assessment Can Serve as an Outcome Measure. *Educational Psychology Review*, *26*(3), 403-424. doi: 10.1007/s10648-014-9269-z

*Oslo kommune. (2012). *LUS-håndboken: Bruk av leseutviklingsskjema i grunnskolen*. Oslo: Oslo kommune, Utdanningsetaten.

*Ottem, E., & Frost, J. (2005). *Språk 6-16. Screening test*. Oslo: Bredtvet kompetansesenter

PsykTestBarn. (2016). EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests. *Test Review Form and Notes for Reviewers.* Version 4.2.6. Retrieved from http://www.psyktestbarn.no/CMS/ptb.nsf

Rivera, M., Al-Otaiba, S., & Koorland, M. (2006). Reading Instruction for Students With Emotional and Behavioral Disorders and At Risk of Antisocial Behaviors in Primary Grades: Review of Literature. *Behavioral Disorders, 31*(3), 323-339.

Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment]. (2014). *Dyslexi hos barn och ungdomar - tester och innsatser. En systematisk litteraurøversik. [Dyslexia in Children and Adolescence - Tests and Efforts: A Systematic Review]*. Stockholm: Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment]

Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & Scherer, R. (2016). Taking a future perspective by learning from the past – A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educational Research Review, 19,* 58-84. doi: http://dx.doi.org/10.1016/j.edurev.2016.05.002

[*]Sivertsen, R. (1990). *Aston Index. Prøve for observasjon og vurdering av lese-, skrive- og språkvansker.* Brandbu: Skolepsykologi-materiellservice.

[*]Skaathun, A. (2013). *Lesesenterets staveprøve*. Stavanger: Universitetet i Stavanger. Available from http://lesesenteret.uis.no/boeker-hefter-og-materiell/boeker-og-hefter/lesesenterets-staveprove-article85269-12686.html . ISBN: 978-82-7649-079-4

*Solheim, O. J. (2015a). *Kartleggingsprøven i lesing for 1. trinn.* Rapport basert på ordinær gjennomføring våren 2015. Stavanger:[**]

*Solheim, O. J. (2015b). *Kartleggingsprøven i lesing for 2. trinn.* Rapport basert på ordinær gjennomføring våren 2015. Stavanger: Universitetet i Stavanger[**]

*Solheim, O. J. (2015c). *Kartleggingsprøven i lesing for 3. trinn.* Rapport basert på ordinær gjennomføring våren 2015. Stavanger: Universitetet i Stavanger[**]

[**] Received on request from the Norwegian Directorate for Education and Training

Standards & Testing Agency. (2015). Retrieved October 2015, from https://www.gov.uk/government/organisations/standards-and-testing-agency

Statistic Norway. (2016). Elever i grunnskolen. Spesialundervisning. Retrieved from http://www.ssb.no/utgrs/#

Stewart, R. M., Benner, G. J., Martella, R. C., & Marchand-Martella, N. E. (2007). Three-Tier Models of Reading and Behavior: A Research Review. *Journal of Positive Behavior Interventions, 9*(4), 239-253.

*Støle, H., Mangen, A., & Stangeland, E. B. (2015). *Den nasjonale prøven i lesing på 5.trinn 2015.* Stavanger: Universitetet i Stavanger**

*Sutherland, M. J., & Smith, C. D. (1991). Assessing Literacy Problems in Mainstream Schooling: a critique of three literacy screening tests. *Educational Review, 43*(1), 39-48. doi: 10.1080/0013191910430104

Terras, M. M., Thompson, L. C., & Minnis, H. (2009). Dyslexia and psycho-social functioning: an exploratory study of the role of self-esteem and understanding. *Dyslexia, 15*(4), 304-327. doi: 10.1002/dys.386

The National Assessment Governing Board. (2013). *Reading framework for the 2013 National Assessment of Educational Progress.* US Department of Education. Washington, DC: Author. Retrieved from http://files.eric.ed.gov/fulltext/ED542063.pdf

Thorndike, R. M., & Thorndike-Christ, T. (2014). *Measurement and evaluation in psychology and education* (8th ed. ed.). Harlow: Pearson.

*Topstad, I. (2000): *"Leseklar? Kartlegging av språklig bevissthet i 1. – 2. Klasse".* Kristiansand: Arbeid med ord læremidler A/S

** Received on request from the Norwegian Directorate for Education and Training

*Topstad, I. (2001): *Kartlegging av leseferdighet 2. klasse.* Kristiansand: Arbeid med ord Læremidler A/S

Undheim, A. M., Wichstrøm, L., & Sund, A. M. (2011). Emotional and behavioral problems among school adolescents with and without reading difficulties as measured by the youth self-report: a one-year follow-up study. *Scandinavian Journal of Educational Research, 55*(3), 291-305.

*Utdanningsdirektoratet [The Norwegian Directorate for Education and Training]. (2016, March 30). *Elevenes sosiale kompetanse - hva kan du som lærer gjøre? Kartlegging av klassens sosiale kompetanse.* [Students' Social Competence]. Retrieved from http://www.udir.no/laring-og-trivsel/laringsmiljo/psykososialt-miljo/sosial-kompetanse/struktur-og-regler/

*Utdanningsdirektoratet [The Norwegian Directorate for Education and Training]. (2016, February 23). *Elevundersøkelsen.* [Student Survey]. Retrieved from http://www.udir.no/tall-og-forskning/brukerundersokelser/elevundersokelsen/

*Utdanningsdirektoratet [The Norwegian Directorate for Education and Training]. (2015). *Nasjonale prøver. 2015. Lesing 5. Trinn* [National Tests. Reading Grade 5.] Oslo: Utdanningsdirektoratet**

*Utdanningsdirektoratet [The Norwegian Directorate for Education and Training]. (2011). *Rammeverk for kartleggingsprøver på barnetrinnet*. Oslo: Utdanningsdirektoratet**

*Utdanningsdirektoratet [The Norwegian Directorate for Education and Training]. (2016). *Metodegrunnlaget for nasjonale prøver.* Oslo: Utdanningsdirektoratet

** Received on request from the Norwegian Directorate for Education and Training

Welsh, M., Parke, R., Widaman, K., & O'Neill, R. (2001). Linkages between students' social and academic competence: A longitudinal analysis. *Journal of School Psychology, 39*(6), 463-482.

Table 1

Assessment Instruments Identified for Inclusion or Exclusion in the Systematic Literature Review

Assessment Instrument	Include	Exclude	Exclusion Reason
Social Functioning Instruments			
ADDES: Attention Deficit Disorder Evaluation Scale		X	Identify attention deficit disorders
ECBI: Eyberg Child Behavior Inventory		X	Parent scale on conduct disorders
Elevundersøkelsen [Students' Self-report Olweus' Program]		X	Identify bullying
Elevundersøkelse [Students' Self-report Respect Program]		X	Identify bullying
Elevens selvrapport [Student's Self-report]	X		
Elevenes sosiale kompetanse [Students' Social Competence]	X		
Elevundersøkelsen [Student Survey]	X		
Ingen utenfor [None Outside]		X	Activity Materials
Innblikk [Insight]		X	Identify bullying
Klassetrivsel.no – Sociometric		X	Data entering system
Lions Quest		X	Intervention program
Mitt valg [My Choice]		X	Intervention program
Mobbeundersøkelsen [Bullying Survey Zero-program]		X	Identify bullying
Psykologisk 1.hjelp [Psychological First Aid]		X	Guidelines to mental health
Snakk med meg! [Talk with Me]		X	Identify bullying
Sociometric	X		No information available*
Steg for steg [Second Step]		X	Intervention program
SWIS (School-Wide Information System)		X	Data entering system
Teacher-made Students' Self-report		X	Not applicable
Teacher-made Teacher ratings		X	Not applicable
Zippy's Friends		X	Intervention program
TOTAL Social Functioning:	4	17	
Reading Instruments			
AMO [Automatic Most Frequent Words]	X		No information available*
Arbeid med ord [Working with Words]	X		
Arbeidsprøven - Dynamisk kartlegging [The Working Test]	X		
Aski Raski (Intervention and Assessment)	X		
Aston Index	X		
Bokstavtesten [Letter Test]	X		No information available*
Carlsten's Reading Test	X		
DAMMS leseunivers [Reading Univers]	X		
Diagnostisk lese- og skriveprøve [Diagnostic Reading Test]	X		
God leseutvikling [Good Reading Development]	X		
HOA lesetest [reading test]	X		No information available*
IL-basis	X		
Kartleggeren.no	X		
KOAS: Kartlegging av ordavkodingsstrategier [Worddecoding]	X		
KTI: Kontrollert tegneiakttagelse		X	Mapping language skills
Leselos	X		
Lesesenterets staveprøve [Spelling Test]	X		
LOGOS	X		
LUS (Leseutviklingsskjema) [Reading Development Form]	X		
National Tests Reading Proficiency (NTRP)	X		
NSL (Norsk Som Læringsspråk) [Norwegian as Learning Language]		X	Language skills 2nd languager
OL64; OL120/MiniSL1; MiniSL2	X		
Ordkjedetest [Word Chain Test]	X		
OS400	X		No information available*

(continued)

Table 1 (continued)

Assessment Instrument	Include	Exclude	Exclusion Reason
Reading Instruments			
Osloprøven i lesing [The Oslo Reading Test]	X		No information available*
På vei til å bli en god leser [To be a good reader]r		X	Assess deaf children
ReleMo		X	Intervention
Ringeriksmaterialet	X		
Setningsleseprøve SL60 /SL40 [Sentence Reading Test]	X		
SOL (Systematisk Observasjon av Lesing [Systematic Observation of Reading])	X		
Språk 6-16 [Language 6-16]	X		
STAS (Standardisert Test i Avkoding og Staving [Standardized Test in Decoding and Spelling])	X		
Tempolex		X	Intervention
TRAS		X	Language skills in pre-school
20 spørsmål om språk [20 Questions about Language]		X	Language skills and relations
Teacher-made Teacher Ratings		X	Not applicable
TOTAL Reading	28	8	

Note. *No information available = Assessment Instruments met the inclusion criteria but no documentation available for the EFPA review.

Table 2

Assessment Instruments and Connected Publications Included in the EFPA Review

Instrument	Author	Publication Type
Social Functioning		
1. Elevens selvrapport: Systematisk kartlegging av elevens subjektive forståelse av egen livssituasjon [Student's Self-report]	Duna & Frost, 1999	Manual
2. Elevundersøkelsen [Students' Survey of Self-assessment of Learning and Well-being]	Utdanningsdirektoratet [The Norwegian Directorate for Education and Training], 2016	Information and Material
3. Kartlegging av klassens sosiale miljø - "Elevenes sosiale kompetanse"[Students' Social Competence]	Utdanningsdirektoratet [The Norwegian Directorate for Education and Training], 2016	Information and Material
Reading Proficiency		
1. arbeidmedord.no: «Leseklar» og «Kartlegging av leseferdighet» [Working with Words]	Topstad, 2000, 2001	Manuals
2. «Arbeidsprøven - Dynamisk kartlegging» [The Working Test]	Duna, Frost, Godøy, & Monsrud, 2003	Manual
3. Aski Raski	Ask, 2002-2016	Information and Material
4. Aston Index	Sivertsen, 1990 Sutherland & Smith, 1991 Gallefoss, 1996	Manual & Study Study – Article Master's Thesis
5. Carlsten leseprøve [Carlsten Reading test]	Carlsten, 2016	Manuals
6. DAMMS leseunivers – Ressursperm [Reading Univers]	Aschim, 2006	Manual
7. Diagnostisk lese- og skriveprøve [Diagnostic Reading Test]	Johnsen, 1980	Manual
8. God leseutvikling [Good Reading Development]	Lundberg & Herrling, 2008	Manual
9. IL-basis	Evensen, 2011 Frost & Nielsen, 2000	Master's Thesis Manual
10. Kartleggeren.no	Fagbokforlaget, 2016	Information and Material
11. KOAS: Kartlegging av ordavkodingsstrategiene [Worddecoding]	Høien & Lundberg, 1988, 1991	Manual & Study
12. Leselos	Engen & Helgevold, 2012	Manual
13. Lesesenterets staveprøve [Spelling Test]	Skaathun, 2013	Manual
14. LOGOS	Høien, 2007	Manual
15. LUS: LeseUtviklingsSkjema [Reading Development Form]	Oslo kommune, 2012 Allard et al., 2006	Manual Manual

(continued)

Table 2 (continued)

Instrument	Author	Publication Type
Reading Proficiency		
16. NTRP: Nasjonale kartleggingsprøver lesing 1.-3. trinn/ Nasjonale prøver i lesing 5.trinn [National Tests Reading Proficiency]	The Norwegian Directorate for Education and Training, 2011, 2015, 2016; Solheim, 2015a, 2015b, 2015c; Støle, Mangen, & Stangeland, 2015**	Information and Material
17. OL64; OL120 / MiniSL1; MiniSL2	Nielsen, 2010; Nielsen, Kreiner, Poulsen, & Søegård, 2008	Manual Manual
18. Ordkjedetest [Word Chain Test]	Asbjørnsen, Obrzut, Eikeland, & Manger, 2010; Høien & Tønnesen, 2008	Article –Study Manual
19. Ringeriksmaterialet	Lyster & Tingleff, 2002	Manual
20. S40 setningsleseprøve [Sentence Reading Test]	Høien & Tønnesen, 2008	Manual
21. Setningsleseprøve SL60 /SL40 [Sentence Reading Test]	Nielsen, Kreiner, Poulsen, & Søegård, 1995	Manual
22. SOL (Systematisk Observasjon av Lesing [Systematic Observation of Reading])	Gjesdal kommune, 2011	Information and Material
23. Språk 6-16 [Language 6-16]	Ottem & Frost, 2005, 2007	Manual
24. STAS (Standardisert Test i Avkoding og Staving [Standardized Test in Decoding and Spelling])	Klinkenberg & Skaar, 2003	Manual

** Received on request from the Norwegian Directorate for Education and Training

Table 3.

Number of Elementary Schools Using Educational Assessment Instruments (EAI) for Children's Social Functioning and Reading Proficiency

		Use EAI for social functioning		
		No	Yes	Total
Use EAI for reading	No	22 (9.4%)	4 (1.7%)	26 (11.1%)
	Yes	139 (59.4%)	69 (29.5%)	208 (88.9%)
	Total	161 (68.8%)	73 (31.2%)	234 (100%)

Note. EAI = Educational Assessment Instruments

Table 4.

Assessing Frequencies of Social Functioning and Reading in Number of Elementary Schools

Frequences of assessing reading		Frequences of assessing social functioning					
		< once a year	once a year	Two times per year	Three times per year	> Three times per year	Total
	Once a year	2 (0.9)	6 (2.6)	5 (2.1)	1 (0.4)	4 (1.7)	18 (7.7)
	Two times per year	18 (7.7)	11 (4.7)	46 (19.7)	0	31 (13.2)	106 (45.3)
	Three times per year	11 (4.7)	8 (3.4)	15 (6.4)	3 (1.3)	15 (6.4)	52 (22.2)
	> Three times per year	10 (4.3)	3 (1.3)	19 (8.1)	1 (0.4)	25 (10.7)	58 (24.8)
	Total	41 (17.5)	28 (12)	85 (36.3)	5 (2.1)	75 (32.1)	234 (100)

Note. Numbers in brackets are the percentage of schools.

Table 5.

Number of Norwegian Elementary Schools Using Information Derived from the Results of Assessing Children to Lead Decisions to Promote Children's Social and Reading Skills

	Use of Assessing Information to Promote Social Skills	Use of Assessing Information to Promote Reading Skills
No	11 (4.7%)	0
Yes	214 (91.5%)	231 (98.7%)
Don't know	9 (3.8%)	3 (1.3%)
Total	234 (100%)	234 (100%)

Table 6
Descriptions and Characteristics for the Assessment Instruments of Social Functioning and Reading Proficiency

Instrument (Year reviewed version)	Purpose of instrument	Type of measure	Grade/ Age group	Response Mode	Number of sub-scales	Number of items (bracket is the scale #)	Scoring	Items Format	Time to administer	Test Related Qualification Required	Number of studies	Sample Size
						Social Skills						
Elevens selvrapport [Student's Self- report] (1999)	Assess self-perception, motivation, social and emotional function for academic struggling students	Individual	Gr. 1-7	PP; DO	4	17(1); 26(2); 7(3); 6(4)	TN; # of total Yes/No/ Don't know	OI; Open; Likert	No Info	No	No Info	No Info
Elevenes sosiale kompetanse [Students' Social Competence] (2016)	Assess social climate/relations in class, students' social functioning/ competence	Group Individual	Gr. 1-10	PP	1	6	No Info	Likert; Open	No Info	No	No Info	No Info
Elev-undersøkelsen [Student Survey] (2016)	Self-assess learning and well-being	Group Individual	Gr. 5-10 Uppergr.	CB	10	55 + 27(Gr.6-10)	Raw-score	Likert	No Info	No	No Info	No Info
						Reading Skills						
Arbeid med ord [Working with Words] (2000/2001)	Assess phonological awareness, letter knowledge to judge development in reading	Individual	Gr. 1-2	PP	2	17(1); 14 (2)	TN; R/W	Open	No Info	No	No Info	No Info
Arbeidsprøven [The Working Test] (2008)	Assess listening reading comprehension, sentence memory, vocabulary, letter knowledge, spelling, writing	Individual	Gr. 1-2	PP; DO	14	8(1); 10(2); 10(3); 9(4); 10(5); 1(6); 4(7); 1(8); 1(9); 3(10); 3(11); 3(12); 5(13); 1(14)	TN; R/W	MC; Likert	No Info	No	No Info	No Info

(continued)

Table 6 (continued)

Instrument (Year reviewed version)	Purpose of instrument	Type of measure	Grade/ Age group	Response Mode	Number of sub-scales	Number of items (bracket is the scale #)	Scoring	Items Format	Time to administer	Test Related Qualification Required	Number of studies	Sample Size
Aski Raski (2016)	Assess decoding skills: graphems, spelling, consonant, morphems	Individual	Gr. 2-4	PP; CB	12	240 (20 per subtest)	R/W; Cut-off	MC	5-10 min	No	No Info	No Info
Aston Index (1988)	Assess difficulties in reading/ writing/ language/ auditory perception/ memory, motor/ laterality to identify risk for dyslexia	Individual	Age 6-14	PP; DO	17	1-26 per sub-scale	TN; R/W; Raw-Score Norms	MC	No Info	Yes	3	≤ 300
Carlsten (2016)	Assess reading and writing	Screening Group Individual	Gr. 1-10 Upper gr.	PP; DO	8	25 (2)	Cut-off; TN	Dictate; Texts	No Info	No	No Info	No Info
DAMMs leseunivers [Reading Univers] (2014)	Assess development in reading and writing	Individual	Gr. 1-7	PP; DO	11	No Info	R/W; TN; TR	Likert; Open; MC	No Info	No	No Info	No Info
Diagnostisk lese og skriveprøve [Diagnostic Reading and Writing Test] (1980))	Assess development in reading and writing	Individual	Gr. 1-3	PP; DO	2	10(1); 10(2)	TN; R/W	Dictate; Texts; DO	No Info	No	No Info	No Info
God lese-utvikling [Good Reading Development] (2008)	Assess phonological awareness, decoding, fluency, comprehension, reading motivation along with training	Screening	Gr. 1-7	PP; DO	5	11(1); 13(2); 8(3); 17(4); 12(5)	TN; R/W	MC	No Info	No	No Info	No Info

(continued)

Table 6 (continued)

Instrument (Year reviewed version)	Purpose of instrument	Type of measure	Grade/ Age group	Response Mode	Number of sub-scales	Number of items (bracket is the scale #)	Scoring	Items Format	Time to administer	Test Related Qualification Required	Number of studies	Sample Size
IL-basis (2000)	Assess listening comprehension, language awareness, word comprehension, letter knowledge, writing letter/ words	Group Individual	Gr. 1-2	PP	14 + 3	1(1); 5(2); 5(3); 6(4); 11(5); 6(6); 5(7); 12(8); 24(9); 24(10); 6(11); 8(12); 6(13); 1(14) + 1(x3)	TN; R/W	MC; Write/ Draw	10-20 min	No	No Info	No Info
Kartleggeren.no (2016)	Assess reading skills, orthography, vocabulary	Screening	Gr. 5-10	CB	3	4(1); 3(2); 4(3)	R/W	MC	45 min	No	No Info	No Info
KOAS (1991)	Assess cognitive processes/ strategies in reading words: decoding, orthography, phonology	Individual	Gr. 3-7	CB	5	72(1)	TN; R/W; Z-Score; Norms	MC	No Info	Yes	1	300
Lesesenterets staveprøve [Spelling Test] (2013)	Assess spelling skills	Screening/ Group Diagnost/ Individual	Gr. 3-10	PP; DO	1	32	R/W Norms	Dictate; MC	No Info	No	1	≤ 454
Leselos (2010)	Assess reading development; reading words, decoding letters/ lettersequences/ morphems, fluency, textreading	Individual	Gr. 1-10	PP; DO; CB	5	17	TR: (1) master by modeling, (2) master with other, (3) master independcy	Check-Form	No Info	No	No Info	No Info
LOGOS (2012)	Assess reading fluency, listening /reading comprehension, conceptual, vocabulary, decoding	Individual	Gr. 1-10	PP; CB	3	13(1); 18(2); 15(3)	TN; Z-Score; Cut-off; Norms	Open; MC	60-90min	Yes	1	≤ 482

(continued)

Table 6 (continued)

Instrument (Year reviewed version)	Purpose of instrument	Type of measure	Grade/ Age group	Response Mode	Number of sub-scales	Number of items (bracket is the scale #)	Scoring	Items Format	Time to administer	Test Related Qualification Required	Number of studies	Sample Size
LUS (2012)	Assess reading development	Individual	Gr.1-10	PP; DO; CB	3	12(1); 7(2); 1(3)	TN	OF	No Info	No	No Info	No Info
NTRP (2015)	Assess reading development and identify at risk students (find information, comprehension, reflection)	Group Individual	Gr. 1-3;	PP	6(Gr1); 4(Gr2/3);	86(Gr1); 54/74(Gr2/3);	R/W; Cut-off	MC	60 min	No	3	≤ 1 097
			Gr. 5		5(Gr.5)	29(Gr5)		MC; Open	90 min		1	55 703
OL64/120; MiniSL1/SL2 (2008)	Assess incipient reading development/motivation (words/sentences)	Screening Group	Gr. 1-3 Gr. 1-2	PP	2; 3	64(1); 120(2); 12(1); 2(2); 8/7(3)	R/W; Cut-off; Norms; Raw-Score	MC; Open; Likert	10-15 min 45 min	No	4	≤ 420
Ordekjedetesten [Word Chain Test] (2008)	Screen difficulties in word recognition and decoding	Group Individual	Gr. 3-10 Adults	PP	1	90	R/W; Stanine; Norms	MC; TL	10min	No	2	≤ 421
Ringeriks-materialet (2002)	Assess language awareness (listening, rhymes, phonological/ morphological/ syntactic awareness, words memory, rapid words, grammar/ syntactic knowledge/ comprehension, letter knowledge)	Individual	Age 5-7	PP	17	11(1); 6(2); 16(3); 10(4); 10(5); 10(6); 6(7); 9(8); 16(9); 9(11); 13(12); 13(13); 6(14); 18(15); 10(16), 1(17)	R/W; Raw-Score	MC	No Info	No	1	273
S-40 (2008)	Assess skills in reading sentences	Group Individual	Gr. 4-10	PP	1	40	R/W; Stanine; Norms; Raw-Score	MC	10 min	No	1	1 984

(continued)

Table 6 (continued)

Instrument (Year reviewed version)	Purpose of instrument	Type of measure	Grade/ Age group	Response Mode	Number of sub-scales	Number of items (bracket is the scale #)	Scoring	Items Format	Time to administer	Test Related Qualification Required	Number of studies	Sample Size
SL60/SL40 (1995)	Assess skills in reading sentences	Group	Gr. 3-4	PP	2	60(1); 40(2)	R/W; Cut-off; Norms; Raw-Score	MC; TL	15 min	Yes	2	≤ 393
SOL (2011)	Assess level of reading development across grades	Individual	Gr. 1-10	DO; CB	10	5-10 per scale	TN	Likert; OF	No Info	Yes	No Info	No Info
Språk 6-16 [Language 6-16] (2007)	Identify language difficulties (words/ sentence memory, conceptual contrasts, word-knowledge	Individual Screening	Age 6-16	PP	4 + 3	13(1); 12(2); 13(3); 4(4) + 19(1); 23(2); 62(3)	R/W; Z-Score; Norms	MC	No Info	Yes	2	≤ 1 214
STAS (2003)	Assess skills in decoding and spelling	Screening Group Individual	Gr. 2-10	PP	8	No Info	R/W; Norms; Raw-Score	MC; TL	No Info	Yes	1	1 022

Note. CB = Computer Based; DO = Direct Observation; MC = Multiple Choice; No Info = No information available to rate; NTRP = National Test Reading Proficiency; OF = Observation Form; OI = Oral Interview; PP = Paper and Pencil; R/W = Right/Wrong; TN = Teacher's Note; TL = Time Limit; TR = Teacher's Ratings

Table 7
Overview of Material Quality and Documented Psychometric Properties of the Assessment Instruments of Social Skills and Reading Skills

Instrument (Year of reviewed version)	Quality of rationale explanation, presentation & information	Quality of Test Material		Quality Computer generated reports	Norms	Reliability	Validity
		Paper Pencil	CBT/ WBT				
			Social Skills				
Elevens selvrapport [Student's Self-report] (1999)	Adeq	Good	NA	NA	NA	No Info	No Info
Elevenes sosiale kompetanse [Students' Social Competence] (2016) *	No Info	Inadeq	NA	NA	No Info	No Info	No Info
Elevundersøkelsen [Student Survey] (2016)	No Info	NA	Good	Good	NA	No Info	No Info
			Reading				
Arbeid med ord [Working with Words] (2000/2001) *	Inadeq	Inadeq	NA	NA	NA	No Info	No Info
Arbeidsprøven [The Working Test] (2008) *	Inadeq	Adeq	NA	NA	NA	No Info	No Info
Aski Raski (2016) *	Inadeq	Inadeq	No Info	Inadeq	NA	No Info	No Info
Aston Index (1988)	Good	Good	NA	NA	Inadeq	No Info	No Info
Carlsten (2016) *	Inadeq	Adeq	NA	NA	No Info	No Info	No Info
DAMMs leseunivers [Reading Univers] (2014)	Inadeq	Good	NA	NA	NA	No Info	No Info
Diagnostisk lese- og skriveprøve [Diagnostic Reading and Writing Test] (1980) *	Inadeq	Inadeq	NA	NA	NA	No Info	No Info
God lese-utvikling [Good Reading Development] (2008)	Adeq	Good	NA	NA	NA	No Info	No Info
IL-basis (2000)	Good	Good	NA	NA	NA	No Info	No Info
Kartleggeren.no (2016)	Adeq	NA	Good	Adeq	No Info	No Info	No Info
KOAS (1991)	Good	NA	Exc	Adeq	Inadeq	No Info	No Info
Lesesenterets staveprøve [Spelling Test] (2013)	Good	Good	NA	NA	Good	Good	No Info
Leselos (2010) *	Inadeq	Adeq	NA	NA	No Info	No Info	No Info
LOGOS (2012)	Good	Good	Exc	Good	Good	Adeq	Inadeq
LUS (2012) *	Inadeq	Adeq	No Info	No Info	NA	No Info	No Info
NTRP (2015) *	Adeq	Adeq	NA	NA	No Info	No Info	No Info

(continued)

Table 7 (continued)

Instrument (Year of reviewed version)	Quality of rationale explanation, presentation & information	Quality of Test Material		Quality Computer generated reports	Norms	Reliability	Validity
		Paper Pencil	CBT/ WBT				
			Reading				
OL64/120; MiniSL1/SL2 (2008) *	Inadeq	Good	NA	NA	Inadeq	No Info	No Info
Ordkjedetesten [Word Chain Test] (2008)	Good	Good	NA	NA	Good	Good	Good
Ringeriksmaterialet (2002) *	Adeq	Good	NA	NA	No Info	Inadeq	No Info
S-40 (2008)	Good	Good	NA	NA	Adeq	Adeq	Adeq
SL60/SL40 (1995)	Adeq	Adeq	NA	NA	Inadeq	No Info	No Info
SOL (2011) *	Adeq	Good	Good	No Info	NA	No Info	No Info
Språk 6-16 [Language 6-16] (2007)	Exc	Exc	NA	NA	Good	Good	Good
STAS (2003) *	Good	Adeq	NA	NA	Inadeq	No Info	Inadeq

Note. * = double rated; Adeq = Adequate; CBT = Computer-Based Test; Exc = Excellent; Inadeq = Inadequate; NA = Not Applicable; No Info = No information available to rate; NTRP = National Test Reading Proficiency; WBT = Web-Based Test

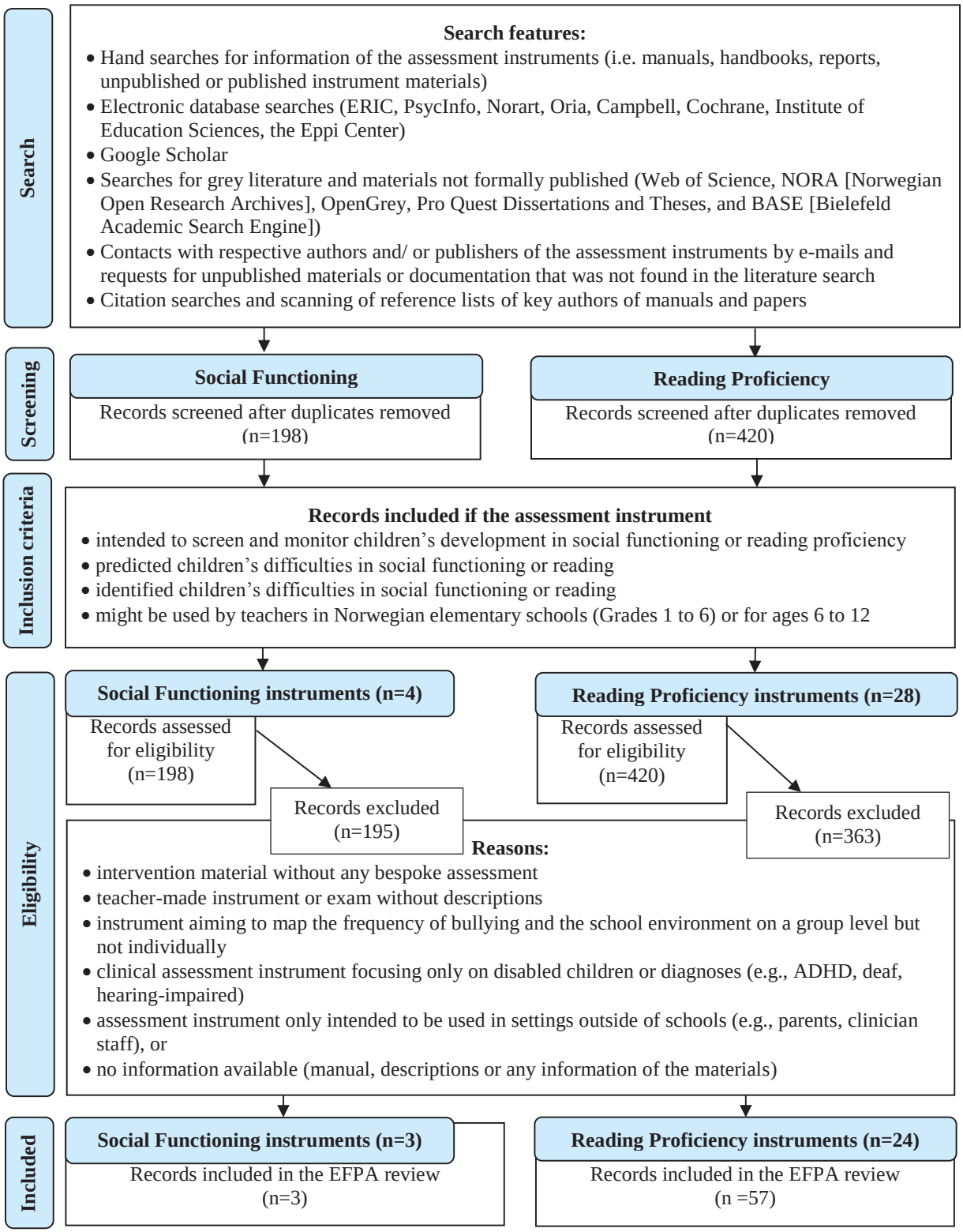


Figure 1. Flow chart for the search and inclusion of studies and materials of the assessment instruments to be included for the EFPA review (modified after Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009). Records refer to the identified publications on the instruments.

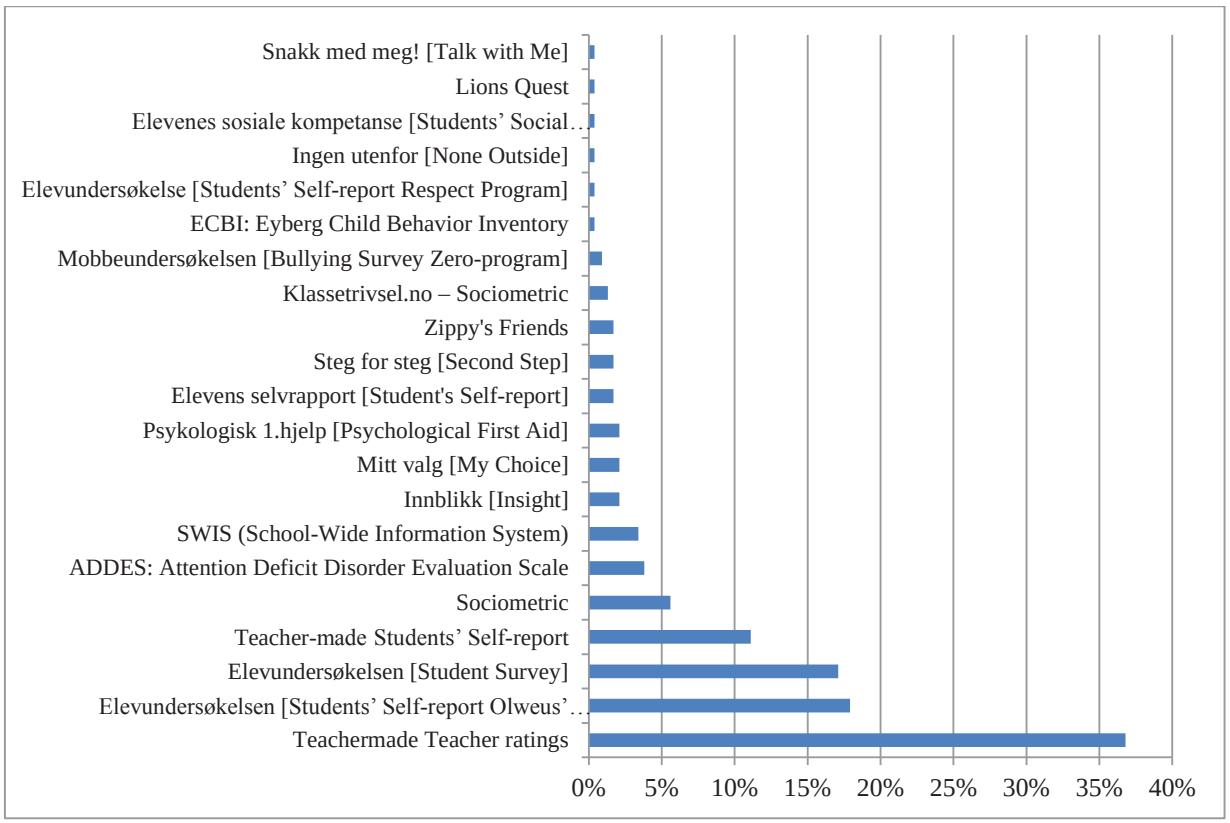


Figure 2. Percentage of schools (n=234) reported use of social functioning assessment instruments

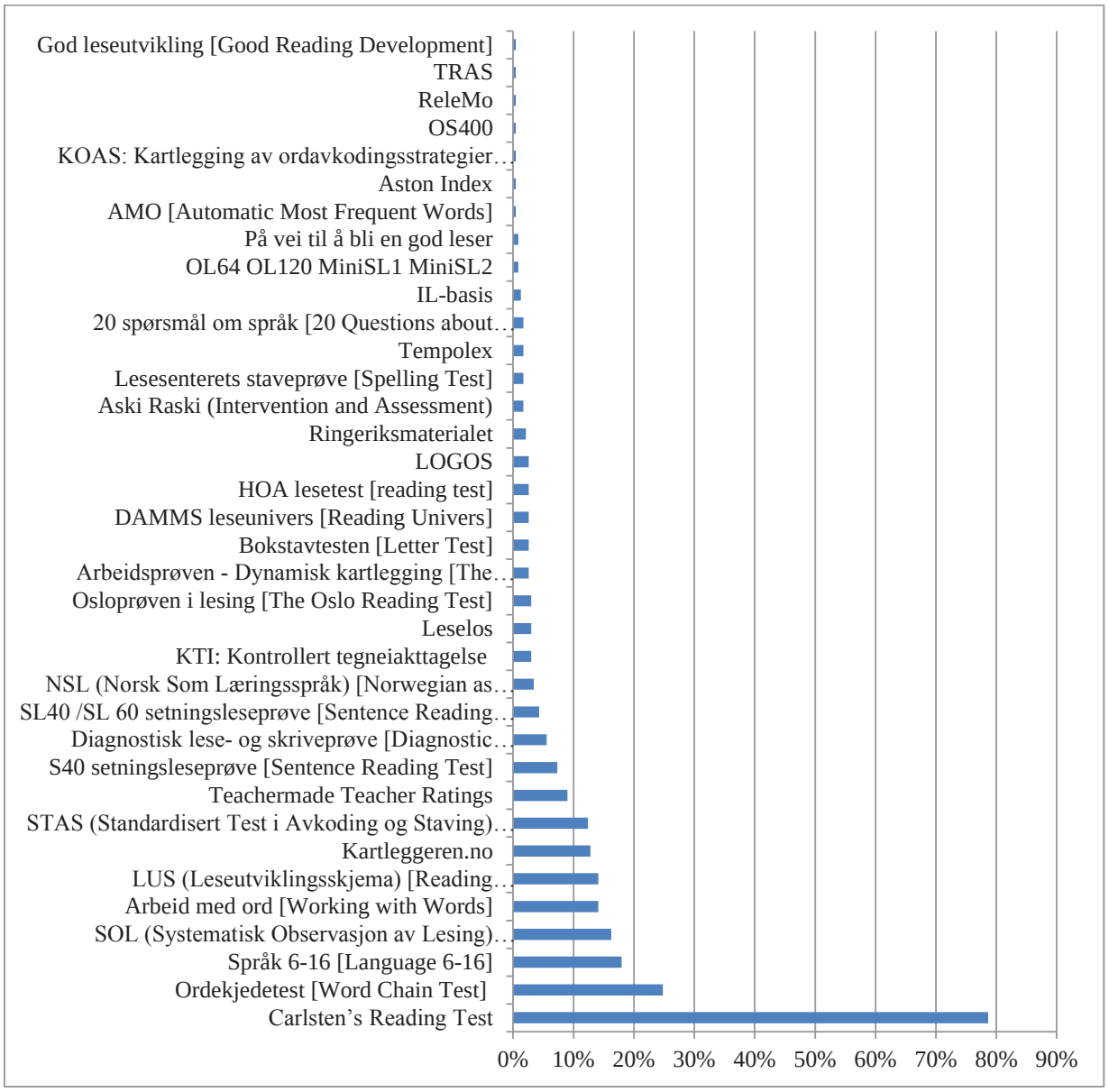


Figure 3. Percentage of schools (n=234) reported use of reading assessment instruments

Appendix A

Search Key Words with Synonyms and Alternative Expressions of Educational Assessment Instruments on Social Functioning

social*	**assess***	**psychometr***	**"elementary school"**	**"at risk"**
social abilit*	assess*	analys*	"early child education"	risk assess*
social achievm*	evaluat*	construct valid*	"early school age"	risk factor*
social behav*	exam*	norm* (normed)	"primary school"	risk manage*
social competenc*	identif*	propert* (properties)		at risk student*
social develop*	map*	reliab*		
social difficult*	measur*	standard* (standardized)		
social function*	predict*	valid*		
social performanc*	"progress monitor*"			
social problem*	rat* (rating; rate)			
social proficienc*	"respons to intervention"			
social skill*	scale*			
	screen*			
	test*			

Note. Key Words are bolded. * = truncation of search words

Appendix B

Search Key Words with Synonyms and Alternative Expressions of Educational Assessment Instruments on Reading Proficiency

read*	**assess***	**psychometr***	**"elementary school"**	**"at risk"**
reading abilit*	assess*	analy*	"early child education"	at risk student*
reading achiev*	exam*	construct valid*	"early school age"	risk assess*
reading competenc*	evaluat*	cut* score*	"primary school"	risk factor*
reading comprehens*	"formative evaluation"	norm*		risk manage*
reading decod*	identif*	propert*		
reading develop*	map*	reliab*		
reading difficult*	measur*	standard*		
reading fluenc*	predict* ("predictive measurement")	"test length"		
reading performanc*	"progress monitor*"	valid*/ test validity		
reading proficienc*	rat* (rating; rate)			
reading skill*	scal*			
"oral reading"	screen*			
	test*			

Note. Key Words are bolded. * = truncation of search words

Appendix B

Search Key Words with Synonyms and Alternative Expressions of Educational Assessment Instruments on Reading Proficiency

read*	**assess***	**psychometr***	**"elementary school"**	**"at risk"**
reading abilit*	assess*	analy*	"early child education"	at risk student*
reading achiev*	exam*	construct valid*	"early school age"	risk assess*
reading competenc*	evaluat*	cut* score*	"primary school"	risk factor*
reading comprehens*	"formative evaluation"	norm*		risk manage*
reading decod*	identif*	propert*		
reading develop*	map*	reliab*		
reading difficult*	measur*	standard*		
reading fluenc*	predict* ("predictive measurement")	"test length"		
reading performanc*	"progress monitor*"	valid*/ test validity		
reading proficienc*	rat* (rating; rate)			
reading skill*	scal*			
"oral reading"	screen*			
	test*			

Note. Key Words are bolded. * = truncation of search words

Paper II:

Arnesen, A., Smolkowski, K., Ogden, T., & Melby-Lervåg, M. (2017). Validation of the Elementary Social Behavior Assessment: Teacher ratings of students' social skills adapted to Norwegian, Grades 1 to 6. *Emotional and Behavioural Difficulties*. doi: 10.1080/13632752.2017.1316473

II

Paper III:

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. doi:10.1002/rrq.159

III

INTERNATIONAL LITERACY ASSOCIATION READING RESEARCH QUARTERLY

Growth in Oral Reading Fluency in a Semitransparent Orthography: Concurrent and Predictive Relations With Reading Proficiency in Norwegian, Grades 2–5

Anne Arnesen
Johan Braeken
University of Oslo, Norway

Scott Baker
Southern Methodist University, Dallas, Texas, USA

Wilhelm Meek-Hansen
Terje Ogden
Norwegian Center for Child Behavioral Development, Oslo, Norway

Monica Melby-Lervåg
University of Oslo, Norway

ABSTRACT

This study investigated an adaptation of the Oral Reading Fluency (ORF) measure of the Dynamic Indicators of Basic Early Literacy Skills into a European context for the Norwegian language, which has a more transparent orthography than English. Second-order latent growth curve modeling was used to examine the longitudinal measurement invariance of the ORF measure, the growth in oral reading fluency within and across grades 2–5, the relative stability of the ORF measure, and the relationship between the ORF measure and high-stakes national tests of reading proficiency. Results showed that the ORF passages measured the same underlying construct, but some passages stood out regarding the invariance pattern. The oral reading fluency growth curve models demonstrated a linear growth in grades 2 and 3 and a nonlinear growth in grades 4 and 5. Initial individual differences varied more than growth rates, which for all were positive but largest in grades 3 and 4. High relative stability in the ORF measure was found across grades. The concurrent and predictive relations of the ORF measure on the Norwegian national reading tests were moderate to strong (range = .44–.75). Findings indicated that the ORF is a reliable and valid measure of reading in Norwegian grades 2–5 and easy and fast to administer. The ORF measure might contribute to early identification of students at risk for reading difficulties in an orthography more transparent than English. Implications for school practice and future research are discussed.

In the United States, elementary schools commonly use a measure of reading fluency, called oral reading fluency (ORF), to screen students for reading difficulties and examine their reading progress over time (S.K. Baker et al., 2008; Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Shinn, 1998). In this approach, students read aloud grade-specific stories in a one-on-one testing setting, and the number of words read correctly in one minute constitutes their reading performance score (Deno et al., 1982; Shinn, 1989, 1998). In a systems-level approach to screening students for reading problems and monitoring their progress over time, the ORF measure is typically administered three times per school year to all students (S.K. Baker et al., 2011; Shinn, 1989; Shinn, Shinn, Hamilton, & Clarke, 2002).

In Europe, however, few systematic studies have been conducted concerning the instruments that schools use to assess reading skills

Reading Research Quarterly, 52(2)
pp. 177–201 | doi:10.1002/rrq.159

and progress. For reading proficiency, there seems to be a large variation in the types of screening instruments that schools use (Standards & Testing Agency, 2016; Statens Beredning för Medicinsk Utvärdering [SBU], 2014). In the United Kingdom, reading assessments have traditionally focused on reading accuracy tests, such as the phonics screening check applied in first grade (see Standards & Testing Agency, 2016). The emphasis on accuracy is likely to be due to the fact that English has an opaque orthography with inconsistent relations between letters and sounds as compared with other European languages. In more transparent European languages, however, assessments of decoding skills have generally focused on reading fluency rather than accuracy (for an overview, see SBU, 2014). To our knowledge, none of the tests used in European settings include monitoring students' progress in reading fluency over time, as is done with the ORF measure. That is, most of the reading fluency measures in European settings are administered as one-shot assessments. This is a serious omission in light of the importance of reading fluency in the overall development of reading proficiency (Kuhn & Stahl, 2003; LaBerge & Samuels, 1974). Also, concerns have been raised about the lack of psychometric validation of the screening tests and their ability to identify struggling readers (Duff, Mengoni, Bailey, & Snowling, 2015; SBU, 2014).

Identifying struggling readers at an early age is important to provide appropriate interventions for these students. Many students fail in developing well-functioning reading skills. For instance, the PISA studies have shown that 24% of the 15-year-old students in the Organisation for Economic Co-operation and Development (OECD) member countries have low performance in reading comprehension (OECD, 2013). This problem is worrisome because reading comprehension is consistently, across many different contexts (e.g., across languages, in many different countries), a strong predictor of learning overall and specific academic outcomes in multiple subjects (García-Madruga, Vila, Gómez-Veiga, Duque, & Elosúa, 2014; Melby-Lervåg & Lervåg, 2014b). Furthermore, because success in education is strongly related to future possibilities and accomplishments for students, promoting students' reading skills is crucial (Gustafsson et al., 2010). Thus, it is prudent to establish practices and systems for screening students for reading problems. This can support data-based decisions for early intervention, and progress monitoring of students' reading proficiency over time can determine whether interventions are having their intended impact.

The purpose of the present study is to examine the psychometric properties of the ORF measure and its relationship with high-stakes reading tests in a large sample of Norwegian students. The ORF measure used is based on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), a measure widely used in the United States. In this study, we adapted the measurement approach for use in Norwegian, a more transparent orthography than English. Only a small number of ORF studies have been conducted in languages other than English. A Spanish ORF measure, also adapted from DIBELS, has been studied in a U.S. educational context on Spanish-speaking immigrant students (D.L. Baker, Stoolmiller, Good, & Baker, 2011). Thus, these results are not very transferable to a European setting with mainly monolingual students. Although a variety of reading fluency measures are used in European countries (see, e.g., Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005; Veenendaal, Groen, & Verhoeven, 2015), no studies have been conducted using an adaption of the ORF measure based on DIBELS.

Reading Comprehension: The Ultimate Goal of Reading Proficiency

The ultimate purpose of reading is to extract meaning from text—in other words, to read with comprehension. Several theories have been suggested to explain the development of reading and reading comprehension (Cromley & Azevedo, 2007; Hoover & Gough, 1990; Kintsch, 1988; LaBerge & Samuels, 1974; Perfetti & Stafura, 2014). However, in elementary school students, the theoretical foundation known as the simple view of reading has the strongest empirical support (Gough & Tunmer, 1986; Hoover & Gough, 1990). According to this theory, reading comprehension is the product of the ability to decode words and sentences fluently, accurately, and with automaticity and being able to understand the meaning of these words in the context in which they are used. A number of studies have shown that decoding and listening comprehension can explain much of the variation in students' reading comprehension (for a review, see García & Cain, 2014). In fact, in a recent study using latent variables controlling for measurement error, the features of the simple view of reading explained as much as 94% of the variation among students, leaving little variation left to be explained by other variables (Foorman, Koon, Petscher, Mitchell, & Truckenmiller, 2015). Thus, learning to decode accurately and fluently, together with understanding the meaning of words, is paramount for developing well-functioning reading skills.

More specifically, *decoding skills* refers to the ability to accurately and automatically decipher the relationship between letters and sounds in words and sentences. *Reading fluency* is commonly defined as reading with accuracy, speed, and expression or prosody

(Rasinski, Reutzel, Chard, & Linan-Thompson, 2011; Schwanenflugel, Hamilton, Kuhn, Wisenbaker, & Stahl, 2004; Veenendaal et al., 2015). Recognizing and identifying words implies accurate decoding, but decoding is not necessarily dependent on knowing the meaning of the words, because it is possible to decode nonsense words or to decode real words but not understand the words' meanings. However, several foundational theories of decoding suggest that when a student knows the meanings of the words in a text and can activate this from his or her lexicon, words are more likely to be read automatically and fluently (Perfetti, 1985; Seidenberg & McClelland, 1989). This has also solid empirical support; it is easier to read fluently when you know the meanings of the words (see van IJzendoorn & Bus, 1994). Thus, the more automatic decoding skills are, the less attention needs to be used to assist in the decoding process. More resources will then be available to focus on comprehension.

In the development of decoding skills, students first learn to master decoding accuracy at the word level, then transfer these skills to passages and texts, and increasingly build reading fluency with connected text. As students get older, they learn to master accurate and fluent decoding skills both at the word and sentence levels (Landerl & Wimmer, 2008). When this is mastered as students get older, the effect of decoding on reading comprehension decreases, and language comprehension skills account for more of the variance in reading comprehension (García & Cain, 2014; Lervåg & Aukrust, 2010). Notably, cross-language studies have found differences in reading development between orthographies with different degrees of transparency (Caravolas et al., 2012; Caravolas, Lervåg, Defior, Seidlová Málková, & Hulme, 2013). Although the predictors of decoding are similar (Caravolas et al., 2012), the developmental pattern is different, and students learn to decode fluently more slowly in English, as compared with more transparent languages such as Spanish, Czech (Caravolas et al., 2013), and Finnish (Parrila et al., 2005).

ORF as a Measure of Reading Proficiency

An important question concerning the ORF measure has been its association with other measures of reading. There is strong theoretical support for reading fluency as a crucial component in reading comprehension. Pikulski and Chard (2005) described reading fluency as the bridge between decoding and reading comprehension. As mentioned previously, in the United States, the ORF measure is widely used to measure students' growth trajectories in decoding accuracy and automaticity with age-appropriate passages of connected text read aloud. A number of studies (e.g., S.K. Baker et al., 2008; Pikulski & Chard, 2005; Stoolmiller, Biancarosa, & Fien, 2013; Wise et al., 2010) have demonstrated strong correlations between reading fluency and reading comprehension (.60–.90).

Shinn et al. (2002) studied the association between the ORF measure and measures of decoding and of reading comprehension using confirmatory factor analysis. Third- and fifth-grade students were tested on reading tasks, including decoding phonetically regular words and pseudowords, answering literal and comprehension questions, completing cloze items, producing written retells of texts read, and ORF. For the third-grade sample, all measures made a significant contribution to a unitary, reading proficiency model. ORF measures correlated higher with the model than any of the other measures. For the fifth-grade sample, reading proficiency was best characterized as composed of two factors—decoding and comprehension—although these factors were very highly correlated (r = .83). The ORF measure fitted best with the decoding factor but also correlated higher with the comprehension factor than did the literal and inferential comprehension subtests of the Stanford Diagnostic Reading Test. Thus, the ORF measure provides a good index of reading proficiency, including comprehension (S.K. Baker et al., 2008).

The common conceptualization of the positive association between reading fluency and comprehension is that stronger fluency helps free up cognitive resources, which students can then direct toward constructing the meaning of the text. D.L. Baker and colleagues (2011) used structural equation modeling (SEM) to study whether reading with comprehension also has a positive effect on reading fluency. They also asked whether this possible influence might vary depending on the transparency of the language. To study this, reading data were collected in Spanish and English with second-grade English learners being taught to read in both languages. Results showed that ORF had an effect on reading comprehension, but reading comprehension also had an effect on reading fluency. In other words, the association was reciprocal. In addition, the pattern of the associations was the same in English and Spanish. The instructional implications suggest that reading comprehension instruction—teaching students to comprehend text—leads not only to comprehension benefits but also to reading fluency benefits.

Notably, there are also results showing that ORF is a better predictor of reading comprehension than decoding nonsense words (i.e., word attack), decoding real words in word lists, speed of word-reading measures (García & Cain, 2014; Wise et al., 2010), letter naming, vocabulary, or phoneme awareness is

(Kim, Petscher, Schatschneider, & Foorman, 2010). Furthermore, several studies have shown that there is a different set of predictors for decoding word lists versus decoding words accurately and fluently in connected text. The most plausible reason for this is that accurate and fluent text reading is more related to language comprehension, whereas reading decontextualized word lists rests primarily on phoneme awareness, rapid automatized naming, and letter knowledge (D.L. Baker et al., 2011; Hulme, Bowyer-Crane, Carroll, Duff, & Snowling, 2012; Stanovich, 2000). Therefore, when trying to account for students' reading proficiency when they are reading decontextualized word lists versus connected text, it is necessary to consider the reading task, distinguishing between reading word lists and reading words in connected texts (Veenendaal et al., 2015).

ORF as a Measure of Reading Growth Across Time

Another issue in ORF research has been the degree of reading fluency growth over time (Fuchs et al., 1993; Hasbrouck & Tindal, 1992, 2006) and the meaning of that growth in terms of improvements in overall reading proficiency (S.K. Baker et al., 2008). Hasbrouck and Tindal analyzed ORF data collected in the fall, winter, and spring of grades 2–5. Student performance increased over the course of the year as expected, and the cross-sectional data showed that students' reading fluency grew fastest in grades 2 and 3.

Although the majority of ORF studies have been concurrent or cross-sectional, some longitudinal studies have examined predictive relationships over time and estimated the increase in the numbers of words read per week. For instance, Fuchs et al. (1993) conducted the first longitudinal study on ORF. Different students were assessed in grades 1–6, but in each grade, the same students were tested repeatedly over time. Slope of performance was positive in each grade but decreased steadily across grades, consistent with findings reported by Deno et al. (1982) and Hasbrouck and Tindal (1992, 2006). This nonlinear pattern of rapid early growth and later slower growth has been replicated in other studies (S.K. Baker et al., 2008; Nese et al., 2013; Stage & Jacobsen, 2001). Speece and Ritchey (2005) showed that students with high rates of growth on ORF in grade 1 were more likely to maintain strong growth rates in grade 2 and read at grade level at the end of grade 2 than students who had low rates of growth. Using growth curve analysis, Speece and Ritchey also showed that students who were at risk for reading problems at the beginning of first grade had predicted ORF scores at the end of the year that were less than half the magnitude of their peers who were not at risk.

Several longitudinal studies in the United States have shown that growth in ORF is related to reading comprehension within and across school years and grades. For instance, S.K. Baker and colleagues (2008) investigated what unique contribution, if any, slope on ORF made to performance on comprehensive measures of reading. They investigated this with students in grades 1–3 who were tracked longitudinally for either 1.5 years (the middle of first grade to the end of second grade) or for two years (the beginning of second grade to the end of third grade). In each group, ORF data were collected five (first- and second-grade group) or six times (second- and third-grade group), in addition to a pretest and posttest on a comprehensive measure of reading (the SAT–10 or the state reading test). After controlling for initial status on the ORF measure and the comprehensive measure of reading at pretest, slope of ORF still added to the accuracy of predicting performance on the comprehensive measure of reading at posttest. Thus, progress in ORF was positively associated with improvement in reading proficiency. In grades 1–3, Wanzek, et al. (2010) found that ORF was a reliable predictor of student success on two high-stakes national and state-normed measures. Thus, several U.S. studies have provided strong support for the predictive validity of ORF for reading comprehension.

Because ORF is an important developmental indicator of reading proficiency and creates a foundation for reading comprehension (for a review, see Breznitz, 2006), monitoring reading fluency can help schools identify students at risk for reading failure (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Pfost, Dörfler, & Artelt, 2012). By understanding how reading fluency develops and how it in turn relates to reading comprehension, schools can give struggling readers targeted support in the early stages (S.K. Baker et al., 2008; Hosp & Suchey, 2014; Pikulski & Chard, 2005). When examining developmental processes in reading, latent growth curve models offer a particularly useful way to predict and explain change over time (Little, 2013; Rogosa, Brandt, & Zimowski, 1982; Stoolmiller, 1995). Most studies of growth in ORF have used first-order growth models with one indicator of ORF per timepoint. Our approach is to use multiple ORF indicators, which allows for not only a more thorough investigation of the measurement properties of the ORF reading passages in terms of longitudinal invariance but also the use of second-order latent growth models to better account for measurement error in the individual ORF scores (see, e.g., Widaman, Ferrer, & Conger, 2010). At the same time, a second-order latent growth model still also allows for an investigation of the interindividual differences in ORF starting levels, the interindividual differences in ORF growth across the school year, and the relation between these two individual difference factors among students.

Aim of the Study and Research Questions

The overall aim of our study was to examine initial status and growth in ORF and to investigate how ORF relates to students' reading performance on high-stakes national tests focusing on general reading proficiency (decoding and reading comprehension). Using a longitudinal design, we assessed students in grades 2–5 during one school year in Norwegian, a semitransparent orthography. The ORF measure was constructed by developing three unique grade-specific narrative and expository passages to be administered on three measurement occasions per school year (fall, winter, and spring). All passages were constructed to be parallel ORF items, similar in difficulty, but the actual content, in terms of the stories and information presented, differed to avoid practice effects. Assessing the longitudinal measurement invariance of the ORF passages allowed us to determine whether this objective was achieved.

Our study adds to the literature in several ways: As we have seen, the ORF measure is frequently used in the United States, and many studies have shown that it is a valid and reliable index of students' reading development (see, e.g., S.K. Baker et al., 2008; Deno et al., 1982; Fuchs et al., 1993; Good & Kaminski, 2002; Shinn, 1998; Stoolmiller et al., 2013). However, the ORF measure has never previously been adapted to a European setting where there is a variety of reading measures. With exceptions concerning bilingual Spanish-speaking students in the United States, an ORF measure based on DIBELS has not been used in transparent orthographies. Also, in Europe, the lack of psychometric validation of screening measures is a concern (SBU, 2014). Although it is crucial to examine at-risk students' progress from interventions over time, progress monitoring is not integrated in other European reading assessments. Finally, longitudinal invariance is taken for granted. However, it is difficult to design grade-level reading passages that are of comparable difficulty. If ignored, trends in the ORF measure across time might simply reflect a specific performance difference on a specific reading passage, instead of real progress and development.

To add to the previous literature, we will more specifically examine these four research questions:

1. Does ORF measure the same construct over time (i.e., demonstrate measurement invariance)?
2. How much growth do students experience on ORF over the course of the school year?
3. How stable is the rank order among students on ORF over time?
4. What is the association between the ORF measure and high-stakes tests of reading proficiency?

Method

Participants

A total of 2,228 students (48% female) participated in the study. The students were distributed across grades 2–5 in 21 schools across Norway in one school year (2012–2013). The schools were strategically selected to be representative of the Norwegian population. Therefore, they were located in both urban and rural districts across the country, and students from a variety of socioeconomic backgrounds were included. The number of students enrolled in each grade level ranged from four to 73 per school. Each grade level included 557 students on average, and 84% of them were monolingual. Furthermore, 11% of the students had two parents who were both bilingual, and 5% of the students had one parent who was bilingual.

Measures

An overview of the longitudinal study design and timing of the collected measures for each of the four grade levels (2–5) is given in Table 1. This also clarifies the range of predictive and concurrent relations between the ORF measure and the national tests in reading that are possible

TABLE 1
Overview of the Longitudinal Study Design and Timing of the Collected Measures for Grades 2–5

Year	2012–2013			2013–2014
Period	**Fall**	**Winter**	**Spring**	**Fall**
Grade 2		*Grade 2*		
ORF	1–3	4–6	7–9	
NTRP			Assessment	
Grade 3		*Grade 3*		
ORF	10–12	13–15	16–18	
NTRP			Assessment	
Grade 4		*Grade 4*		*Grade 5*
ORF	19–21	22–24	25–27	
NTRP				Test
Grade 5		*Grade 5*		
ORF	28–30	31–33	34–36	
NTRP	Test			

Note. NTRP = national tests in reading proficiency; ORF = oral reading fluency reading passages.

to assess in this study. Each school's assessment team had a data coordinator who was responsible for entering the data in an Excel spreadsheet established for this study.

ORF

The ORF measure and procedures are based on those of the ORF subtest drawn from the reading assessments, DIBELS sixth edition (Good & Kaminski, 2002). ORF was measured by three grade-specific narrative and expository passages on three measurement occasions at four-month intervals (fall, winter, and spring) during the 2012–2013 school year (see Appendix A for an overview). The range of words in each passage by grade varied: grade 2 = 190–207; grade 3 = 251–299; grade 4 = 297–310; and grade 5 = 300–326. Each passage was read aloud for one minute following standardized procedures. A trained teacher administered the ORF measure in an individual setting. Students were asked to read the passages aloud as accurately and as best they could until the teacher told them to stop. Students were told that if they got stuck, the teacher would tell them the word so they could keep reading. Words self-corrected within three seconds were scored as accurate. For each of the three passages, the number of words read correctly in one minute was the ORF raw score used in data analysis.

The ORF measures were administered individually to students by a teacher who was part of an assessment team that was established in each school for the purpose of the study. The assessment team consisted of expert teachers in reading, classroom teachers, or special teachers employed in the schools. All teachers who administered the ORF assessments received half-day training in the procedures of administration and scoring. For each grade level in this study, three new reading passages were administered at each measurement occasion. The full set of 36 reading passages (four grade levels × three passages × three occasions) were specifically developed in Norwegian for grades 2–5. Each set of the nine grade-level passages was constructed so each passage was similar to the others in the set in terms of purpose and passage characteristics, such as difficulty, length, and format. According to standard administration of ORF passages (Good & Kaminski, 2002), students who read fewer than 10 words correctly on the first of the three passages were not administered passages 2 and 3. In such cases, the ORF raw score for the latter two passages is not recorded and is therefore missing by design.

In this study, the alternate-form reliabilities were very high for all of the ORF passages within and across grades 2–5, ranging from .92 to .97 (see Table 2). This is in line with reliability findings in U.S.-based studies, where similar reliabilities have been reported as ranging from .89 to .97 for ORF measures (see, e.g., Cummings, Biancarosa, Schaper, & Reed, 2014; Good, Kaminski, & Dill, 2002; Stoolmiller et al., 2013).

National Tests of Reading Proficiency (NTRP)

In Norway, there are two types of national tests of reading proficiency administered to students in elementary school. The first type targets the early grades (1–3) and is a mandatory reading assessment for use in all Norwegian schools. It is group administered annually in the spring and aims to identify the need for support at both the individual and school levels. The second type is used in grade 5 only and functions as part of the quality assessment system for the Norwegian schools. This national test is group administered annually in the fall to all Norwegian students in grade 5. Each year, a new version of the NTRP is developed for both test types (Norwegian Reading Centre, 2013a, 2013b; Skaftun, Stangeland, Solheim, & Mangen, 2013; Solheim, Skaftun, & Walgermo, 2012). For this study, the annual updating of the NTRP implies that the measures differ across the four grade levels in terms of complexity.

For grade 2, the national reading assessment in spring 2013 consisted of the following seven subtests (see Appendix B for descriptions of the subtests): recognizing letters, writing words, reading words, splitting compound words, reading sentences, following written instructions, and reading text. For grade 3, the national reading assessment in spring 2013 consisted of the following four subtests (see Appendix B for descriptions): chains of words, reading narrative text, word knowledge, and reading expository text. Because no NTRP is available for grade 4, the national reading test score is based on the fall 2013 version of the test from when the fourth-grade students moved to grade 5; for grade 5, the score is based on the fall 2012 version (Cronbach αs based on official population data for the two fifth-grade tests are 0.86 and 0.86, respectively). The fifth-grade tests consisted of multiple texts to assess students' decoding and comprehension skills. Test formats included multiple-choice, closed-ended, and open-ended questions. Students had to find information in the texts, interpret the texts, and explain the meaning of them. The test used in 2012 consisted of 28 items, and the test used in 2013 consisted of 29 items.

Data Analysis

For each grade level (2–5), statistical models for data analysis were established in line with the longitudinal study design and within a SEM framework using the lavaan package (Rosseel, 2012) in the statistical software environment R. Full information maximum likelihood was used to handle missing data and make use of all available information for each individual. We applied robust (Huber–White) standard errors for all estimated parameters and a scaled goodness-of-fit chi-square for statistical inference. Model fit was evaluated based on commonly recommended goodness-of-fit indexes (Hu & Bentler,

TABLE 2
Descriptive Statistics for All Reading Passages (RP) That Form the Basis of the Oral Reading Fluency Measure Across Grades 2–5

Oral reading fluency: Grade 2									
	Fall (reliability = .97)			**Winter (reliability = .96)**			**Spring (reliability = .96)**		
	RP 1	**RP 2**	**RP 3**	**RP 4**	**RP 5**	**RP 6**	**RP 7**	**RP 8**	**RP 9**
M	37.78	44.21	37.63	53.81	52.12	53.43	69.18	63.88	62.21
SD	27.96	27.12	24.37	32.35	29.32	29.23	31.39	32.20	31.50
[min, max]	[0, 162]	[2, 162]	[3, 141]	[4, 167]	[3, 162]	[8, 166]	[4, 190]	[3, 187]	[3, 201]
Skewness	1.00	1.15	1.13	0.74	0.91	0.79	0.63	0.59	0.77
Kurtosis	3.91	4.28	4.22	2.87	3.70	3.37	3.17	3.05	3.58
n	411	373	372	466	462	461	459	459	458
Oral reading fluency: Grade 3									
	Fall (reliability = .96)			**Winter (reliability = .95)**			**Spring (reliability = .94)**		
	RP 10	**RP 11**	**RP 12**	**RP 13**	**RP 14**	**RP 15**	**RP 16**	**RP 17**	**RP 18**
M	66.72	73.82	69.64	84.19	82.66	82.54	96.20	88.89	95.46
SD	33.78	37.53	36.10	33.16	33.82	33.24	35.10	33.99	34.98
[min, max]	[0, 184]	[0, 198]	[0, 196]	[9, 190]	[10, 198]	[13, 220]	[13, 220]	[10, 229]	[11, 230]
Skewness	0.65	0.52	0.63	0.26	0.15	0.41	0.47	0.54	0.24
Kurtosis	3.25	2.86	3.00	2.94	2.68	2.93	3.25	3.63	3.42
n	435	434	432	472	472	472	471	471	471
Oral reading fluency: Grade 4									
	Fall (reliability = .93)			**Winter (reliability = .93)**			**Spring (reliability = .95)**		
	RP 19	**RP 20**	**RP 21**	**RP 22**	**RP 23**	**RP 24**	**RP 25**	**RP 26**	**RP 27**
M	98.70	109.61	95.59	100.10	122.14	102.86	123.62	128.59	112.90
SD	34.26	37.32	36.59	35.11	38.29	32.38	37.61	35.71	37.88
[min, max]	[4, 194]	[6, 214]	[5, 203]	[16, 204]	[17, 213]	[20, 199]	[22, 221]	[28, 232]	[25, 208]
Skewness	−0.20	−0.11	−0.12	0.21	−0.14	0.20	−0.08	−0.13	0.16
Kurtosis	2.77	2.76	2.84	2.69	2.74	2.90	2.82	3.23	2.57
n	475	475	475	532	533	532	443	443	441
Oral reading fluency: Grade 5									
	Fall (reliability = .92)			**Winter (reliability = .93)**			**Spring (reliability = .94)**		
	RP 28	**RP 29**	**RP 30**	**RP 31**	**RP 32**	**RP 33**	**RP 34**	**RP 35**	**RP 36**
M	107.95	103.52	120.00	117.88	126.11	128.49	122.59	118.37	117.79
SD	29.30	33.36	37.04	31.66	34.78	30.79	31.25	32.97	36.67
[min, max]	[19, 193]	[20, 184]	[25, 213]	[30, 205]	[24, 214]	[32, 224]	[26, 27]	[28, 227]	[15, 226]
Skewness	0.06	−0.10	−0.10	−0.09	−0.17	−0.25	−0.06	0.02	−0.03
Kurtosis	3.06	2.41	2.53	2.81	2.68	3.00	3.17	3.05	2.61
n	461	461	461	482	482	482	443	443	443

Note. *M* = mean; *SD* = standard deviation. Reliability was measured by calculating the mean of the correlations between the passages at the timepoint and the following timepoints.

1999), including the chi-square test of exact model fit, the root mean square error of approximation (RMSEA: ≤0.08 = acceptable, ≤0.05 = good) to assess close fit, the comparative fit index (CFI: ≥0.95 = good) contrasting to a null independence model, and the standardized root mean square residual (SRMR: ≤0.05 = good).

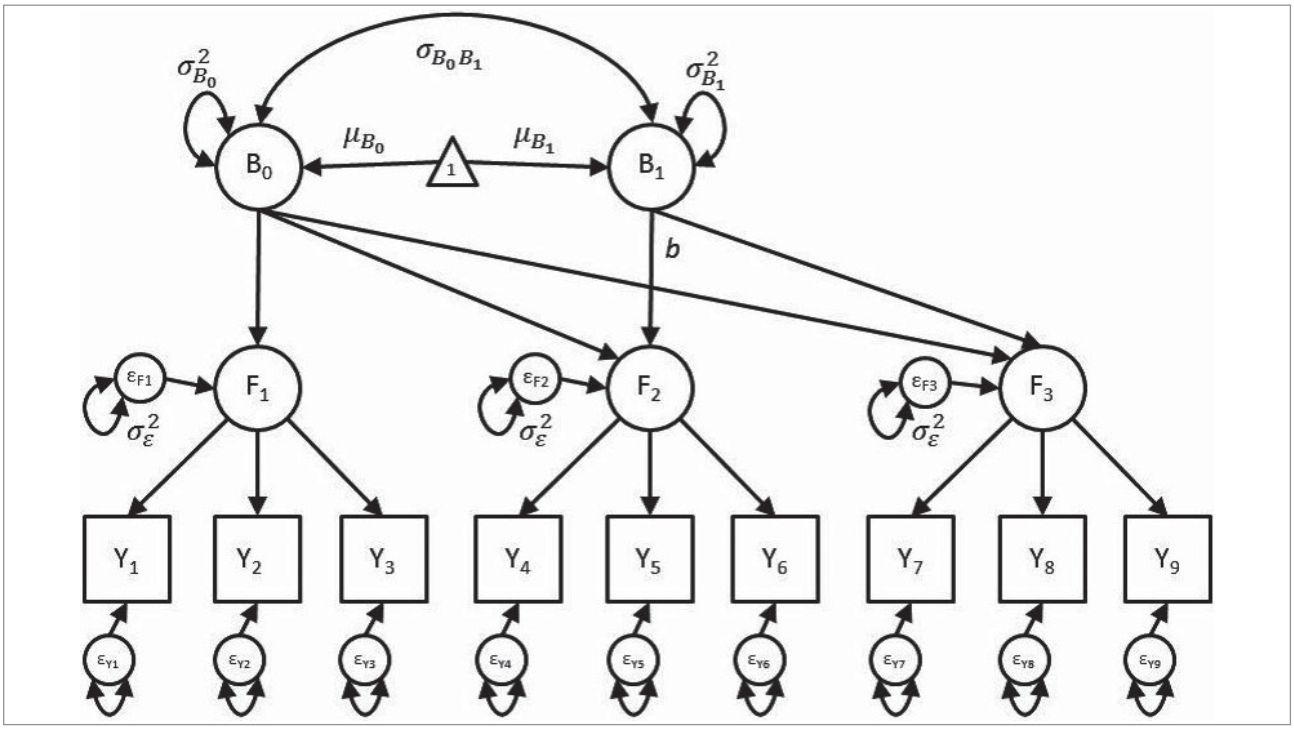
Longitudinal Measurement Invariance

Although the ORF reading passages were designed to be of comparable difficulty, we first verified this design feature for each grade level by investigating the longitudinal measurement invariance of the latent variable measurement model with all nine reading passages. We followed a model comparison procedure (for an overview, see Millsap, 2011) assessing the viability of restricting specific model parameters to be parallel invariant across the nine reading passages. When full invariance was not obtainable, we aimed to establish partial invariance by freeing up some invariance constraints under the condition that at each measurement timepoint, at least one reading passage was kept parallel invariant. The reason to pursue (at least partial) invariance is that otherwise differences in the ORF measure across time might simply reflect an idiosyncratic performance difference on a specific reading passage (e.g., an intended parallel-designed passage that unexpectedly turns out to be extremely difficult or easy in practice might overrule the general ORF trend across time). Model comparison is based on assessment of absolute goodness of fit and on the relative fit profile across the sequence of invariance models as indicated by differences in fit indexes such as the CFI (see, e.g., Cheung & Rensvold, 2002; Little 2013).

ORF Growth Models

Having established longitudinal (partial) invariance, a growth model is posited on top of the latent ORF factors in the measurement model. A path diagram of such a second-order latent growth model is given in Figure 1. The second-order latent growth model not only accounts for the measurement error in the individual ORF reading passage scores but also allows us to investigate the interindividual differences in ORF starting levels of the students in the same grade level (i.e., as indicated by the variance $\sigma^2_{B_0}$ of the random intercept factor B_0), the interindividual differences in ORF growth across the school year (i.e., as indicated by the variance $\sigma^2_{B_1}$ of the random slope factor B_1), and the relation between these two individual difference factors (i.e., as indicated by their covariance $\sigma_{B_0B_1}$). The mean parameter μ_{B_0} of the random intercept factor represents the average ORF starting level. To allow for a potential

FIGURE 1
Path Diagram of the Second-Order Latent Growth Model for the Nine (3 × 3) Parallel-Invariant Oral Reading Fluency Passages in Each of the Grades (2–5)



Note. Observed variables are represented by squares, latent variables by circles, and constants by triangles. The value of paths corresponding to nonannotated directed arrows is fixed at 1.

nonlinear growth trajectory, the loading for the in-between timepoint on the random slope factor B_1 is estimated freely, allowing the corresponding parameter b to be interpreted as the proportional change in ORF relative to the average change in ORF from the first timepoint to the last timepoint as represented by the mean parameter μ_{B_1} of the random slope factor. Variances of the residual time-specific ORF factors are constrained to be equal (i.e., parameter σ_ϵ^2).

Note that partial invariance would imply that some of the observed reading passage scores do not follow the general implied growth curve trend exactly and show a slightly differential pattern in either the observed mean score, as reflected in a nonzero intercept parameter (i.e., an additional direct path from the constant to one of the squares), or in the observed score (co)variance, as reflected in a freely estimated loading of the reading passage on its corresponding time-specific ORF factor.

The ORF Measure and the NTRP

The periodic changes in the Norwegian NTRP prevent a clear-cut comparison across time of the NTRP scores and of their link with the ORF measures. Yet, it results in the added benefit of having a variety of reading proficiency measures to evaluate the value of the ORF measure against it (see Table 1).

Missing Data Analysis

In general, missingness and dropout can be expected in every longitudinal study. Yet, its impact depends on whether data are systematically missing according to processes that can bias the measures of interest (e.g., only low-scoring students dropping out) or whether missingness is merely due to some random idiosyncratic events or planned because of the design. Random events here are relocation of students, students or administering teachers being absent due to illness, and practical administration issues preventing two schools from conducting data collection at the first timepoint for grade 2 and one school not completing data collection at the second and third timepoints in any grade. For grade 4 specifically, national reading examination tests were not available at the time of the ORF measure's administration but only one year after, when students moved to grade 5, such that there was less incentive for local data coordinators to follow through with delivering this extra set of NTRP data for all students.

Initial exploratory analyses indicate that having one or more missing scores at the later two timepoints is not related to performance at the first ORF timepoint. Given the low stakes of the ORF assessment, it is reasonable to assume that missingness is indeed random and not due to expected negative consequences of the ORF assessment for schools, teachers, or students involved. There is one design factor present: 37 of 528 students in grade 2 reading the first of three ORF passages at the first timepoint with less than 10 words read correctly per minute were exempted from the remaining two passages for that timepoint, in line with the ORF measure's administration protocol. Only one or two such cases occurred in later measurement occasions and in later grades.

A complete set of nine ORF scores was available for 308 students in grade 2 (58%; 30% missing between one and three scores, 12% missing more than three scores), 362 students in grade 3 (66%; 22% missing between one and three scores, 12% missing more than three scores), 413 students in grade 4 (70%; 7% missing between one and three scores, 23% missing more than three scores), and 384 students in grade 5 (68%; 11% missing between one and three scores, 21% missing more than three scores). National test scores in reading proficiency were available for 384 students in grade 2 (73%), 351 students in grade 3 (64%), 165 students in grade 4(28%), and 302 students in grade 5 (53%). The missingness in measures was partially overlapping, with 247 (47%), 280 (51%), 110 (19%), and 239 (42%) students having outcomes both on all ORF measures and on all NTRP in grades 2–5, respectively. This implies that when taking into account these practical data collection limitations, a very conservative estimate of the effective sample sizes in the different grades still amounts to about 250, which provides a large enough data coverage base for analysis of ORF–NTRP interrelations (for grade 4, standard errors can be expected to be slightly larger due to the relatively smaller complete overlap). An overview of the sample size for each measure across the year per grade is available in Tables 5–8.

Results

Descriptive Statistics

The ORF Measure

Descriptive statistics for all reading passages that form the basis of the ORF measure across the four grade levels are presented in Table 2. It is readily apparent that mean performance in the number of words read correctly per minute increases gradually within each grade level across the year and also across grades, although this pattern becomes less pronounced when comparing grades 4 and 5. The standard deviations within and across grades are rather similar and large, indicating a similar spread of scores across grades and measurement timepoints and large individual differences across students. Patterns of higher scores as students move up in grade with smaller differences at higher grades, and a relatively consistent spread among students across grades with somewhat larger standard deviations in the upper grades, is consistent with previous research on ORF. Less variability among standard deviations might

be interpreted positively because it shows similarity in the spread of scores. This finding is also consistent with previous studies on ORF in English-speaking students (see, e.g., S.K. Baker et al., 2008). Skewness and kurtosis statistics are within acceptable ranges for further SEM.

NTRP

Descriptive statistics for all NTRP across the four grade levels are presented in Table 3. The scores on the national reading test are higher in grade 4 than grade 5, but scores on the two versions of this test are not directly comparable because the content of the subtests changes from year to year. The sample descriptive statistics for the fifth graders on the national reading tests map closely to official population statistics, a finding that further supports the representativeness of the study sample. For the national reading assessments in the lower grades, no official statistics were available. For the NTRP scores, skewness and kurtosis statistics are also within acceptable ranges for further SEM, except for the first subtest in grade 2. Due to the clear ceiling effect on this "recognizing letters" measure (i.e., almost all students obtain the maximum score of 25), this subtest will not be considered in further analyses.

Longitudinal Measurement Invariance

Investigating change in ORF across time and interrelations across time with external variables such as those of the NTRP requires that we have measured the same ORF construct with the same metric at each occasion. Because three ORF scores are available at each occasion, we can explicitly evaluate this required longitudinal measurement invariance. If the ORF measurement instrument does not exhibit evidence of longitudinal invariance, then the interpretation of change in mean scores and correlations between timepoints may be ambiguous (Horn & McArdle, 1992).

Table 4 provides an overview of the measurement invariance model results, treating all reading passages within a grade level as parallel ORF indicators. In each grade (2–5), the configural reference model (Horn & McArdle, 1992; Little, 2013) provided an excellent goodness of fit to the data, reflecting that the nine ORF passages were measuring the same underlying construct. Restricting the loadings of the ORF passage scores to be equal across time had only a small impact on the resulting fit to the data. This implies that the assumption of metric invariance was met such that latent ORF scores can be considered to be expressed in the same units across time. Restricting the intercepts of the ORF scores to be equal across time had a dramatic impact on the resulting fit to the data. This indicates that although all reading passages were designed to be comparable in principle, there were particular passages that stood out empirically and biased the general trend in ORF latent means across time. Yet, by relaxing some of the restrictions for these differentially functioning reading passages, a well-fitting partial scalar invariance model was still obtained for every grade level, allowing for meaningful unambiguous comparisons and further analyses of ORF across time.

Growth in ORF

Having established longitudinal partial invariance in each grade level (2–5), a growth model was posited on top of the latent ORF factors in the measurement model (see Figure 1) of each grade level. The resulting second-order latent growth models showed good fit to the data: Grade 2: $\chi^2(33) = 81.17$, $p < .001$, CFI = 0.991, RMSEA = 0.053, $p = .335$, SRMR = 0.022; grade 3: $\chi^2(36) = 118.53$, $p < .001$, CFI = 0.984, RMSEA = 0.065, $p = .018$, SRMR = 0.041; grade 4: $\chi^2(32) = 157.28$, $p < .001$, CFI = 0.977, RMSEA = 0.082, $p = .082$, SRMR = 0.053; and grade 5: $\chi^2(31) = 314.00$, $p < .001$, CFI = 0.959, RMSEA = 0.094, $p = .128$, SRMR = 0.094.

Average Growth Trajectory

Figure 2 provides an overview of the estimated average ORF growth trajectory across grades 2–5 if we examine the results of the four grades together. The Norwegian students began the year reading an average of about 38, 66, 97, and 104 words correct per minute (WCPM) in grades 2–5, respectively. The average growth in number of WCPM was about 26, 31, 26, and 14 in grades 2–5, respectively. The average peak performance in the growth trajectories was 65, 97, 123, and 129 WCPM in grades 2–5, respectively.

The students in grade 2 started off reading about 38 WCPM (i.e., random intercept mean $\mu_{B_0} = 38.29$ [1.27], $p < .001$), which rapidly increased (i.e., random slope $\mu_{B_1} = 26.11$ [0.70], $p < .001$) across the year up to about 65 WCPM in the spring. The growth trajectory is approximately linear, with 57% (i.e., loading $b = 0.57$ [0.02]) of the total average change in ORF in grade 2 already occurring by winter. A similar pattern of results occurred in grade 3 ($\mu_{B_0} = 66.42$ [1.52], $p < .001$; $\mu_{B_1} = 31.48$ [0.77], $p < .001$; $b = 0.58$ [0.02]). In grade 4, the growth trajectory starts at about the same level ($\mu_{B_0} = 97.39$ [1.44], $p < .001$) as the spring ORF results for grade 3 but still shows continuing ORF growth ($\mu_{B_1} = 25.92$ [0.76], $p < .001$), although initially there is now a slower increase between fall and winter ($b = 0.15$ [0.03]), with an increase to spring accounting for 85% of the average total growth. In grade 5, the ORF growth trajectory seems to decrease ($\mu_{B_1} = 14.13$ [0.84], $p < .001$), with the initial average level in the fall for grade 5 ($\mu_{B_0} = 104.36$ [1.42], $p < .001$) being in the zone of the winter results for grade 4. The growth trajectory in grade 5 is no longer systematically increasing, with the peak ORF performance occurring in the winter ($b = 1.74$ [0.08]) and not in the spring as would be expected.

TABLE 3
Descriptive Statistics for National Tests of Reading Proficiency (NTRP) Across Grades 2–5

	Grade 2								Grade 3			Grade 4	Grade 5
	Recognizing letters	Writing words by listening and spelling	Reading words	Splitting compound words	Reading sentences	Following written instructions	Reading text	Chains of words	Reading narrative text	Word knowledge	Reading expository text	NTRP 2013	NTRP 2012
Timepoint	Spring								Spring			Fall of next year	Fall
M	24.71	13.46	15.43	12.92	14.53	7.86	3.35	24.29	5.33	14.97	3.99	21.22	18.57
SD	1.27	2.26	4.35	5.59	3.74	2.54	1.47	9.04	2.30	3.94	1.69	6.25	6.32
[min, max]	[24, 25]	[5, 21]	[4, 21]	[0, 21]	[2, 18]	[0, 10]	[0, 8]	[3, 65]	[0, 14]	[2, 20]	[0, 7]	[0, 33]	[0, 35]
Skewness	−6.08	−1.06	−0.37	0.01	−0.89	−1.30	−0.40	0.41	−0.18	−1.09	−0.01	−0.47	−0.40
Kurtosis	45.08	4.36	2.17	1.88	2.86	3.98	2.93	3.60	2.74	3.76	2.53	2.88	2.52
n	384	384	384	384	384	384	384	352	352	351	351	165	302
Official population statistics													
M												21.5	18.50
SD												6.70	6.20
Cronbach's α												0.86	0.86
N												55,272	54,296

Note. *M* = mean; *SD* = standard deviation. Due to the clear ceiling effect in the first test score of grade 2 (i.e., almost everyone obtains the maximum score), this "recognizing letters" measure will not be considered in further analyses.

TABLE 4
Oral Reading Fluency Longitudinal Measurement Invariance Results for Grades 2–5

Measurement invariance model	χ^2	*df*	*p*	Comparative fit index (CFI)	Root mean square error of approximation	*p*	Standardized root mean square residual	ΔCFI
Grade 2								
Configural	33	24	.102	0.998	0.027	.983	0.004	—
Metric	146	30	<.001	0.978	0.086	<.001	0.050	0.020
Scalar	562	36	<.001	0.902	0.167	<.001	0.064	0.096
Partial	64	31	<.001	0.994	0.045	.684	0.012	0.004
Grade 3								
Configural	17	24	.833	1.000	0.000	1	0.003	—
Metric	77	30	<.001	0.991	0.054	.307	0.036	0.009
Scalar	347	36	<.001	0.939	0.127	<.001	0.048	0.061
Partial	112	34	<.001	0.985	0.065	.021	0.037	0.015
Grade 4								
Configural	56	24	<.001	0.994	0.048	.571	0.005	—
Metric	186	30	<.001	0.971	0.094	<.001	0.063	0.023
Scalar	1,268	36	<.001	0.772	0.242	<.001	0.108	0.222
Partial	138	30	<.001	0.980	0.079	<.001	0.038	0.014
Grade 5								
Configural	47	24	.004	0.997	0.041	.785	0.005	—
Metric	328	30	<.001	0.957	0.133	<.001	0.093	0.040
Scalar	1,072	36	<.001	0.850	0.227	<.001	0.111	0.147
Partial	142	29	<.001	0.984	0.084	<.001	0.070	0.013

Note. In line with the intended oral reading fluency test design, the measurement invariance models treat all reading passages (RPs) as parallel items. Freed invariance constraints for the grade 2 partial model: Loading RP 3 and RP4 and intercept RP 2, RP 3, and RP 7; freed invariance constraints for the grade 3 partial model: Intercept RP 11 and RP 17; freed invariance constraints for the grade 4 partial model: Loading RP 23 and RP 24 and Intercept RP 20, RP 23, RP 26, and RP27; freed invariance constraints for the grade 5 partial model: Loading RP 28, RP 32, and RP 36 and intercept RP 30, RP 31, RP 32, and RP 34.

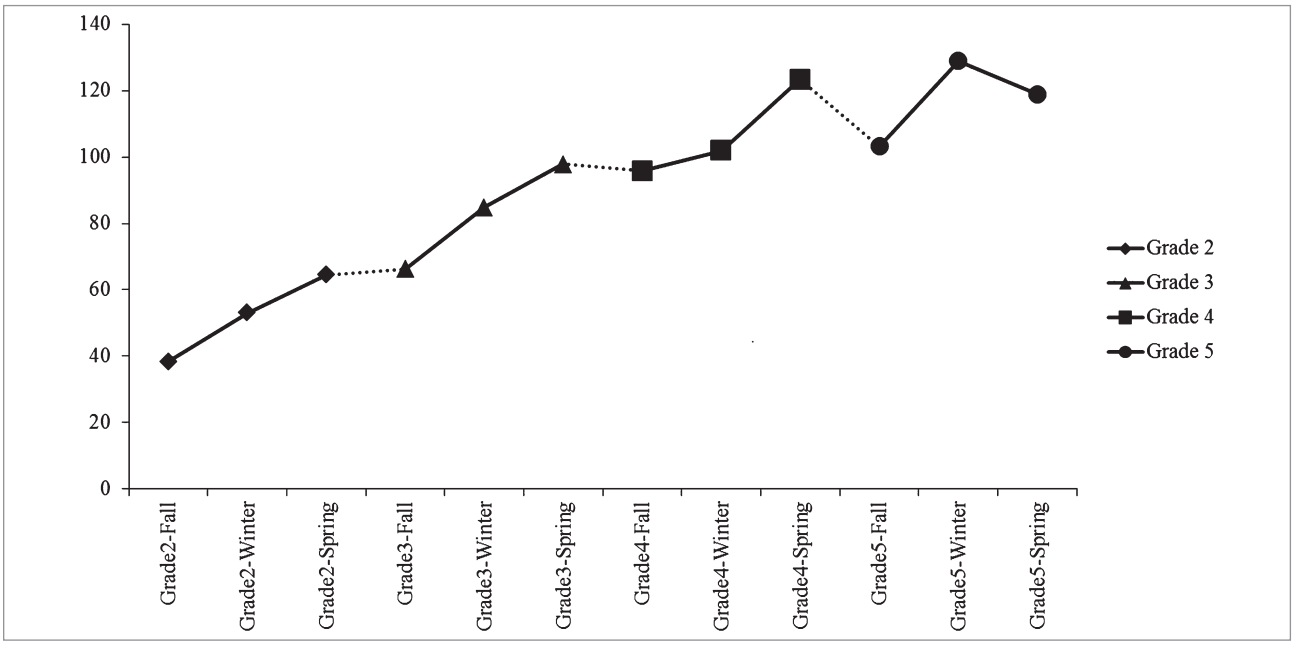
Individual Differences in ORF Development

The boxplots in Figure 3 provide an overview of the individual differences in estimated initial ORF levels and ORF growth rates (i.e., random intercept and slope, B_0 and B_1) in the four grade levels. As expected, initial levels ($\sigma^2_{B_0}$ = 706.39 [59.36], 1,145.34 [75.90], 1,088.59 [64.58], and 967.97 [61.65], respectively) vary much more than growth rates ($\sigma^2_{B_1}$ = 31.86 [31.30], 65.50 [29.61], 59.59 [19.24], and 1.45 [4.77], respectively) across individuals at all grade levels. Estimated population variation in the growth rate across individuals is larger in grades 3 and 4, whereas in grades 2 and 5, the variance could not be estimated very precisely and is smaller (grade 2) to almost nonexistent (grade 5). For grades 3 and 4, there is a small correlation between initial level and growth rate ($r_{B_0B_1}$ = –.16, p = .177; $r_{B_0B_1}$ = .289, p = .001). For grades 2 and 5, interpreting a correlation in the presence of a lack of variation of one of its components is not informative. The spaghetti plot in Figure 4 presents the resulting estimated individual growth trajectories. Consistent with students' natural development of reading skills, growth rates are positive for all individuals (sample minimum of estimated growth rates = 18.82, 14.54, 5.34, and 12.02 for grades 2–5, respectively).

Relative Stability of the ORF Measure and Concurrent and Predictive Relations Between the ORF Measure and the NTRP

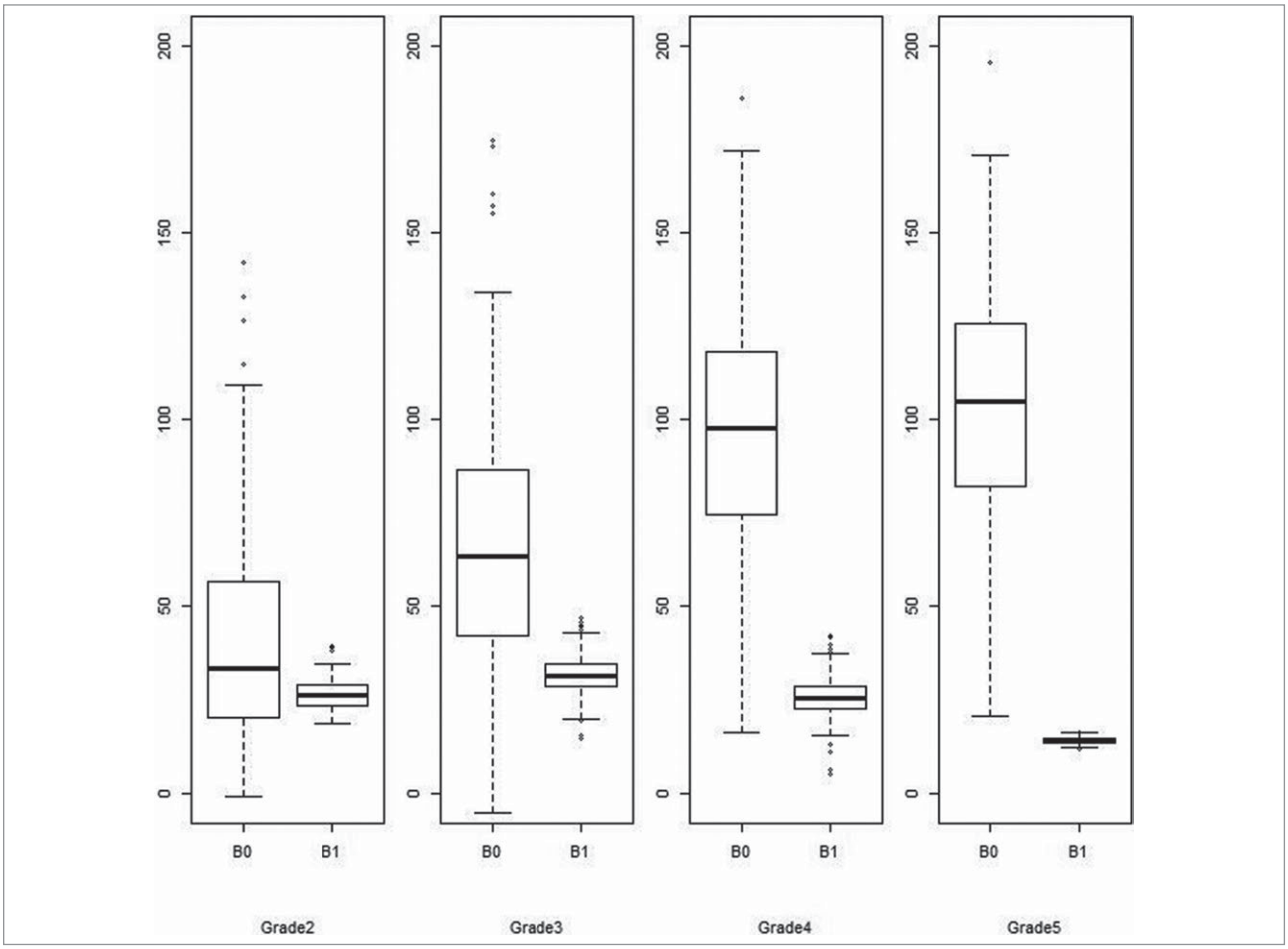
The relative stability of the ORF measure was high, as reflected by correlations of above .9 between the three

FIGURE 2
Oral Reading Fluency Growth Curve Across Grades 2–5



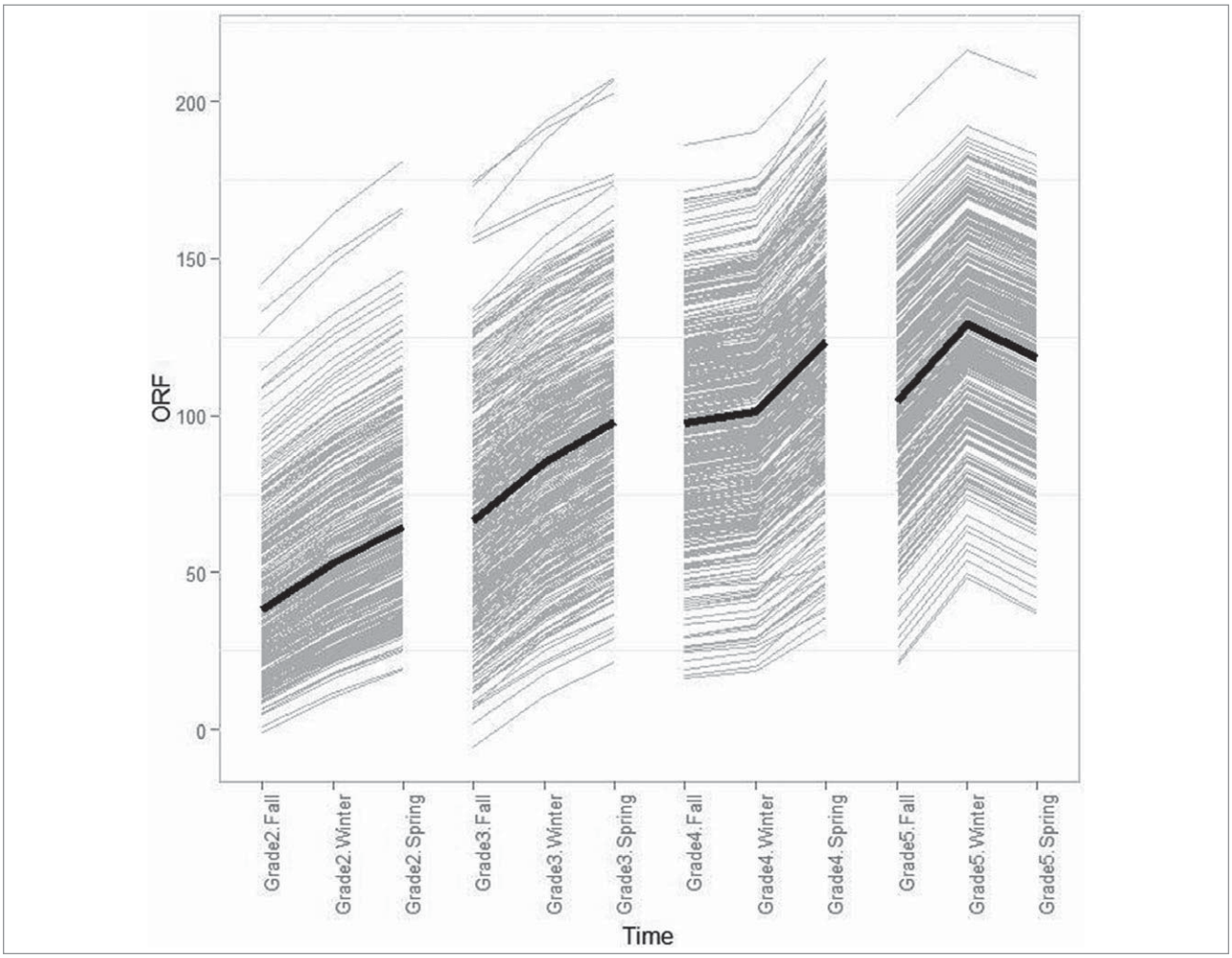
Note. The dotted line indicates the transition from oral reading fluency data based on one grade to another.

FIGURE 3
Individual Differences in Estimated Oral Reading Fluency Growth Parameters Across Grades 2–5



Note. The random intercept factors B_0 represent the starting level, and the random slope factors B_1 represent the growth rate.

FIGURE 4
Estimated Oral Reading Fluency (ORF) Growth Trajectories Across Grades 2–5



Note. Thin, gray lines represent individuals, and thick, black lines represent grade averages.

ORF factors across measurements within a grade (i.e., r_{F_{t-1},F_t} = .92–.94, .94–.96, .95–.97, and .95–.97, for grades 2–5, respectively). Hence, although ORF increases across the school year in an absolute sense, the relative rank ordering in terms of the students' ORF did not change much (see Figure 4).

The correlations between the ORF measures and the NTRP for grades 2–5 are shown in Tables 5–8. Grade 2 students were administered a national reading assessment consisting of seven subtests in the spring, which allows us to asses both predictive relations with the ORF measure (winter and fall measurement occasions) and concurrent relations with it (spring occasion). The first subtest, recognizing letters, is uninformative because almost all students earn the maximum score, and was consequently dropped from further analyses. The six remaining subtests, which required more elementary operations or targeted subskills needed for reading fluency, more strongly related to the ORF measures (reading words: r = .68, .69, and .73, respectively; splitting compound words: r = .69, .70, and .75, respectively; reading sentences: r = .63, .67, and .73, respectively) than did the subtests requiring more complex operations or higher level skills (writing words by listening and spelling: r = .48, .48, and .52, respectively; following written instructions: r = .57, .61, and .66, respectively; reading text: r = .49, .47, and .53, respectively). Concurrent correlations (i.e., the third r value indicative of the spring measurement occasion) were slightly larger than predictive relations (i.e., the first two r values indicative of the fall and winter measurement occasion), with a noticeable increase in the relation between the ORF measure and the reading sentences subtest.

Grade 3 students were administered a national reading assessment consisting of four subtests in the spring, which allows us to asses both predictive relations with the ORF measure (winter and fall measurement

TABLE 5
Correlations Between All Observed Measures With the Sample Size at All Timepoints for Grade 2

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. ORF 1	411															
2. ORF 2	**.97**	373														
3. ORF 3	**.97**	**.97**	372													
4. ORF 4	**.89**	**.89**	**.89**	466												
5. ORF 5	**.89**	**.90**	**.89**	**.97**	462											
6. ORF 6	**.89**	**.90**	**.90**	**.96**	**.97**	461										
7. ORF 7	**.89**	**.89**	**.88**	**.90**	**.90**	**.91**	459									
8. ORF 8	**.89**	**.88**	**.88**	**.91**	**.90**	**.91**	**.96**	459								
9. ORF 9	**.89**	**.89**	**.88**	**.91**	**.90**	**.91**	**.96**	**.97**	458							
10. NTRP 1	.10	.08	.07	.14	.13	.14	.18	.19	.17	384						
11. NTRP 2	.48	.42	.38	.48	.47	.47	.53	.52	.52	**.17**	384					
12. NTRP 3	.66	.61	.60	.68	.67	.67	.73	.72	.72	**.29**	**.46**	384				
13. NTRP 4	.67	.63	.63	.71	.69	.70	.74	.73	.75	**.20**	**.46**	**.75**	384			
14. NTRP 5	.62	.55	.56	.66	.65	.67	.72	.72	.71	**.27**	**.52**	**.77**	**.75**	384		
15. NTRP 6	.55	.49	.47	.60	.59	.59	.67	.65	.64	**.21**	**.46**	**.63**	**.60**	**.74**	384	
16. NTRP 7	.48	.38	.40	.47	.45	.46	.54	.53	.50	**.22**	**.46**	**.43**	**.40**	**.49**	**.55**	384

Note. NTRP = national tests of reading proficiency; ORF = oral reading fluency measure. The bold numbers are correlations within construct. Correlations greater than .13 are significant at the .05 level. The diagonal is the sample size at each measure.

TABLE 6
Correlations Between All Observed Measures With the Sample Size at All Timepoints for Grade 3

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13
1. ORF 1	435												
2. ORF 2	**.96**	434											
3. ORF 3	**.96**	**.96**	432										
4. ORF 4	**.91**	**.90**	**.90**	472									
5. ORF 5	**.90**	**.90**	**.90**	**.95**	472								
6. ORF 6	**.90**	**.90**	**.90**	**.95**	**.96**	472							
7. ORF 7	**.89**	**.89**	**.89**	**.91**	**.91**	**.91**	471						
8. ORF 8	**.88**	**.89**	**.89**	**.90**	**.91**	**.91**	**.94**	471					
9. ORF 9	**.88**	**.88**	**.89**	**.91**	**.91**	**.91**	**.95**	**.94**	471				
10. NTRP 1	.72	.71	.68	.72	.75	.72	.74	.70	.73	352			
11. NTRP 2	.43	.45	.45	.52	.50	.52	.48	.48	.50	**.39**	351		
12. NTRP 3	.27	.28	.28	.35	.34	.37	.34	.33	.34	**.29**	**.45**	351	
13. NTRP 4	.36	.39	.40	.45	.45	.46	.46	.45	.45	**.35**	**.49**	**.42**	351

Note. NTRP = national tests of reading proficiency; ORF = oral reading fluency measure. The bold numbers are correlations within construct. All correlations are significant at the .01 level. The diagonal is the sample size for each measure.

TABLE 7
Correlations Between All Observed Measures With the Sample Size at All Timepoints for Grade 4

Measure	1	2	3	4	5	6	7	8	9	10
1. ORF 1	476									
2. ORF 2	**94**	475								
3. ORF 3	**.93**	**.94**	475							
4. ORF 4	**.89**	**.90**	**.91**	532						
5. ORF 5	**.92**	**.92**	**.92**	**.93**	533					
6. ORF 6	**.88**	**.89**	**.90**	**.92**	**.93**	532				
7. ORF 7	**.90**	**.89**	**.89**	**.91**	**.92**	**.90**	443			
8. ORF 8	**.90**	**.89**	**.89**	**.90**	**.91**	**.90**	**.95**	443		
9. ORF 9	**.88**	**.88**	**.89**	**.91**	**.90**	**.91**	**.94**	**.94**	441	
10. NTRP	.52	.48	.50	.50	.48	.46	.48	.47	.50	165

Note. NTRP = national tests of reading proficiency; ORF = oral reading fluency measure. The bold numbers are correlations within construct. All correlations are significant at the .01 level. The diagonal is the sample size for each measure.

TABLE 8
Correlations Between All Observed Measures With the Sample Size at All Timepoints for Grade 5

Measure	1	2	3	4	5	6	7	8	9	10
1. ORF 1	462									
2. ORF 2	**.92**	461								
3. ORF 3	**.92**	**.93**	461							
4. ORF 4	**.88**	**.88**	**.89**	482						
5. ORF 5	**.88**	**.89**	**.89**	**.93**	482					
6. ORF 6	**.88**	**.90**	**.88**	**.92**	**.94**	482				
7. ORF 7	**.88**	**.87**	**.89**	**.91**	**.91**	**.90**	443			
8. ORF 8	**.87**	**.87**	**.88**	**.91**	**.91**	**.91**	**.95**	443		
9. ORF 9	**.89**	**.90**	**.90**	**.92**	**.92**	**.92**	**.94**	**.94**	443	
10. NTRP	.55	.54	.55	.49	.53	.53	.49	.51	.54	302

Note. NTRP = national tests of reading proficiency; ORF = oral reading fluency measure. The bold numbers are correlations within construct. All correlations are significant at the .01 level. The diagonal is the sample size for each measure.

occasions) and concurrent relations with it (spring). Relations between the ORF measure and the simple chains of words subtest (r = .73, .75, and .75, respectively) were stronger than with the three other subtests that targeted higher level reading skills (reading narrative text: r = .49, .53, and .51, respectively; reading expository text: r = .44, .47, and .47, respectively) and vocabulary (word knowledge: r = .31, .36, and .35, respectively).

The fifth-grade national reading test was completed by grade 5 students at the time of the ORF measure's administration in the fall (concurrent relation) and by grade 4 students in the fall of the year after the ORF measure's administration when the students had moved to grade 5 (predictive relation). The concurrent and predictive relations between the ORF measure and the two fifth-grade samples on the national test are estimated to be .55 and .54, respectively. These somewhat lower correlations are in line with expectations, given that the national reading test focuses on higher level reading competences, such as finding information, reading comprehension, text interpretation, and reflection. An increase in ORF scores during grade 5 occurs, but the rate of acceleration is flatter.

Discussion

The main purpose of the study was to examine initial status and growth on ORF and how ORF relates to students' reading performance in and across multiple

grades on two high-stakes national compulsory reading tests focused on decoding and comprehension in Norwegian, a semitransparent orthography. The study differed from previous research in that it examined the use of an ORF measure in a European context with a more transparent language than English using longitudinal second-order latent growth curve modeling. Overall, this study makes several contributions to the existing research regarding growth trajectories on the ORF measure and its relations with general reading proficiency.

First, we examined whether the ORF measure has longitudinal measurement invariance. This is important because invariance in text difficulty and complexity can help determine that the growth trajectories are due to the reading development and not passage characteristics. We found that the configural reference model for the ORF passages was measuring the same underlying construct. This was predicted because all passages were constructed to be parallel items, similar in difficulty but with different content (stories) to avoid retest effects. However, when the intercepts of the ORF scores were restricted to be equal across time, particular passages stood out empirically as more or less difficult and affected the general trend in ORF latent means. By relaxing some of the restrictions for these reading passages, a well-fitting partial scalar invariance model could still be obtained for each grade level (2–5). This allowed for meaningful, unambiguous comparisons and further analyses of the ORF measure across time. We explored potential reasons for why some passages differed empirically from the others by linking the deviations in intercept (mean) and loading of the passages to (a) technical measures of readability (e.g., LIX readability formula: Gilliland, 1972; Flesch readability formula: Flesch & Paterson, 1948) and (b) the content type and topic of the reading passage (see Appendix A). However, we did not find a link with the technical measures nor with content type.

The invariance results support previous research and show the difficulty of creating parallel reading passages (e.g., Cummings, Park, & Bauer Schaper, 2013). Although the grade-level passages are developed to be equal in difficulty, the reason why some passages stood out empirically might be that different passages mirror students' interests and familiarity because of content variation in the passages and the types of text structures used (e.g., narrative, expository). Our findings underline the importance of measuring ORF across a set of reading passages instead of basing results on only a one-passage measure. In fact, an observed median score of the three passages at each measurement point is recognized as a better indicator of a student's ORF performance than just one passage (DIBELS: Good & Kaminski, 2002).

Second, we examined students' growth in ORF. As for the examination of growth trajectories in a semitransparent language, the main findings—that linear growth represented grades 2 through 3, and more nonlinear growth represented grades 4 and 5—were as expected and extend previous research in students' growth of reading fluency. The findings that the Norwegian students' performance increased over the course of the year and fastest in grades 2 and 3 are similar to previous studies (see, e.g., Hasbrouck & Tindal, 1992, 2006). The slower growth in grade 5 might be interpreted as an indication of reaching a performance ceiling, especially for some students, and hence a flattening out in level of ORF rate in Norwegian. However, further studies including students in higher grades should be conducted to determine this.

The nonlinear growth pattern is supported by previous research of ORF growth in English for students in the later grades (e.g., S.K. Baker et al., 2008; Fuchs et al., 1993; Nese et al., 2013). Furthermore, it is consistent with theory and research regarding the development of automaticity in reading (LaBerge & Samuels, 1974). The nonlinear growth pattern indicates that as reading is first developing, changes in fluency are reflecting that the decoding process is becoming more automatic. As students become more proficient in reading, individual differences seem to deal more with reading comprehension of the particular passage than with reading comprehension in general (e.g., García & Cain, 2014; Pikulski & Chard, 2005; Stanovich, 2000). This does not necessarily mean that administering the ORF measure in grade 5 is not useful but that expectations for linear growth might be unrealistic in practice (Nese et al., 2013).

The context for this study is that reading fluency was measured in a semitransparent orthography, whereas most other studies of ORF measure reading fluency in English, which has a less transparent orthography. Previous studies comparing the development of reading fluency between students from different orthographies have found that there are similar mechanisms and predictors underlying the development of decoding but that students learning to read in English have slower decoding growth rates, at least during the first three years of school (Caravolas et al., 2013). If we compare our findings with growth rates found in studies of English readers, we see that the Norwegian students began the year reading fewer WCPM but experience stronger growth during the school year. For instance, Tindal, Nese, Stevens, and Alonzo (2015) found that U.S. students in grade 3 began the year reading an average of nearly 81 WCPM (15 WCPM more than Norwegian students), just above 100 WCPM in grade 4 (three WCPM more than Norwegians), and 125 WCPM in grade 5 (21 WCPM more than Norwegians).

The slope in each of these grades ranged between 0.65 and 0.73 WCPM, whereas the slope in the Norwegian grades 3–5 ranged between 0.76 and 0.84 WCPM.

The lower initial value of WCPM in Norwegian students might be due to the fact that learning to read starts at a later age than for students in the U.S. school system. However, the stronger ORF growth in Norwegian students supports previous research findings that learning to read in a more transparent language is easier, particularly in the early stages of reading development, than learning to read in a nontransparent orthography (Caravolas et al., 2013). Another possible explanation is that reading instruction during the school year in Norwegian elementary schools is somehow different from other contexts and more aligned with practices that accelerate reading fluency growth. Another important factor is that U.S. students start their formal reading instruction one year earlier than Norwegian students.

As expected, individual differences in initial ORF level varied much more than growth rates. However, all individuals had positive growth rates. The estimated population variation in growth across individuals was largest in grades 3 and 4, and the correlation between initial level and growth in these grades was small. In grades 2 and 5, the variance in growth rates across individuals was small and almost nonexistent. One interpretation is that the effects of reading instruction are constant for students across different reading-proficiency levels and that the students are relatively homogeneous in terms of their reading proficiency. The initial differences might be useful as a baseline indicator to identify students at risk for reading problems, which was demonstrated in several previous studies (e.g., Silberglitt & Hintze, 2007; Speece & Ritchey, 2005; Wang, Porfeli, & Algozzine, 2008).

Based on early screening to identify struggling readers, Parrila et al. (2005) demonstrated that it is possible for teachers to reduce individual differences in basic reading skills during early reading development. Teachers can respond early to individual differences among students with specific interventions, followed by systematic monitoring of students' growth (Stecker, Fuchs, & Fuchs, 2005). Although variability in ORF growth should be expected, results of the present study and others can be used to define normative rates of growth that can help identify students with low initial ORF levels and/or slow ORF growth so they can receive more intensive reading support. Furthermore, it will be important to define ORF benchmark and/or cutoff scores to identify struggling readers for intervention and progress monitoring. However, thresholds for appropriate levels of automaticity and reading rate by grade level might best be considered by using receiver operating characteristic curves using generated specifications related to sensitivity and specificity. Establishing thresholds using professional judgment and various objective approaches (e.g., students scoring below the 20th percentile) might help teachers identify the "right" students for intervention or extra support, but the sensitivity of yielding "true positives" and the specificity of yielding "true negatives" is also important to consider (Smolkowski, Cummings, & Strycker, 2016).

Finally, we investigated the relative stability of ORF growth and the concurrent and predictive relations between the ORF measure and the NTRP. The high stability of the ORF measure, reflected by correlations between the three ORF factors at all three timepoints in all four grade levels (>.92) in a semitransparent orthography, confirms previous evidence generated in less transparent orthographies (e.g., S.K. Baker et al., 2008; Kim et al., 2010; Nese et al., 2013). Overall, the findings extend evidence to more transparent languages by demonstrating moderate to strong positive correlations between the ORF measure and the NTRP across the school year in grades 2–5 (range = .44–.75). The findings are in line with previous studies on correlations between the ORF measure and high-stakes criterion measures of reading in English. However, high-stakes criterion measures among state tests and national tests have varying levels of difficulty and psychometric quality, of course. For instance, Wanzek et al. (2010) demonstrated in a longitudinal study of predictive validity across grades 1–3 in the United States that the ORF measure was a reliable predictor of students' reading proficiency on two different high-stakes measures in grade 3. However, greater student growth on the ORF measure through the three grades was needed to achieve success on the nationally normed test (SAT-10) compared with what was needed on the state-normed test (Texas Assessment of Knowledge and Skills).

Furthermore, in a study across grades 1–3, S.K. Baker et al. (2008) found slightly different correlations between the ORF measure and two high-stakes measures, the SAT-10 (range = .63–.80) and the Oregon state test (range = .58–.68). Findings that the ORF measure provided a stronger relation to the NTRP in earlier grades than in later grades in Norway have also been confirmed by previous research on how the ORF measure relates to high-stakes criterion measures in reading in a less transparent orthography and how relations between ORF and reading performance decrease over time (e.g., S.K. Baker et al., 2008).

Regarding implications for practice, our study showed that the ORF measure is an important developmental indicator of reading proficiency and is useful in monitoring students' reading fluency, which can help schools identify students who are at risk for reading failure (Fuchs et al., 2001; Pfost et al., 2012). In a prevention and early intervention framework of reading

development, ORF is an efficient measure that schools can use to help teachers efficiently identify students who are on track and those who are not. This can lead to providing struggling readers with targeted support in the early stages of reading development, when their growth trajectories in some areas, such as reading fluency, tend to be developing rapidly (S.K. Baker et al., 2008; Hosp & Suchey, 2014; Pikulski & Chard, 2005). By identifying struggling readers early and following their development within the year, teachers can initiate early reading supports and do not have to wait for high-stakes tests, such as the NTRP results, at the end of the school year. It is worth noting that the absence of the NTRP in Norwegian in grade 4 probably increases the risk of not identifying struggling readers. That is, two years can elapse before reading data are provided. In summary, we conclude that the Norwegian version of the ORF measure is a reliable and valid screening instrument that is easy and efficient to administer in schools and contributes to the early identification of students at risk for reading difficulties across years in the elementary grades.

Because poor reading skills can be a significant impediment to success in formal education, interventions are crucial. As shown in the present study, and also demonstrated in previous studies, the ORF measure can serve as an index of students' reading development, not only as a measure of reading fluency per se. Many studies have shown that reading fluency problems can be effectively remediated through repeated reading interventions (for a review, see Chard, Vaughn, & Tyler, 2002; National Institute of Child Health and Human Development, 2000). However, these kinds of repeated reading interventions can lead to reading instruction practices where the focus is on reading for speed and where other important components are excluded (Rasinski et al., 2011; Rayner, Schotter, Masson, Potter, & Treiman, 2016). Thus, because the ORF measure also serves as an index of reading problems beyond reading fluency, repeated reading should not be the only intervention for these students. Many studies have shown that the best way to ensure strong comprehension and with a sufficient reading speed is to also work on vocabulary, in line with the simple view of reading. However, learning to read is a complex process that includes several aspects beyond the simple view (e.g., sociocultural, neurological, genetic). For instance, from a sociocultural point of view, Purcell-Gates (2002) argued that the simple view of reading is not without controversy. Because students enter different school contexts from different socioeconomic backgrounds, they will face different learning and reading difficulties. Ultimately, it is important not only to identify each student's specific needs by analyzing potential needs from different perspectives but also to differentiate the interventions appropriately so all students are supported effectively.

Also, substantial research has shown that for young students struggling with learning to read, small-group interventions that emphasize all major aspects of reading development (phonics, fluency, comprehension, and vocabulary) consistently produce benefits in measured aspects of reading, including comprehension and fluency (Hulme & Melby-Lervåg, 2015; Melby-Lervåg & Lervåg 2014a). Furthermore, even for students with fluency problems solely, interventions could focus on fluency, including instructions for comprehension and prosody. Notably, students with dyslexia with no additional language problems will also get a low score on the ORF measure but will not necessarily need vocabulary and language comprehension training as a part of their intervention. It is therefore important to monitor progress and provide more in-depth diagnostic assessments to determine in a more precise way what specific areas of difficulty a student is having trouble with. Thus, teachers can distinguish between using the ORF measure for screening and progress-monitoring assessments and when additional assessment data are needed for other purposes, such as determining specific program areas and intervention content. Furthermore, additional diagnostic data will be necessary when more intensive interventions for students are needed because they are not responding sufficiently to universal or less intense interventions (e.g., vocabulary, decoding).

Future Studies

In future studies, it could be useful to validate the ORF measure in a transparent orthography against a larger battery of diagnostic reading tests in addition to more general national or statewide assessment tests. By validating the ORF measure against a battery of individually administered reading tests, it would be possible to examine how sensitive and specific the ORF test is when it comes to detecting reading problems at an early stage (see Duff et al., 2015; Snowling & Hulme, 2012). The ORF measure might also be validated against other groups of students with known characteristics (e.g., dyslexia, individualized education plans in reading). Although we did not find significant differences on the ORF measure by student group based on variables such as gender or in the small group of bilingual students in this study, these variables in addition to students' socioeconomic status will be important to investigate more thoroughly in future studies.

The ORF measure as implemented in this study was a teacher-administered measure, which can affect data collection quality based on teachers' prior knowledge of students, potential bias against or in favor of the measure or specific students, or potential bias in relation to the use of test data (e.g., accountability, teacher

evaluations). Thus, the ORF measure should be validated against tests that are not administered or collected by teachers, or when possible, ORF data should be collected by impartial examiners. However, because the ORF measure is intended as a teacher instrument, it is important to also have teachers as test administrators. Still, in future studies, inter-rater reliabilities should be included. Finally, in both Europe and the United States, large-scale full-classroom assessments have also been met by considerable controversy among educators (e.g., Goodman, 2006; National Union of Teachers, 2012). Hence, usability and usefulness of the ORF measure from a teacher perspective needs to be evaluated. To our knowledge, this type of study has not yet been conducted. In contrast to the mandatory tests used in schools, the ORF measure is not a one-time check but a tool that can be integrated into teachers' work throughout the year to measure students' progress. This is potentially more useful because it can be more directly linked to intervention and used as a measure of progress.

NOTE

We gratefully acknowledge the students and their teachers for participating in the study.

REFERENCES

Baker, D.L., Stoolmiller, M., Good, R., & Baker, S. (2011). Effect of reading comprehension on passage fluency in Spanish and English for second-grade English learners. *School Psychology Review*, *40*(3), 331–351.

Baker, S.K., Smolkowski, K., Katz, R., Fien, H., Seeley, J.R., Kame'enui, E.J., & Beck, C.T. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review*, *37*(1), 18–37.

Baker, S.K., Smolkowski, K., Smith, J.M., Fien, H., Kame'enui, E.J., & Beck, C.T. (2011). The impact of Oregon Reading First on student reading outcomes. *The Elementary School Journal*, *112*(2), 307–331. doi:10.1086/661995

Breznitz, A. (2006). *Fluency in reading: Synchronization of processes*. Mahwah, NJ: Erlbaum.

Caravolas, M., Lervåg, A., Defior, S., Seidlová Málková, G., & Hulme, C. (2013). Different patterns, but equivalent predictors, of growth in reading in consistent and inconsistent orthographies. *Psychological Science*, *24*(8), 1398–1407. doi:10.1177/0956797612473122

Caravolas, M., Lervåg, A., Mousikou, P., Efrim, C., Litavsky, M., Onochie-Quintanilla, E., … Hulme, C. (2012). Common patterns of prediction of literacy development in different alphabetic orthographies. *Psychological Science*, *23*(6), 678–686. doi:10.1177/0956797611434536

Chard, D.J., Vaughn, S., & Tyler, B.-J. (2002). A synthesis of research on effective interventions for building fluency with elementary students with learning disabilities. *Journal of Learning Disabilities*, *35*(5), 386–406. doi:10.1177/00222194020350050101

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

Cromley, J.G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325. doi:10.1037/0022-0663.99.2.311

Cummings, K.D., Biancarosa, G., Schaper, A., & Reed, D.K. (2014). Examiner error in curriculum-based measurement of oral reading. *Journal of School Psychology*, *52*(4), 361–375. doi:10.1016/j.jsp.2014.05.007

Cummings, K.D., Park, Y., & Bauer Schaper, H.A. (2013). Form effects on DIBELS Next Oral Reading Fluency progress-monitoring passages. *Assessment for Effective Intervention*, *38*(2), 91–104. doi:10.1177/1534508412447010

Deno, S.L., Mirkin, P.K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*(1), 36–45.

Duff, F.J., Mengoni, S.E., Bailey, A.M., & Snowling, M.J. (2015). Validity and sensitivity of the phonics screening check: Implications for practice. *Journal of Research in Reading*, *38*(2), 109–123. doi:10.1111/1467-9817.12029

Flesch, R., & Paterson, D.G. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. doi:10.1037/h0057532

Foorman, B.R., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology*, *107*(3), 884–899. doi:10.1037/edu0000026

Fuchs, L.S., Fuchs, D., Hamlett, C.L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review*, *22*(1), 27–48.

Fuchs, L.S., Fuchs, D., Hosp, M.K., & Jenkins, J.R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*(3), 239–256. doi:10.1207/S1532799XSSR0503_3

García, J.R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, *84*(1), 74–111. doi:10.3102/0034654313499616

García-Madruga, J.A., Vila, J.O., Gómez-Veiga, I., Duque, G., & Elosúa, M.R. (2014). Executive processes, reading comprehension and academic achievement in 3th grade primary students. *Learning and Individual Differences*, *35*, 41–48. doi:10.1016/j.lindif.2014.07.013

Gilliland, J. (1972). *Readability*. London, UK: University of London Press.

Good, R.H., & Kaminski, R.A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Education Achievement.

Good, R.H., Kaminski, R.A., & Dill, S. (2002). DIBELS oral reading fluency and retell fluency. In R.H. Good & R.A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed., pp. 30–38). Eugene, OR: Institute for the Development of Education Achievement.

Goodman, K.S. (with Flurkey, A., Kato, T., Kamii, C., Manning, M., Seay, S., Thome, C., … Wilde, S.). (2006). *The truth about DIBELS: What it is, what it does*. Portsmouth, NH: Heinemann.

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

Gustafsson, J.E., Allodi Westling, M., Alin Åkerman, B., Eriksson, C., Eriksson, L., Fischbein, S., … Persson, R.S. (2010). *School, learning and mental health: A systematic review*. Stockholm, Sweden: The Royal Swedish Academy of Sciences, The Health Committee.

Hasbrouck, J., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children*, *24*(3), 41–44. doi:10.1177/004005999202400310

Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher*, *59*(7), 636–644. doi:10.1598/RT.59.7.3

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

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. doi:10.1080/03610739208253916

Hosp, J., & Suchey, N. (2014). Reading assessment: Reading fluency, reading fluently, and comprehension—Commentary on the special topic. *School Psychology Review*, *43*(1), 59–68.

Hu, L.T., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi:10.1080/10705519909540118

Hulme, C., Bowyer-Crane, C., Carroll, J.M., Duff, F.J., & Snowling, M.J. (2012). The causal role of phoneme awareness and letter-sound knowledge in learning to read. *Psychological Science*, *23*(6), 572–577. doi:10.1177/0956797611435921

Hulme, C., & Melby-Lervåg, M. (2015). Effects from interventions for psychological learning and behavioural disorders in children. In A. Thapar, D.S. Pine, J.F. Leckman, S. Scott, M.J. Snowling, & E. Taylor (Eds.), *Rutter's child and adolescent psychiatry* (6th ed., pp. 533–545). Oxford, UK: John Wiley & Sons.

Kim, Y.S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, *102*(3), 652–667. doi:10.1037/a0019643

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review*, *95*(2), 163–182. doi:10.1037/0033-295X.95.2.163

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

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

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology*, *100*(1), 150–161. doi:10.1037/0022-0663.100.1.150

Lervåg, A., & Aukrust, V.G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *The Journal of Child Psychology and Psychiatry*, *51*(5), 612–620. doi:10.1111/j.1469-7610.2009.02185.x

Little, T.D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford.

Melby-Lervåg, M., & Lervåg, A. (2014a). Effects from educational interventions on reading comprehension and its underlying components. *Child Development Perspectives*, *8*(2), 96–100. doi:10.1111/cdep.12068

Melby-Lervåg, M., & Lervåg, A. (2014b). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, *140*(2), 409–433. doi:10.1037/a0033890

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

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, reports of the subgroups* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

National Union of Teachers. (2012). *'Five years old is too young to fail!': Year one phonics screening check report*. Retrieved from https://www.teachers.org.uk/files/phonics-survey-conference-2012-apr-12-je.doc

Nese, J.F.T., Biancarosa, G., Cummings, K., Kennedy, P., Alonzo, J., & Tindal, G. (2013). In search of average growth: Describing within-year oral reading fluency growth across grades 1–8. *Journal of School Psychology*, *51*(5), 625–642. doi:10.1016/j.jsp.2013.05.006

Norwegian Reading Centre, University of Stavanger. (2013a). *Kartleggingsprøve i lesing på 2.trinn—forslag til endelig prøve* [Mandatory reading assessment in grade 2—proposed final test; Technical report]. Stavanger, Norway: Author.

Norwegian Reading Centre, University of Stavanger. (2013b). *Kartleggingsprøve i lesing på 3.trinn—forslag til endelig prøve* [Mandatory reading assessment in grade 3—proposed final test; Technical report]. Stavanger, Norway: Author.

Organisation for Economic Co-operation and Development (2013). *OECD economic surveys: France 2013*. Paris, France: Author.

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J.-E., & Kirby, J.R. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology*, *97*(3), 299–319. doi:10.1037/0022-0663.97.3.299

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

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

Pfost, M., Dörfler, T., & Artelt, C. (2012). Reading competence development of poor readers in a German elementary school sample: An empirical examination of the Matthew effect model. *Journal of Research in Reading*, *35*(4), 411–426. doi:10.1111/j.1467-9817.2010.01478.x

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

Purcell-Gates, V. (2002). The irrelevancy—and danger—of the 'simple view' of reading to meaningful standards. In R. Fisher, M. Lewis, & G. Brooks (Eds.), *Raising standards in literacy* (pp. 105–116). London, UK: RoutledgeFalmer.

Rasinski, T.V., Reutzel, C.R., Chard, D., & Linan-Thompson, S. (2011). Reading fluency. In M.L. Kamil, P.D. Pearson, E.B. Moje, & P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 286–319). New York, NY: Routledge.

Rayner, K., Schotter, E.R., Masson, M.E.J., Potter, M.C., & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest*, *17*(1), 4–34. doi:10.1177/1529100615623267

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726–748. doi:10.1037/0033-2909.92.3.726

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:10.18637/jss.v048.i02

Schwanenflugel, P.J., Hamilton, A.M., Kuhn, M.R., Wisenbaker, J., & Stahl, S.A. (2004). Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology*, *96*(1), 119–129. doi:10.1037/0022-0663.96.1.119

Seidenberg, M.S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. doi:10.1037/0033-295X.96.4.523

Shinn, M.R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York, NY: Guilford.

Shinn, M.R. (1998). *Advanced applications of curriculum-based measurement*. New York, NY: Guilford.

Shinn, M.R., Shinn, M.M., Hamilton, C., & Clarke, B. (2002). Using curriculum-based measurement in general education classrooms to promote reading success. In M.R. Shinn, H.M. Walker, & G. Stoner (Eds.), *Interventions for academic and behavior problems*

II: Prevention and remedial approaches (pp. 113–142). Bethesda, MD: National Association of School Psychologists.
Silberglitt, B., & Hintze, J.M. (2007). How much growth can we expect? A conditional analysis of R-CBM growth rates by level of performance. *Exceptional Children*, *74*(1), 71–84. doi:10.1177/001440290707400104
Skaftun, A., Stangeland, E.B., Solheim, O.J., & Mangen, A. (2013). *Den nasjonale prøven i lesing på 5.trinn, 2013* [The national reading test in grade 5, 2013; Technical report]. Stavanger, Norway: The Norwegian Reading Centre, University of Stavanger.
Smolkowski, K., Cummings, K.D., & Strycker, L. (2016). An introduction to the statistical evaluation of fluency measures with signal detection theory. In K.D. Cummings, & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and application* (Vol. 1, pp. 187–221). New York, NY: Springer.
Snowling, M.J., & Hulme, C. (2012). Annual research review: The nature and classification of reading disorders—a commentary on proposals for DSM–5. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *53*(5), 593–607.
Solheim, O.J., Skaftun, A., & Walgermo, B.R. (2012). *Den nasjonale prøven i lesing på 5.trinn, 2012* [The national reading test in grade 5, 2012; Technical report]. Stavanger, Norway: The Norwegian Reading Centre, University of Stavanger.
Speece, D.L., & Ritchey, K.D. (2005). A longitudinal study of the development of oral reading fluency in young children at risk for reading failure. *Journal of Learning Disabilities*, *38*(5), 387–399. doi:10.1177/00222194050380050201
Stage, S., & Jacobsen, M. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review*, *30*(3), 407–419.
Standards & Testing Agency. (2016). *National curriculum assessments: Past papers*. Retrieved from https://www.gov.uk/government/collections/key-stage-2-tests-past-papers#phonics-screening-check
Stanovich, K.E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York, NY: Guilford.
Statens Beredning för Medicinsk Utvärdering (2014). *Dyslexi hos barn och ungdomar—tester och innsatser: En systematisk litteraturöversikt* [Dyslexia in children and adolescence—tests and efforts: A systematic review]. Stockholm, Sweden: Author.
Stecker, P.M., Fuchs, L.S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*(8), 795–819. doi:10.1002/pits.20113
Stoolmiller, M. (1995). Using latent growth curve models to study developmental processes. In J.M. Gottman (Ed.), *The analysis of change* (pp. 103–138). Mahwah, NJ: Erlbaum.
Stoolmiller, M., Biancarosa, G., & Fien, H. (2013). Measurement properties of DIBELS Oral Reading Fluency in grade 2: Implications for equating studies. *Assessment for Effective Intervention*, *38*(2), 76–90. doi:10.1177/1534508412456729
Tindal, G., Nese, J.F.T., Stevens, J.J., & Alonzo, J. (2015). Growth on oral reading fluency measures as a function of special education and measurement sufficiency. *Remedial and Special Education*, *37*(1), 28–40. doi:10.1177/0741932515590234
van IJzendoorn, M.H., & Bus, A.G. (1994). Meta-analytic confirmation of the nonword reading deficit in developmental dyslexia. *Reading Research Quarterly*, *29*(3), 266–275. doi:10.2307/747877
Veenendaal, N.J., Groen, M.A., & Verhoeven, L. (2015). What oral text reading fluency can reveal about reading comprehension. *Journal of Research in Reading*, *38*(3), 213–225. doi:10.1111/1467-9817.12024
Wang, C., Porfeli, E., & Algozzine, B. (2008). Development of oral reading fluency in young children at risk for failure. *Journal of Education for Students Placed at Risk*, *13*(4), 402–425. doi:10.1080/10824660802427702
Wanzek, J., Roberts, G., Linan-Thompson, S., Vaughn, S., Woodruff, A.L., & Murray, C.S. (2010). Differences in the relationship of oral reading fluency and high-stakes measures of reading comprehension. *Assessment for Effective Intervention*, *35*(2), 67–77.
Widaman, K.F., Ferrer, E., & Conger, R.D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. doi:10.1111/j.1750-8606.2009.00110.x
Wise, J.C., Sevcik, R.A., Morris, R.D., Lovett, M.W., Wolf, M., Kuhn, M., … Schwanenflugel, P. (2010). The relationship between different measures of oral reading fluency and reading comprehension in second-grade students who evidence different oral reading fluency difficulties. *Language, Speech, and Hearing Services in Schools*, *41*(3), 340–348. doi:10.1044/0161-1461(2009/08-0093)

Submitted February 11, 2016
Final revision received June 24, 2016
Accepted June 27, 2016

ANNE ARNESEN (corresponding author) is a doctoral student in the Department of Special Needs Education at the University of Oslo, Norway; e-mail anne.arnesen@isp.uio.no. She is interested in aspects of evidence-based assessments for early identification of children at risk for difficulties in reading and social behavior and how Response to Intervention can promote their development and learning.

JOHAN BRAEKEN is an associate professor in the Centre for Educational Measurement at the University of Oslo, Norway; e-mail johan.braeken@cemo.uio.no.

SCOTT BAKER is a research professor at the Center on Research and Evaluation at the Simmons School of Education and Human Development, Southern Methodist University, Dallas, Texas, USA; e-mail skbaker@smu.edu. He is interested in the impact of interventions on child outcomes, mechanisms that underlie effective interventions, and how intervention impact varies by factors intrinsic and extrinsic to the child.

WILHELM MEEK-HANSEN is a special advisor at the Norwegian Center for Child Behavioral Development, Oslo, Norway; e-mail wilhelm.meek-hansen@atferdssenteret.no. He is interested in the relationship between children's emotional and behavioral problems and reading difficulties that impact their social and academic development.

TERJE OGDEN is the research director at the Norwegian Center for Child Behavioral Development, Oslo, Norway; e-mail terje.ogden@atferdssenteret.no.

MONICA MELBY-LERVÅG is a professor in the Department of Special Needs Education at the University of Oslo, Norway; e-mail monica.melby-lervag@isp.uio.no. She is interested in language and reading development in children with dyslexia, specific language difficulties, and minority languages.

APPENDIX A

Textual Properties of the Reading Passages Related to Empirical Performance Deviations From the Intended Parallel Construction

	ΔDIF[a]		LIX readability formula		Flesch readability formula		Summary	Content	
Reading passage	Intercept	Loading	Full	Average read	Reading ease	Grade level	Grade level	Topic	Type
Grade 2									
1	0.00	1.00	11	20	92.9	2.2	2	Dear diary	Narrative
2	3.13***	1.00	12	9	96.2	2.4	2	Ill	Expository
3	0.58	0.90***	13	8	96.6	2.2	2	Pen friend	Narrative
4	0.00	1.04***	12	10	92.3	2.4	2	Little and nice	Expository
5	0.00	1.00	11	9	97.2	1.7	2	In the cabin	Narrative
6	0.00	1.00	15	11	88.8	3.2	3	On way to school	Expository
7	6.24***	1.00	10	7	97.5	2.3	2	Afraid of darkness	Narrative
8	0.00	1.00	12	11	98.6	1.8	2	Angry	Expository
9	0.00	1.00	13	9	93.3	2.3	2	Life along river	Expository
Grade 3									
10	0.00	1.00	14	16	93.9	2.2	3	Soccer tournament	Narrative
11	5.81***	1.00	17	13	86.1	3.8	4	In the library	Expository
12	0.00	1.00	19	20	87.2	3.3	4	School camp	Narrative
13	0.00	1.00	18	14	85.8	3.9	4	Boys don't play	Expository
14	0.00	1.00	18	18	87.0	3.3	4	Grandfather fishing	Narrative
15	0.00	1.00	18	19	85.4	4.0	4	Wild animals	Narrative
16	0.00	1.00	15	12	90.5	3.0	3	Stranger	Narrative
17	−6.94***	1.00	17	11	87.9	3.4	4	An iceland is born	Expository
18	0.00	1.00	18	17	86.6	3.6	4	Moving to town	Narrative
Grade 4									
19	0.00	1.00	20	19	81.6	4.5	5	Author visiting	Narrative
20	12.40***	1.00	20	19	82.4	4.6	4	Moving to London	Narrative
21	0.00	1.00	21	22	83.0	4.6	5	Kayaking	Expository
22	0.00	1.00	21	27	80.8	4.5	5	Bike ride	Narrative
23	6.83***	1.14***	20	18	85.3	4.8	4	Dog in the house	Expository
24	0.00	1.01*	19	18	80.7	4.4	5	Cheeta (cat)	Expository
25	0.00	1.00	23	22	81.1	4.8	5	Water well	Narrative
26	4.97***	1.00	20	24	79.4	5.1	4	Emil and Eilert	Narrative
27	−10.52***	1.00	22	21	83.2	4.7	5	Cairo	Expository

(*continued*)

Textual Properties of the Reading Passages Related to Empirical Performance Deviations From the Intended Parallel Construction (*continued*)

	ΔDIF[a]		LIX readability formula		Flesch readability formula		Summary	Content	
Reading passage	Intercept	Loading	Full	Average read	Reading ease	Grade level	Grade level	Topic	Type
Grade 5									
28	0.00	1.02***	22	23	81.9	3.9	5	Soccer game	Narrative
29	0.00	1.00	24	29	77.0	5.3	6	Vikings	Expository
30	15.36***	1.00	22	27	78.3	5.7	6	Save the children	Expository
31	−10.63***	1.00	26	26	75.2	5.8	5	Way guide	Narrative
32	−17.28***	1.12***	25	22	77.4	5.0	6	Climbing wall	Narrative
33	0.00	1.00	27	23	77.2	5.1	5	Dolphins	Expository
34	4.87***	1.00	19	23	76.9	5.0	5	Family trip	Narrative
35	0.00	1.00	25	25	77.0	5.5	6	Water is source	Expository
36	0.00	1.00	26	30	73.6	5.9	6	The body is not	Expository

[a]Wald tests were run for the ΔDIF parameters.
*$p < .05$. ***$p < .001$.

APPENDIX B

Subtest	Number of items	Timed	Description
National reading assessment subtests for grade 2: Spring 2013			
Recognizing letters	25	1 minute	The students were presented printed capital letters and asked to identify the same printed lowercase letters and vice versa.
Writing words	16	—	To measure skills in spelling, the examiner asked the students to write words when listening to the teacher read words aloud.
Reading words	21	2 minutes	The students were asked to look at a picture and identify which of four words represented the picture.
Splitting compound words	21	1 minute	To measure morphemic awareness and word-decoding skill, the students were asked to divide compound words by putting a line between two meaningful units in words that varied in terms of the number of letters and difficulty.
Reading sentences	18	2 minutes	To measure reading comprehension at the sentence level, the students were asked to read a sentence of increasing length (two to eight words) and identify which of four pictures best represented the meaning of the sentence.
Following written instructions	10	2.5 minutes	The students were asked to read instructions (one or two sentences) and demonstrate their reading comprehension by marking on a picture of elements the one that corresponded to each instruction (e.g., "Please make a circle around the bus station").
Reading text	6	—	To measure reading comprehension, the students read one short text silently and then answered six multiple-choice questions about the text, which was taken from Aesop's fable "The Bear and the Two Friends."

(*continued*)

(*continued*)

Subtest	Number of items	Timed	Description
National reading assessment subtests for grade 3: Spring 2013			
Chains of words	66	5 minutes	To measure decoding and word recognition skills, the students were presented an unbroken chain of four meaningful words (e.g., "onfivebeatcheese") and asked to read the unit as separate words ("*on five beat cheese*").
Reading narrative text	9	15 minutes	To measure reading comprehension, this subtest consisted of a narrative text that students read silently and then answered nine questions about it. The text was from a Norwegian illustrated children's book. Three questions measured literacy comprehension, in that the students could find the information to answer the questions in the text. Five other questions measured inferential comprehension, in that the students had to integrate information from the text with their own background knowledge about the topic or infer meaning in the text from things not stated explicitly in the text to answer the question correctly; one question required students to make a reasonable interpretation of the text based on multiple pieces of information in the text.
Word knowledge	20	—	Multiple-choice items measured vocabulary. Each item consisted of four words. The teacher read the first target word aloud and then each of three option words, one of which was a synonym for the target word. The students marked the word that was the correct synonym.
Reading expository text	7	15 minutes	To measure reading comprehension, the students silently read a text about making pancakes and answered seven multiple-choice questions about the text. Information for five of the questions was directly expressed in the text. For the other two questions, the students had to combine information from different places in the text and rely on their own background knowledge and experience to answer the questions correctly.

APPENDICES

Appendices related to Study 1

Appendix A: The EFPA review model

Appendix E: Survey

Appendix F: Protocol Systematic Literature Search

Appendices related to Study 2

Appendix C: Receipt NSD

Appendix G: The ESBA and the SSRS-T

Appendices related to Study 3

Appendix B: Percentile tables

Appendix D :Receipt NSD

Appendix H : The ORF measure

A

EFPA REVIEW MODEL FOR THE DESCRIPTION AND EVALUATION OF PSYCHOLOGICAL AND EDUCATIONAL TESTS

TEST REVIEW FORM AND NOTES FOR REVIEWERS

VERSION 4.2.6

Version 4.2.6 is a major revision of Version 3-42 (2008) by a task force of the Board of Assessment of EFPA consisting of:

Arne Evers (chair, the Netherlands)
Carmen Hagemeister (Germany)
Andreas Høstmælingen (Norway)
Patricia Lindley (UK)
José Muñiz (Spain)
Anders Sjöberg (Sweden)

Approved by the EFPA General Assembly, 13-07-2013

CONTENTS

1 Introduction 3

PART 1 DESCRIPTION OF THE INSTRUMENT 5

2 General description 6
3 Classification 8
4 Measurement and scoring 14
5 Computer generated reports 16
6 Supply conditions and costs 20

PART 2 EVALUATION OF THE INSTRUMENT 23

7 Quality of the explanation of the rationale, the presentation and the information provided 26
7.1 Quality of the explanation of the rationale 26
7.2 Adequacy of documentation available to the user 26
7.3 Quality of procedural instructions provided for the user 28
8 Quality of the test materials 31
8.1 Quality of the test materials of paper-and-pencil tests 31
8.2 Quality of the test materials of Computer Based Tests (CBT) or Web Based Tests (WBT) 33
9 Norms 33
9.1 Norm-referenced interpretation 38
9.2 Criterion referenced interpretation 43
10 Reliability 53
11 Validity 54
11.1 Construct validity 58
11.2 Criterion validity 61
11.3 Overall validity 62
12 Quality of computer generated reports 66
13 Final evaluation 68

PART 3 BIBLIOGRAPHY 72

APPENDIX An aide memoire of critical points for comment when an instrument has been translated and/or adapted from a non-local context

1 Introduction

The main goal of the EFPA Test Review Model is to provide a description and a detailed and rigorous assessment of the psychological assessment tests, scales and questionnaires used in the fields of Work, Education, Health and other contexts. This information will be made available to test users and professionals in order to improve tests and testing and help them to make the right assessment decisions. The EFPA Test Review Model is part of the information strategy of the EFPA, which aims to provide evaluations of all necessary technical information about tests in order to enhance their use (Evers et al., 2012; Muñiz & Bartram, 2007). Following the *Standards for Educational and Psychological Testing* the label *test* is used for any "... evaluative device or procedure in which a sample of examinee's behaviour in a specified domain is obtained and subsequently evaluated and scored using a standardized process" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 3). Therefore, this review model applies to all instruments that are covered under this definition, whether called a scale, questionnaire, projective technique, or whatever.

The original version of the EFPA test review model was produced from a number of sources, including the BPS Test Review Evaluation Form (developed by Newland Park Associates Limited, NPAL, and later adopted by the BPS Steering Committee on Test Standards); the Spanish Questionnaire for the Evaluation of Psychometric Tests (developed by the Spanish Psychological Association) and the Rating System for Test Quality (developed by the Dutch Committee on Tests and Testing of the Dutch Association of Psychologists). Much of the content was adapted with permission from the review proforma originally developed in 1989 by Newland Park Associates Ltd for a review of tests used by training agents in the UK (see Bartram, Lindley & Foster, 1990). This was subsequently used and further developed for a series of BPS reviews of instruments for use in occupational assessment (e.g., Bartram, Lindley, & Foster, 1992; Lindley et al., 2001). The first version of the EFPA review model was compiled and edited by Dave Bartram (Bartram, 2002a, 2002b) following an initial EFPA workshop in March 2000 and subsequent rounds of consultation. A major update and revision was carried out by Patricia Lindley, Dave Bartram, and Natalie Kennedy for use in the BPS review system (Lindley et al, 2004). This was subsequently adopted by EFPA in 2005 (Lindley et al., 2005) with minor revisions in 2008 (Lindley et al., 2008). The current version of the model has been prepared by a Task Force of the EFPA Board of Assessment, whose members are Arne Evers (Chair, the Netherlands), Carmen Hagemeister (Germany), Andreas Høstmælingen (Norway), Patricia Lindley (UK), José Muñiz (Spain), and Anders Sjöberg (Sweden). In this version the notes and checklist for translated and adapted tests produced by Pat Lindley and the Consultant Editors of the UK test reviews have been integrated (Lindley, 2009). The texts of some major updated passages are based on the revised Dutch rating system for test quality (Evers, Lucassen, Meijer, & Sijtsma, 2010; Evers, Sijtsma, Lucassen, & Meijer, 2010).).

The EFPA test review model is divided into three main parts. In the first part (Description of the instrument) all the features of the test evaluated are described in detail. In the second part (Evaluation of the instrument) the fundamental properties of the test are evaluated: Test materials, norms, reliability, validity, and computer generated reports, including a global final evaluation. In the third part (Bibliography), the references used in the review are included.

As important as the model itself is the proper implementation of the model. The current version of the model is intended for use by two independent reviewers, in a peer review process similar to the usual evaluation of scientific papers and projects. A consulting editor will oversee the reviews and may call in a third reviewer if significant discrepancies between the two reviews are found. Some variations in the procedure are possible, whilst ensuring the competence and independence of the reviewers, as well as the consulting editor. EFPA recommends that the evaluations in these reviews are directed towards qualified

practising test users, though they should also be of interest to academics, test authors and specialists in psychometrics and psychological testing.

Another key issue is the publication of the results of a test's evaluation. The results should be available for all professionals and users (either paid or for free). A good option is that results are available on the website of the National Psychological Association, although they could also be published by third parties or in other media such as journals or books.

The intention of making this model widely available is to encourage the harmonisation of review procedures and criteria across Europe. Although harmonisation is one of the objectives of the model, another objective is to offer a system for test reviews to countries which do not have their own review procedures. It is realized that local issues may necessitate changes in the EFPA Test Review Model or in the review procedures when countries start to use the Model. Therefore, the Model is called a *Model* to stress that local adaptations are possible to guarantee a better fit with local needs.

Comments on the EFPA test review model are welcomed in the hope that the experiences of users will be instrumental in improving and clarifying the processes.

PART 1 DESCRIPTION OF THE INSTRUMENT

2 General description

This section of the form should provide the basic information needed to identify the instrument and where to obtain it. It should give the title of the instrument, the publisher and/or distributor, the author(s), the date of original publication and the date of the version that is being reviewed.

The questions 2.1.1 through 2.7.3 should be straightforward. They are factual information, although some judgment will be needed to complete information regarding content domains.

	Reviewer[1]	
	Date of current review	
	Date of previous review *(if applicable)*[2]	
2.1.1	**Instrument name (local version)**	
2.1.2	**Shortname of the test** *(if applicable)*	
2.2	**Original test name** *(if the local version is an adaptation)*	
2.3	**Authors of the original test**	
2.4	**Authors of the local adaptation**	
2.5	**Local test distributor/publisher**	
2.6	**Publisher of the original version of the test** *(if different to current distributor/publisher)*	
2.7.1	**Date of publication of current revision/edition**	
2.7.2	**Date of publication of adaptation for local use**	
2.7.3	**Date of publication of original test**	

[1] Each country can decide either to publish the reviewers' names when the integrated review is published or to opt for anonymous reviewing.
[2] This information should be filled in by the editor or the administration.

General description of the instrument Short stand-alone non-evaluative description (200-600 words) A concise non-evaluative description of the instrument should be given here. The description should provide the reader with a clear idea of what the instrument claims to be - what it contains, the scales it purports to measure etc. It should be as neutral as possible in tone. It should describe what the instrument is, the scales it measures, its intended use, the availability and type of norm groups, general points of interest or unusual features and any relevant historical background. This description may be quite short (200-300 words). However, for some of the more complex multi-scale instruments, it will need to be longer (300-600 words). It should be written so that it can stand alone as a description of the instrument. As a consequence it may repeat some of the more specific information provided in response to sections 2 – 6. It should outline all versions of the instrument that are available and referred to on subsequent pages. This item should be answered from information provided by the publisher and checked for accuracy by the reviewer.

3 Classification

3.1	**Content domains** *(select all that apply)* You should identify the content domains specified by the publisher. Where these are not clear, this should be indicated and you should judge from the information provided in the manual (standardisation samples, applications, validation etc.) what the most appropriate answers are for 3.1.	□ Ability - General □ Ability - Manual skills/dexterity □ Ability - Mechanical □ Ability Learning/memory □ Ability - Non-verbal/abstract/inductive □ Ability - Numerical □ Ability - Perceptual speed/checking □ Ability - Sensorimotor □ Ability Spatial/visual □ Ability - Verbal □ Attention/concentration □ Beliefs □ Cognitive styles □ Disorder and pathology □ Family function □ Group function □ Interests □ Motivation □ Organisational function, aggregated measures, climate etc □ Personality – Trait □ Personality – Type □ Personality – State □ Quality of life □ Scholastic achievement (educational test) □ School or educational function □ Situational judgment □ Stress/burnout □ Therapy outcome □ Values □ Well-being □ Other (please describe):
3.2	**Intended or main area(s) of use** *(please select those that apply)* You should identify the intended areas of uses specified by the publisher. Where these are not clear, this should be indicated and you should judge from the information provided in the manual (standardisation samples, applications, validation etc) what the most appropriate answers are for 3.2.	□ Clinical □ Advice, guidance and career choice □ Educational □ Forensic □ General health, life and well-being □ Neurological □ Sports and Leisure □ Work and Occupational □ Other (please describe):
3.3	**Description of the populations for which the test is intended** This item should be answered from information provided by the publisher. For some tests this may be very general (e.g. adults), for others it may be more specific (e.g. manual workers, or boys	

	aged 10 to 14). Only the stated populations should be mentioned here. Where these may seem inappropriate, this should be commented on in the Evaluation part of the review.	
3.4	**Number of scales and brief description of the variable(s) measured by the instrument** This item should be answered from information provided by the publisher. Please indicate the number of scales (if more than one) and provide a brief description of each scale if its meaning is not clear from its name. Reviews of the instrument should include discussion of other derived scores where these are commonly used with the instrument and are described in the standard documentation - e.g. primary trait scores as well as Big Five secondary trait scores for a multitrait personality test, or subtest, factor and total scores on an intelligence test.	
3.5	**Response mode** This item should be answered from information provided by the publisher. If any special pieces of equipment (other than those indicated in the list of options, e.g. digital recorder) are required, they should be described here. In addition, any special testing conditions should be described. 'Standard testing conditions' are assumed to be available for proctored/supervised assessment. These would include a quiet, well-lit and well-ventilated room with adequate desk-space and seating for the necessary administrator(s) and candidate(s).	□ Oral interview □ Paper & pencil □ Manual (physical) operations □ Direct observation □ Computerised □ Other (indicate):
3.6	**Demands on the test taker** This item should be answered from information provided by the publisher. Which capabilities and skills are necessary for the test taker to work on the test as intended and to allow for a fair interpretation of the test score? It is usually clear if a total lack of some prerequisite impairs the ability to complete the test (such as being blind and being given a normal paper-and-pencil test) but the requirements listed should be classified as follows: • "Irrelevant / not necessary" means that this skill is not necessary at all – such as manual capabilities to answer oral questions verbally. • "Necessary information given" means that the possible amount of limitation is stated. • "Information missing" means that there might be limitations on test users without the specific capability or skill (known from theory or empirical results) but this is not clear from information provided by the test publisher e.g. if the test uses language that is not the test taker's first language.	Manual capabilities (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Handedness (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Vision (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Hearing (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Command of test language (understanding and speaking) (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Reading (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Writing (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing
3.7	**Items format** *(select one)* This item should be answered from information provided by the publisher. Two types of multiple choice formats are differentiated. The first type concerns tests in which the respondent has to select the right answer from a number of alternatives as in ability testing (e.g., a figural reasoning test). The second type deals with questionnaires in which there is no clear right answer. This format requires test takers to make choices between sets of two or more items drawn from different scales (e.g., scales in a vocational interest inventory or a personality questionnaire). This format is also called 'multidimensional', because the alternatives belong to different scales or dimensions. In this case it is possible that the statements have to be ranked or the most- and least-like-me options be selected. This format may result in ipsative scales (see question 3.8). In Likert scale ratings the test taker also has to choose from a number of alternatives, but the essential difference with the multiple choice format is that the scales used are unidimensional (e.g., ranging from 'never' to 'always' or from 'very unlikely' to 'very likely') and that the test taker does not have to choose between alternatives from different dimensions. A scale should also be marked as a Likert scale when there are only two alternatives on one dimension (e.g., yes/no or always/never).	□ Multiple choice (ability testing, or right/wrong) Number of alternatives: .... □ Multiple choice (mixed scale alternatives) Number of alternatives: .... □ Likert scale ratings Number of alternatives: .... □ Open □ Other (please describe)

3.8	**Ipsativity** As mentioned in 3.7 multiple choice mixed scale alternatives *may* result in ipsative scores. Distinctive for ipsative scores is that the score on each scale or dimension is constrained by the scores on the other scales or dimensions. In fully ipsative instruments the sum of the scale scores is constant for each person. Other scoring procedures can result in ipsativity (e.g. subtraction of each person's overall mean from each of their scale scores)	□ Yes, multiple choice mixed scale alternatives resulting in partially or fully ipsative scores □ Yes, other item formats with scoring procedures resulting in partially or fully ipsative scores □ No, multiple choice mixed scale alternatives NOT resulting in ipsative scores □ Not relevant
3.9	**Total number of test items and number of items per scale or subtest** This item should be answered from information provided by the publisher. If the instrument has several scales or subtests, indicate the total number of items and the number of items for each scale or subtest. Where items load on more than one scale or subtest, this should be documented.	
3.10	**Intended mode of use (conditions under which the instrument was developed and validated)** (*select all that apply*) This item is important as it identifies whether the instrument has been designed with the intention of it being used in unsupervised or uncontrolled administration conditions. Note that usage modes may vary across versions of a tool. This item should be answered from information provided by the publisher and checked for accuracy. Note. The four modes are defined in the *International Guidelines on Computer-Based and Internet Delivered Testing* (International Test Commission, 2005, pp. 5-6).	□ *Open mode*: Where there is no direct human supervision of the assessment session and hence there is no means of authenticating the identity of the test-taker. Internet-based tests without any requirement for registration can be considered an example of this mode of administration. □ *Controlled mode*: No direct human supervision of the assessment session is involved but the test is made available only to known test-takers. Internet tests will require test-takers to obtain a logon username and password. These often are designed to operate on a one-time-only basis. □ *Supervised (proctored) mode:* Where there is a level of direct human supervision over test-taking conditions. In this mode test-taker identity can be authenticated. For Internet testing this would require an administrator to log-in a candidate and confirm that the test had been properly administered and completed. □ *Managed mode:* Where there is a high level of human supervision and control over the test-taking environment. In CBT testing this is normally achieved by the use of dedicated testing centres, where there is a high level of control over access, security, the qualification of test administration staff and the quality and technical specifications of the test equipment.

3.11	**Administration mode(s)** *(select all that apply)* This item should be answered from information provided by the publisher. If any special pieces of equipment (other than those indicated in the list of options, e.g. digital recorder) are required, they should be described here. In addition, any special testing conditions should be described. 'Standard testing conditions' are assumed to be available for proctored/ supervised assessment. These would include a quiet, well-lit and well-ventilated room with adequate desk-space and seating for the necessary administrator(s) and candidate(s).	□ Interactive individual administration □ Supervised group administration □ Computerised locally-installed application – supervised/proctored □ Computerised web-based application – supervised/proctored □ Computerised locally-installed application – unsupervised/self-assessment □ Computerised web-based application – unsupervised/self-assessment □ Other (indicate):
3.12	**Time required for administering the instrument** *(please specify for each administration mode)* This item should be answered from information provided by the publisher. The response to this item can be broken down into a number of components. In most cases, it will only be possible to provide general estimates of these rather than precise figures. The aim is to give the potential user a good idea of the time investment associated with using this instrument. Do NOT include the time needed to become familiar with the instrument itself. Assume the user is experienced and qualified. • Preparation time (the time it takes the administrator to prepare and set out the materials for an assessment session; access and login time for an online administration). • Administration time per session: this includes the time taken to complete all the items and an estimate of the time required to give instructions, work through example items and deal with any debriefing comments at the end of the session. • Scoring: the time taken to obtain the raw-scores. In many cases this may be automated. • Analysis: the time taken to carry out further work on the raw scores to derive other measures and to produce a reasonably comprehensive interpretation (assuming you are familiar with the instrument). Again, this may be automated. • Feedback: the time required to prepare and provide feedback to a test taker and other stakeholders.	Preparation: Administration: Scoring: Analysis: Feedback:

	It is recognised that time for the last two components could vary enormously - depending on the context in which the instrument is being used. However, some indication or comments will be helpful.	
3.13	**Indicate whether different forms of the instrument are available and which form(s) is (are) subject of this review** Report whether or not there are alternative versions (genuine or pseudo-parallel forms, short versions, computerised versions, etc.) of the instrument available and describe the applicability of each form for different groups of people. In some cases, different forms of an instrument are meant to be equivalent to each other - i.e. alternative forms. In other cases, various forms may exist for quite different groups (e.g. a children's form and an adult's form). Where more than one form exists, indicate whether these are equivalent/alternate forms, or whether they are designed to serve different functions - e.g. short and long version; ipsative and normative version. Also describe whether or not parts of the whole test can be used instead of the whole instrument. If computerised versions do exist, describe briefly the software and hardware requirements. Note that standalone computer based tests (CBT) and online packages, if available, should be indicated.	

4 Measurement and scoring

4.1	**Scoring procedure for the test** *(select all that apply)* This item should be completed by reference to the publisher's information and the manuals and documentation. Bureau services are services provided by the supplier - or some agent of the supplier - for scoring and interpretation. In general these are optional services. If scoring and/or interpretation can be carried out ONLY through a bureau service, then this should be stated in the review - and the costs included in the recurrent costs item.	□ Computer scoring with direct entry of responses by test taker □ Computer scoring by Optical Mark Reader entry of responses from the paper response form □ Computer scoring with manual entry of responses from the paper response form □ Simple manual scoring key – clerical skills only required □ Complex manual scoring – requiring training in the scoring of the instrument □ Bureau-service – e.g. scoring by the company selling the instrument □ Other (please describe):
4.2	**Scores** This item should be completed by reference to the publisher's information and the manuals and documentation. Brief description of the scoring system to obtain global and partial scores, correction for guessing, qualitative interpretation aids, etc).	
4.3	**Scales used** (*select all that apply)* This item should be completed by reference to the publisher's information and the manuals and documentation.	*Percentile Based Scores* □ Centiles □ 5-grade classification: 10:20:40:20:10 centile splits □ Deciles □ Other (please describe): *Standard Scores* □ Z-scores □ IQ deviation quotients etc (e.g. mean 100, SD=15 for Wechsler or 16 for Stanford-Binet) □ College Entrance Examination Board (e.g. SAT mean=500, SD=100) □ Stens □ Stanines, C-scores □ T-scores □ Other (please describe): □ Critical scores, expectancy tables or other specific decision oriented indices □ Raw score use only □ Other (please describe):

4.4	**Score transformation for standard scores**	□ Normalised – standard scores obtained by use of normalisation look-up table □ Not-normalised – standard scores obtained by linear transformation □ Not applicable

5 Computer generated reports

Note that this section is purely *descriptive*. Evaluations of the reports should be given in the Evaluation part of the review

For instances where there are multiple generated reports available please complete items 5.2 – 5.13 for each report or substantive report section (copy pages as necessary). This classification system could be used to describe two reports provided by a system, for example, Report 1 may be intended for the test taker or other un-trained users, and Report 2 for a trained user who is competent in the use of the instrument and understands how to interpret it.

5.1	**Are computer generated reports available with the instrument?** If the answer to 5.1 is 'YES' then the following classification should be used to classify the types of reports available. For many instruments, there will be a range of reports available. Please complete a separate form for each report	□ Yes (complete items below) □ No (move to item 6.1)
5.2	**Name or description of report** *(see introduction to this section)*	
5.3	**Media** *(select all that apply)* Reports may consist wholly of text or contain text together with graphical or tabular representations of scores (e.g. sten profiles). Where both text and data are presented, these may simply be presented in parallel or may be linked, so that the relationship between text statements and scores is made explicit.	□ Text only □ Unrelated text and graphics □ Integrated text and graphics □ Graphics only
5.4	**Complexity** *(select one)* Some reports are very simple, for example just substituting a text unit for a sten score in a scale-by-scale description. Others are more complex, involving text units which relate to patterns or configurations of scale scores and which consider scale interaction effects.	□ Simple (For example, a list of paragraphs giving scale descriptions) □ Medium (A mixture of simple descriptions and some configural descriptions) □ Complex (Contains descriptions of patterns and configurations of scale scores, and scale interactions)
5.5	**Report structure** *(select one)* Structure is related to complexity.	□ Scale based – where the report is built around the individual scales. □ Factor based – where the report is constructed around higher order factors - such as the 'Big Five' for personality measures. □ Construct based – where the report is built around one or more sets of constructs (e.g. in a work setting these could be such as team types, leadership styles, or tolerance to stress; in a clinical setting these could be different kinds of psychopathology; etc.) which are

		linked to the original scale scores. □ Criterion based where the reports focuses on links with empirical outcomes (e.g. school performance, therapy outcome, job performance, absenteeism etc). □ Other (please describe):
5.6	**Sensitivity to context** *(select one)* When people write reports they tailor the language, form and content of the report to the person who will be reading it and take account of the purpose of the assessment and context in which it takes place. In a work and organizational context a report produced for selection purposes will be different from one written for guidance or development; a report for a middle-aged manager will differ from that written for a young person starting out on a training scheme and so on. In an educational context a report produced for evaluation of a students' global ability to learn and function in a learning environment will be different from a report produced to assess whether or not a student has a specific learning disorder. A report directed to other professionals suggesting learning goals and interventions will differ from reports directed to parents informing them of their child's strengths and weaknesses. In a clinical context a report produced for diagnostic purposes will be different from a report evaluating a patient's potential for risk-taking behaviour. A report produced with the purpose of providing feedback to patients will be different from a report produced with the purpose of informing authorities whether or not it is safe to release a patient from involuntary treatment.	□ One version for all contexts □ Pre-defined context-related versions; number of contexts: .... □ User definable contexts and editable reports
5.7	**Clinical-actuarial** *(select all that apply)* Most report systems are based on clinical judgment. That is, one or more people who are 'expert-users' of the instrument in question will have written the text units. The reports will, therefore, embody their particular interpretations of the scales. Some systems include actuarial reports where the statements are based on empirical validation studies linking scale scores to, for example, job performance measures, clinical classification, etc.	□ Based on clinical judgment of one expert □ Based on clinical judgment of group of experts □ Based on empirical/actuarial relationships
5.8	**Modifiability** *(select one)*	□ Not modifiable (fixed print-only output) □ Limited modification (limited to certain areas,

	The report output is often fixed. However, some systems will produce output in the form of a file that can be processed by the user. Others may provide online interactive access to both the end user and the test taker.	e.g. biodata fields) □ Unlimited modification (e.g. through access to Word processor document file) □ Interactive report which provides test taker with an opportunity to insert comments or provides ratings of accuracy of content (e.g. through shared online access to an interactive report engine)
5.9	**Degree of finish** *(select one)* Extent to which the system is designed to generate integrated text - in the form of a ready-to-use report - or a set of 'notes', comments, hypotheses etc..	□ Publication quality □ Draft quality
5.10	**Transparency** *(select one)* Systems differ in their openness or transparency to the user. An open system is one where the link between a scale score and the text is clear and unambiguous. Such openness is only possible if both text and scores are presented and the links between them made explicit. Other systems operate as 'black boxes', making it difficult for the user to relate scale scores to text.	□ Clear linkage between constructs, scores and text □ Concealed link between constructs, scores and text □ Mixture of clear/concealed linkage between constructs, scores and text
5.11	**Style and tone** *(select one)* Systems also differ in the extent to which they offer the report reader guidance or direction. In a work and organizational context a statement as "Mr X is very shy and will not make a good salesman..." is stipulative, whereas other statements are designed to suggest hypotheses or raise questions, such as "From his scores on scale Y, Mr X appears to be very shy compared to a reference group of salespersons. If this is the case, he could find it difficult working in a sales environment. This needs to be explored further with him". In an educational context a stipulative statement might be: "The results show that X's mathematical skills are two years below the average of his peers", whereas a statement designed to suggest hypotheses might be: "The results indicate X is easily distracted by external stimuli while performing tasks. Behavioural observations during testing support this. This should be taken under consideration when designing an optimal learning environment for X". In a clinical context a stipulative statement might be: "Test scores indicate the patient has severe visual neglect, and is not able to safely operate a motor vehicle", whereas a state-	□ Directive/stipulative □ Guidance/suggests hypotheses □ Other (please describe):

	ment designed to suggest hypotheses might be: "Mrs X's test scores indicate she may have problems establishing stable emotional relationships. This should be explored further before a conclusion regarding diagnosis is drawn".	
5.12	**Intended recipients** *(select all that apply)* Reports are generally designed to address the needs of one or more categories of users. Users can be divided into four main groups: *a) Qualified test users.* These are people who are sufficiently knowledgeable and skilled to be able to produce their own reports based on scale scores. They should be able to make use of reports that use technical psychometric terminology and make explicit linkages between scales and descriptions. They should also be able to customize and modify reports. *b) Qualified system users.* While not competent to generate their own reports from a set of scale scores, people in this group are competent to use the outputs generated by the system. The level of training required to attain this competence will vary considerably, depending on the nature of the computer reports (e.g. trait-based versus competency-based, simple or complex) and the uses to which its reports are to be put (low stakes or high stakes). *c) Test Takers.* The person who takes the instrument will generally have no prior knowledge of either the instrument or the type of report produced by the system. Reports for them will need to be in language that makes no assumptions about psychometric or instrument knowledge. *d) Third parties.* These include people - other than the candidate - who will be privy to the information presented in the report or who may receive a copy of the report. They may include potential employers, a person's manager or supervisor or the parent of a young person receiving careers advice. The level of language required for people in this category would be similar to that required for reports intended for Test Takers.	□ Qualified test users □ Qualified system users □ Test takers □ Third Parties
5.13	**Do distributors offer a service to modify and/or develop customised computerised reports?** *(select one)*	□ Yes □ No

6 Supply conditions and costs

This defines what the publisher will provide, to whom, under what conditions and at what costs. It defines the conditions imposed by the supplier on who may or may not obtain the instrument materials. If one of the options does not fit the supply conditions, provide a description of the relevant conditions

6.1	**Documentation provided by the distributor as part of the test package** *(select all that apply)*	□ User Manual □ Technical (psychometric) manual □ Supplementary technical information and updates (e.g. local norms, local validation studies etc.) □ Books and articles of related interest
6.2	**Methods of publication** *(select all that apply)* For example, technical manuals may be kept up-to-date and available for downloading from the Internet, while user manuals are provided in paper form or on a CD/DVD.	□ Paper □ CD or DVD □ Internet download □ Other (specify):
Items 6.3 - 6.5 cover costs. This information is likely to be the most quickly out of date. It is recommended that the supplier or publisher is contacted as near the time of publication of the review as possible, to provide current information for these items.		
6.3.1	**Start-up costs** Price of a complete set of materials (all manuals and other material sufficient for at least one sample administration). Specify how many test takers could be assessed with the materials obtained for start-up costs, and whether these costs include materials for recurrent assessment. This item should try to identify the 'set-up' cost. That is the costs involved in obtaining a full reference set of materials, scoring keys and so on. It only includes training costs if the instrument is a 'closed' one - where there will be an unavoidable specific training cost, regardless of the prior qualification level of the user. In such cases, the training element in the cost should be made explicit. The initial costs do NOT include costs of general-purpose equipment (such as computers, DVD players and so on). However, the need for these should be mentioned. In general, define: any special training costs; costs of administrator's manual; technical manual(s); specimen or reference set of materials; initial software costs, etc.	

6.3.2	**Recurrent costs** Specify, where appropriate, recurrent costs of administration and scoring separately from costs of interpretation (see 6.4.1 – 6.5). This item is concerned with the on-going cost of using the instrument. It should give the cost of the instrument materials (answer sheets, non-reusable or reusable question booklets, profile sheets, computer usage release codes or 'dongle' units, etc.) per person per administration. Note that in most cases, for paper-based administration such materials are not available singly but tend to be supplied in packs of 10, 25 or 50. Itemise any annual or per capita licence fees (including software release codes where relevant), costs of purchases or leasing re-usable materials, and per candidate costs of non-reusable materials.	
6.4.1	**Prices for reports generated by user installed software**	
6.4.2	**Prices for reports generated by postal/fax bureau service**	
6.4.3	**Prices for reports by Internet service**	
6.5	**Prices for other bureau services: correcting or developing automatic reports**	
6.6	**Test-related qualifications required by the supplier of the test** *(select all that apply)* This item concerns the user qualifications required by the supplier. For this item, where the publisher has provided user qualification information, this should be noted against the categories given. Where the qualification requirements are not clear this should be stated under 'Other' *not* under 'None'. 'None' means that there is an explicit statement regarding the lack of need for qualification. For details of the EFPA Level 2 standard, consult the latest version of these on the EFPA website.	□ None □ Test specific accreditation □ Accreditation in general achievement testing: measures of maximum performance in attainment (equivalent to EFPA Level 2) □ Accreditation in general ability and aptitude testing: measures of maximum performance in relation to potential for attainment (equivalent to EFPA Level 2) □ Accreditation in general personality and assessment: measures of typical behaviour, attitudes and preferences (equivalent to EFPA Level 2) □ Other (specify):

6.7	**Professional qualifications required for use of the instrument** *(select all that apply)* This item concerns the user qualifications required by the supplier. For this section, where the publisher has provided user qualification information, this should be noted against the categories given. Where the qualification requirements are not clear this should be stated under 'Other' *not* under 'None'. 'None' means that there is an explicit statement regarding the lack of need for qualification. For details of the EFPA user standards, consult the latest version of these on the EFPA website.	□ None □ Practitioner psychologist with qualification in the relevant area of application □ Practitioner psychologist □ Research psychologist □ Non-psychologist academic researcher □ Practitioner in relevant related professions (therapy, medicine, counselling, education, human resources etc.). Specify: ..... □ EFPA Test User Qualification Level 1 or national equivalent □ EFPA Test User Qualification Level 2 or national equivalent □ Specialist qualification equivalent to EFPA Test User Standard Level 3 □ Other (indicate):

PART 2 EVALUATION OF THE INSTRUMENT

Sources of information

Potentially there are four sources of information that might be consulted in carrying out this evaluation:

1. The manual and /or reports that are supplied by the publisher for the user:
   These are always supplied by the publisher /distributor before the instrument is accepted by the reviewing organisation and form the core materials for the review.

2. Open information that is available in the academic or other literature:
   This is generally sourced by the reviewer and the reviewer may make use of this information in the review and the instrument may be evaluated as having (or having not) made reference to the information in its manual.

3. Information held by the publisher that is not formally published or distributed:
   The distributor/publisher may make this available at the outset or may send it when the review is sent back to the publisher to check for factual accuracy. The reviewer should make use of this information but note very clearly at the beginning of the comments on the technical information that "the starred rating in this review refers to materials held by the publisher/distributor that is not [normally] supplied to test users". If these contain valuable information, the overall evaluation should recommend that the publisher publishes these reports and/or make them available to test purchasers.

4. Information that is commercial in confidence:
   In some instances, publishers may have technically important material that they are unwilling to make public for commercial reasons. In practice there is very little protection available for intellectual property to test developers (copyright law being about the only recourse). Such information could include reports that cover the development of particular scoring algorithms, test or item generation procedures and report generation technology. Where the content of such reports might be important in making a judgment in a review, the association or organization responsible for the review could offer to undertake to enter into a non-disclosure agreement with the publisher. This agreement would be binding on the reviewers and editor. The reviewer could then evaluate the information and comment on the technical aspects and the overall evaluation to the effect that "the starred rating in this review refers to materials held by the publisher/ distributor that have been examined by the reviewers on a commercial in confidence basis. These are not supplied to end users."

Explanation of ratings

All sections are scored using the following rating system (see table on next page). Detailed descriptions giving anchor-points for each rating are provided.
Where a [0] or [1] rating is provided on an attribute that is regarded as *critical* to the safe use of an instrument, the review will recommend that the instrument should only be used in exceptional circumstances by highly skilled experts or in research.
The instrument review needs to indicate which, given the nature of the instrument and its intended use, are the critical technical qualities. It is suggested that the convention to adopt is that ratings of these critical qualities are then shown in bold print.
In the following sections, overall ratings of the adequacy of information relating to validity, reliability and norms are shown, by default, in bold.

Any instrument with one or more [0] or [1] ratings regarding attributes that are regarded as critical to the safe use of that instrument, shall not be deemed to have met the minimum standard.

Rating	Explanation*
[n/a]	This attribute is not applicable to this instrument
0	Not possible to rate as no, or insufficient information is provided
1	Inadequate
2	Adequate
3	Good
4	Excellent

** A five point scale is defined by EFPA but each user can concatenate the points on the scale (for example combining points 3 and 4 into a single point). The only constraint is that there must be a distinction made between inadequate (or worse) on the one hand and adequate (or better) on the other. Descriptive terms or symbols such as stars or smiley faces may be used in place of numbers. Where the five point scale is replaced or customized, the user should provide a key that links the points and the nomenclature to the five point scale of EFPA.*

7 Quality of the explanation of the rationale, the presentation and the information provided

In this section a number of ratings need to be given to various aspects or attributes of the documentation supplied with the instrument (or package). The term 'documentation' is taken to cover all those materials supplied or readily available to the qualified user: e.g. the administrator's manual; technical handbooks; booklets of norms; manual supplements; updates from publishers/suppliers and so on.

Suppliers are asked to provide a complete set of such materials for each Reviewer. If you think there is something which users are supplied with which is not contained in the information sent to you for review, please contact your review editor.

7.1 Quality of the explanation of the rationale

If the instrument is a computer-adaptive test particular attention should be paid to the items 7.1.1 to 7.1. 6.

Items to be rated n/a or 0 to 4		Rating					
7.1.1	Theoretical foundations of the constructs	n/a	0	1	2	3	4
7.1.2	Test development (and/or translation or adaptation) procedure	n/a	0	1	2	3	4
7.1.3	Thoroughness of the item analyses and item analysis model	n/a	0	1	2	3	4
7.1.4	Presentation of content validity	n/a	0	1	2	3	4
7.1.5	Summary of relevant research	n/a	0	1	2	3	4
7.1.6	**Overall rating of the quality of the explanation of the rationale** This overall rating is obtained by using judgment based on the ratings given for items 7.1.1 – 7.1.5	n/a	0	1	2	3	4

7.2 Adequacy of documentation available to the user (user and technical manuals, norm supplements, etc.)

The focus here is on the quality of coverage provided in the documentation accessible to qualified users. Note that sub-section 7.2 is about the comprehensiveness and clarity of the documentation available to the user (user and technical manuals, norm supplements etc.) in terms of its coverage and explanation. In terms of the quality of the instrument as evidenced by the documentation, areas in this part are elaborated on under: 7.1, 7.3, 9, 10 and 11.

Items to be rated n/a or 0 to 4, 'benchmarks' are provided for an 'excellent' (4) rating.		Rating					
7.2.1	Rationale (see rating 7.1.6) Excellent: Logical and clearly presented description of what it is designed to measure and why it was constructed as it was.	n/a	0	1	2	3	4
7.2.2.1	Development Excellent: Full details of item sources, development of stimulus material according to accepted guidelines (e.g. Haladyna, Downing, & Rodriguez, 2002; Moreno, Martinez, & Muñiz, 2006), piloting, item analyses, comparison studies and changes made during development trials.	n/a	0	1	2	3	4
7.2.2.2	Development of the test through translation/adaptation Excellent: Information in the manual showing that the translation/adaptation process was done according to international guidelines (ITC, 2000) and included: • Input from native speakers of new language • Multiple review by both language and content (of test) experts • Back translation from new language into original language • Consideration of cultural and linguistic differences.	n/a	0	1	2	3	4
7.2.3	Standardisation Excellent: Clear and detailed information provided about sizes and sources of standardisation sample and standardisation procedure.	n/a	0	1	2	3	4
7.2.4	Norms Excellent: Clear and detailed information provided about sizes and sources of norms groups, representativeness, conditions of assessment etc.	n/a	0	1	2	3	4
7.2.5	Reliability Excellent: Excellent explanation of reliability and standard error of measurement (SEM), and a comprehensive range of internal consistency, temporal stability and/or inter-scorer and inter-judge reliability measures and the resulting SEM's provided with explanations of their relevance, and the generalisability of the assessment instrument.	n/a	0	1	2	3	4
7.2.6	Construct validity Excellent: Excellent explanation of construct validity with a wide range of studies clearly and fairly described.	n/a	0	1	2	3	4
7.2.7	Criterion validity Excellent: Excellent explanation of criterion validity with a wide range of studies clearly and fairly described.	n/a	0	1	2	3	4

7.2.8	Computer generated reports Excellent: Clear and detailed information provided about the format, scope, reliability and validity of computer generated reports.						
7.2.9	**Adequacy of documentation available to the user (user and technical manuals, norm supplements, etc.)** This rating is obtained by using judgment based on the ratings given for items 7.2.1 – 7.2.8	n/a	0	1	2	3	4

7.3 Quality of the procedural instructions provided for the user

Items to be rated n/a or 0 to 4, 'benchmarks' are provided for an 'excellent' (4) rating		Rating					
7.3.1	For test administration Excellent: Clear and detailed explanations and step-by-step procedural guides provided, with good detailed advice on dealing with candidates' questions and problem situations.	n/a	0	1	2	3	4
7.3.2	For test scoring Excellent: Clear and detailed information provided, with checks described to deal with possible errors in scoring. If scoring is done by the computer, is there evidence that the scoring is done correctly?	n/a	0	1	2	3	4
7.3.3	For norming Excellent: Clear and detailed information provided, with checks described to deal with possible errors in norming. If norming is done by the computer, is there evidence that score transformation is correct and the right norm group is applied?	n/a	0	1	2	3	4
7.3.4	For interpretation and reporting Excellent: Detailed advice on interpreting different scores, understanding normative measures and dealing with relationships between different scales, with illustrative examples and case studies; also advice on how to deal with the possible influence of inconsistency in answering, response styles, faking, etc.	n/a	0	1	2	3	4

7.3.5	For providing feedback and debriefing test takers and others Excellent: Detailed advice on how to present feedback to candidates including the use of computer generated reports (if available)	n/a	0	1	2	3	4
7.3.6	For providing good practice issues on fairness and bias Excellent: Detailed information reported about gender and ethnic bias studies, with relevant warnings about use and generalisation of validities	n/a	0	1	2	3	4
7.3.7	Restrictions on use Excellent: Clear descriptions of who should and who should not be assessed, with well-explained justifications for restrictions (e.g. types of disability, literacy levels required etc.)	n/a	0	1	2	3	4
7.3.8	Software and technical support Excellent: In the case of Computer Based Testing (CBT) or Web Based Testing (WBT): the information with respect to browser requirements, the installation of any required computer software and the operation of the software is complete (also covering possible errors and different systems), and availability of technical support is clearly described.	n/a	0	1	2	3	4
7.3.9	References and supporting materials Excellent: Detailed references to the relevant supporting academic literature and cross-references to other related assessment instrument materials.	n/a	0	1	2	3	4
7.3.10	**Quality of the procedural instructions provided for the user** This overall rating is obtained by using judgment based on the ratings given for items 7.3.1 – 7.3.9	n/a	0	1	2	3	4

7.4	**Overall adequacy** This overall rating for section 7 is obtained by using judgment based on the overall ratings given for the subsections 7.1, 7.2, and 7.3.	n/a	0	1	2	3	4

Reviewers' comments on the documentation: (comment on rationale, presentation and information provided)

8 Quality of the test materials

8.1 Quality of the test materials of paper-and-pencil tests

(this sub-section can be skipped if not applicable)

Items to be rated n/a or 0 to 4		Rating					
8.1.1	General quality of test materials (test booklets, answer sheets, test objects, etc.)	n/a	0	1	2	3	4
8.1.2	Ease with which the test taker can understand the task	n/a	0	1	2	3	4
8.1.3	Clarity and comprehensiveness of the instruction (including sample items and practice trials) for the test taker	n/a	0	1	2	3	4
8.1.4	Ease with which responses or answers can be made by the test taker	n/a	0	1	2	3	4
8.1.5	Quality of the formulation of the items and clarity of graphical content in the case of non-verbal items.	n/a	0	1	2	3	4
8.1.6	**Quality of the materials of paper-and-pencil tests** This overall rating is obtained by using judgment based on the ratings given for items 8.1.1 – 8.1.5	n/a	0	1	2	3	4

8.2 Quality of the test materials of CBT and WBT

(this sub-section can be skipped if not applicable)

Items to be rated n/a or 0 to 4		Rating					
8.2.1	Quality of the design of the software (e.g. robustness in relation to operation when incorrect keys are pressed, internet connections fail etc.)	n/a	0	1	2	3	4
8.2.2	Ease with which the test taker can understand the task	n/a	0	1	2	3	4
8.2.3	Clarity and comprehensiveness of the instructions (including sample items and practice trials) for the test taker, the operation of the software and how to respond if the test is administered by computer	n/a	0	1	2	3	4
8.2.4	Ease with which responses or answers can be made by the test taker	n/a	0	1	2	3	4
8.2.5	Quality of the design of the user interface	n/a	0	1	2	3	4
8.2.6	Security of the test against unauthorized access to items or to answers	n/a	0	1	2	3	4
8.2.7	Quality of the formulation of the items and clarity of graphical content in the case of non-verbal items.	n/a	0	1	2	3	4

8.2.8	**Quality of the materials of CBT and WBT** This overall rating is obtained by using judgment based on the ratings given for items 8.2.1 – 8.2.7	n/a	0	1	2	3	4

Reviewers' comments on quality of the materials

9 Norms

General guidance on assigning ratings for this section

It is difficult to set clear criteria for rating the technical qualities of an instrument. These notes provide some guidance on the sorts of values to associate with inadequate, adequate, good and excellent ratings. However these are intended to act as guides only. The nature of the instrument, its area of application, the quality of the data on which norms are based, and the types of decisions that it will be used for should all affect the way in which ratings are awarded.

To give meaning to a raw test score two ways of scaling or categorizing raw scores can be distinguished (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). First, a set of scaled scores or norms may be derived from the distribution of raw scores of a reference group. This is called norm-referenced interpretation (see sub-section 9.1). Second, standards may be derived from a domain of skills or subject matter to be mastered (domain-referenced interpretation) or cut scores may be derived from the results of empirical validity research (criterion-referenced interpretation)(see sub-section 9.2). With the latter two possibilities raw scores will be categorized in two (for example 'pass' of 'fail') or more different score ranges, e.g. to assign patients in different score ranges to different treatment programs, to assign pupils scoring below a critical score to remedial teaching, or to accept or reject applicants in personnel selection.

9.1 Norm-referenced interpretation

(This sub-section can be skipped if not applicable)

Notes on international norms

Careful consideration needs to be given to the suitability of international (same language) norms. Where these have been carefully established from samples drawn from a group of countries, they should be rated on the same basis as nationally based (single language) norm groups. Where a non-local norm is provided strong evidence of equivalence for both test versions and samples to justify its use should be supplied. Generally such evidence would require studies demonstrating scalar equivalence between the source and target language versions. *Where this has not been reported then it should be commented upon in the Reviewers' comments at the end of section 9.*

An international norm may be the most appropriate for international usage (i.e. comparing people who have taken the test in different languages) but the issues listed below should be considered in determining its appropriateness. In general, use of an international norm requires the demonstration of at least measurement equivalence between the source and target language versions of the test.

The nature of the sample

- The balance of sources of the sample (e.g. a sample that is 95% German with a 2% Italian and 3% British is not a real international sample). A sample could be weighted to better reflect its different constituents.
- The equivalence of the background (employment, education, circumstances of testing etc.) of the different parts of the sample. Norm samples which do not allow this to be evaluated are insufficient.

The type of measure:

- Where there are measures which have little or no verbal content then there will be less impact on translation. This will apply to performance tests and to some extent to abstract and diagrammatic reasoning tests where should be less impact on the scores.

The equivalence of the test version used with the different language samples.

- There should be evidence that all the language versions are well translated/adapted
- Is there any evidence that any of the groups have completed the test in a non-primary language?

Similarities of scores in different samples:

- Evidence should be provided about the relative score patterns of the sample sections from different countries. Where there are large differences these should be accounted for and the implications in use discussed. E.g. if a Spanish sample scores higher on a scale than a Dutch sample is there an explanation of what it means to compare members of either group, or a third group against the average? Is there an interpretation of the difference?

Absence of these sources of evidence need to be commented upon in the Reviewers Comments at the end of the section

Guidance given about generalising the norms beyond those groups included in the international norms should be included in the manual for the instrument

- e.g. if a norm is made up of 20% German, 20% French,20% Italian, 20% British and 20% Dutch, it might be appropriate to use it as a comparison group for Swiss or Belgian candidates but it may not be appropriate to use it as a comparison for a group of Chinese applicants.

9.1	**Norm-referenced interpretation** Where an instrument is designed for use without recourse to norms or reference groups (e.g., ipsative tests designed for intra-individual comparisons only), the 'not applicable' category should be used rather than 'no information given'. However, the reviewer should evaluate whether the reasoning to provide no norms is justified, otherwise the category 'no information given' must be used.	
9.1.1	Appropriateness for local use, whether local or international norms Note that for adapted tests only local (nationally based) or really international norms are eligible for the ratings 2, 3 or 4 even if construct equivalence across cultures is found. Where measurement invariance issues arise separate norms should be provided for (sub)groups and any issues encountered should be explained.	
	Not applicable	n/a
	No information given	0
	Not locally relevant (e.g. inappropriate foreign samples)	1
	Local sample(s) that do(es) not fit well with the relevant application domain but could be used with caution	2
	Local country samples or relevant international samples with good relevance for intended application	3
	Local country samples or relevant international samples drawn from well-defined populations from the relevant application domain	4
9.1.2	Appropriateness for intended applications	
	Not applicable	n/a
	No information given	0

	Norm or norms not adequate for intended applications	1
	Adequate general population norms and/or range of norm tables, or adequate norms for some but not all intended applications	2
	Good range of norm tables	3
	Excellent range of sample relevant, age-related and sex-related norms with information about other differences within groups (e.g. ethnic group mix)	4

9.1.3 Sample sizes (classical norming)

For most purposes, samples of less than 200 test takers will be too small, as the resolution provided in the tails of the distribution will be very small. The SE_{mean} for a z-score with N = 200 is 0.071 of the SD - or just better than one *T*-score point. Although this degree of inaccuracy may have only minor consequences in the centre of the distribution the impact at the tails of the distribution can be quite big (and this may be the score ranges that are most relevant for decisions to be taken on basis of the test scores). If there are international norms then in general, because of their heterogeneity, these need to be larger than the typical requirements of local samples.

Different guideline figures are given for low and high stakes use. Generally high-stakes use is where a non-trivial decision is based at least in part on the test score(s).

	Low-stakes use	High-stakes decisions	
Not applicable			n/a
No information given			0
Inadequate sample size	e.g. < 200	e.g. 200-299	1
Adequate sample size	e.g. 200-299	e.g. 300-399	2
Good sample size	e.g. 300-999	e.g. 400-999	3
Excellent sample size	e.g. ≥ 1000	e.g. ≥ 1000	4

9.1.4 Sample sizes continuous norming

Continuous norming procedures have become more and more popular. They are used particularly for tests that are intended for use in schools (e.g. group 1 to 8 in primary education) or for a specific age range (e.g. an intelligence test for 6-16 year olds). Continuous norming is more efficient as fewer respondents are required to get the same amount of accuracy of the norms. Bechger, Hemker, and Maris (2009) have computed some values for the sizes of continuous norm groups that would give equal accuracy compared to classical norming. When eight subgroups are used N = 70 (8x70) gives equal accuracy compared to *N*s of 200 (8x200) with the classical approach; N = 100 (x8) compares to 300 (x8) and N = 150 (x8) to 400 (x8). In these cases the accuracy on the basis of the continuous norming approach is even better in the middle groups, but somewhat worse in the outer groups. Apart from the greater efficiency, another advantage is that, based on the regression line, values for intermediate norm groups can be computed. However, the approach is based on rather strict statistical assumptions. The test author has to show that these assumptions have been met, or that deviations from these assumptions do not have serious consequences for the accuracy of the norms.
Note that when the number of groups is higher, the number of respondents in each group may be lower and vice versa. For high-stakes decisions, such as school admission, the required number shifts by one step upwards.

	Not applicable	n/a
	No information given	0
	Inadequate sample size (e.g. fewer than 8 subgroups with a maximum of 69 respondents each)	1
	Adequate sample size (e.g. 8 subgroups with 70 - 99 respondents each)	2
	Good sample size (e.g. 8 subgroups with 100 - 149 respondents each)	3
	Excellent sample size (e.g. 8 subgroups with at least 150 respondents each)	4

9.1.5 Procedures used in sample selection (*select one*)

A norm group must be representative of the referred group. A sample can be considered representative of the intended population if the composition of the sample with respect to a number of variables (e.g., age, gender, education) is similar to that of the population, *and* when the sample is gathered with a probability sampling model. In such a model the chance of being included in the sample is equal for each element in the population. In both probability and non-probability sampling different methods can be used.

In probability sampling, when an individual person is the unit of selection, three methods can be differentiated: purely random, systematic (e.g. each tenth member of the population) and stratified (for some important variables, e.g. gender, numbers to be selected are fixed to guarantee representativeness on these variables). However (e.g. for the sake of efficiency), groups of persons can also be sampled (e.g. school classes), or a combination of group and individual sampling can be used. In non-probability sampling four methods are differentiated: pure convenience sampling (simply add every tested person to the norm group, as is done in most samples for personnel selection; post-hoc data may be classified into meaningful sub-groups based on biographical and situational information), quota sampling (as in convenience sampling, but it is specified before how many respondents in each subgroup are needed, as is done in survey research), snow ball sampling (ask you friends to participate, and ask them to ask their friends, etc.) and purposive sampling (e.g., select extreme groups to participate).

No information is supplied	[ ]
Probability sample – random	[ ]
Probability sample – systematic	[ ]
Probability sample – stratified	[ ]
Probability sample – cluster	[ ]
Probability sample – multiphases (e.g. first cluster then random within clusters)	[ ]
Non-probability sample – convenience	[ ]
Non-probability sample – quota	[ ]
Non-probability sample – 'snow ball'	[ ]
Non-probability sample – purposive	[ ]
Other, describe:	[ ]

9.1.6	Representativeness of the norm sample(s)	
	Not applicable	n/a
	No information given	0
	Inadequate representativeness for the intended application domain or the representativeness cannot be adequately established with the information provided	1
	Adequate	2
	Good	3
	Excellent: Data are gathered by means of a random sampling model; a thorough description of the composition of the sample(s) and the population(s) with respect to relevant background variables (such as gender, age, education, cultural background, occupation) is provided; good representativeness with regard to these variables is established	4
9.1.7	Quality of information provided about minority/protected group differences, effects of age, gender etc.	
	Not applicable	n/a
	No information given	0
	Inadequate information	1
	Adequate general information, with minimal analysis	2
	Good descriptions and analyses of groups and differences	3
	Excellent range of analyses and discussion of relevant issues relating to use and interpretation	4
9.1.8	How old are the normative studies?	
	Not applicable	n/a
	No information given	0
	Inadequate, 20 years or older	1
	Adequate, norms between 15 and 19 years old	2
	Good, norms between 10 and 14 years old	3
	Excellent, norms less than 10 years old	4
9.1.9	Practice effects (only relevant for performance tests)	
	Not applicable	n/a
	No information given though practice effects can be expected	[ ]

	General information given	[ ]
	Norms for second test application after typical test-retest-interval	[ ]

9.2 Criterion-referenced interpretation

(This sub-section can be skipped if not applicable)

To determine the critical score(s) one can differentiate between procedures that make use of the judgment of experts (these methods are also referred to as domain-referenced norming, see sub-category 9.2.1) and procedures that make use of actual data with respect to the relation between the test score and an external criterion (referred to as criterion-referenced in the restricted sense, see sub-category 9.2.2).

9.2.1	**Domain-referenced norming**	
9.2.1.1	If the judgment of experts is used to determine the critical score, are the judges appropriately selected and trained? Judges should have knowledge of the content domain of the test and they should be appropriately trained in judging (the work of) test takers and in the use of the standard setting procedure applied. The procedure of the selection of judges and the training offered must be described.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
9.2.1.2	If the judgment of experts is used to determine the critical score, is the number of judges used adequate? The required number of judges depends on the tasks and the contexts. The numbers suggested should be considered as an absolute minimum.	
	Not applicable	n/a
	No information given	0
	Inadequate (less than two judges)	1
	Adequate (two judges)	2
	Good (three judges)	3
	Excellent (four judges or more)	4

9.2.1.3	If the judgment of experts is used to determine the critical score, which standard setting procedure is reported? (*select one*)	
	Nedelsky	[ ]
	Angoff	[ ]
	Ebel	[ ]
	Zieky and Livingston (limit group)	[ ]
	Berk (contrast groups)	[ ]
	Beuk	[ ]
	Hofstee	[ ]
	Other, describe:	[ ]
9.2.1.4	If the judgment of experts is used to determine the critical score, which method to compute inter-rater agreement is reported? (*select one*)	
	Coefficient p_0	[ ]
	Coefficient Kappa	[ ]
	Coefficient Livingston	[ ]
	Coefficient Brennan and Kane	[ ]
	Intra Class Coefficient	[ ]
	Other, describe:	[ ]
9.2.1.5	If the judgment of experts is used to determine the critical score, what is the size of the inter-rater agreement coefficients (e.g. Kappa or ICC)? In the scientific literature there are no unequivocal standards for the interpretation of these kinds of coefficients, although generally values below .60 are considered insufficient. Below the classification of Shrout (1998) is followed. Using the classification needs some caution, because the prevalence or base rate may affect the value of Kappa.	
	Not applicable	n/a
	No information given	0
	Inadequate (e.g. $r < 0.60$)	1
	Adequate (e.g. $0.60 \leq r < 0.70$)	2
	Good (e.g. $0.70 \leq r < 0.80$)	3
	Excellent (e.g. $r \geq 0.80$)	4

9.2.1.6	How old are the normative studies?	
	Not applicable	n/a
	No information given	0
	Inadequate, 20 years or older	1
	Adequate, norms between 15 and 19 years old	2
	Good, norms between 10 and 14 years old	3
	Excellent, norms less than 10 years old	4
9.2.1.7	Practice effects (only relevant for performance tests)	
	No information given though practice effects can be expected	[ ]
	General information given	[ ]
	Norms for second test application after typical test-retest-interval	[ ]
9.2.2	**Criterion-referenced norming**	
9.2.2.1	If the critical score is based on empirical research, what are the results and the quality of this research? To answer this question no explicit guidelines can be given as to which level of relationship is acceptable, not only because what is considered 'high' or 'low' may differ for each criterion to be predicted, but also because prediction results will be influenced by other variables such as base rate or prevalence. Therefore, the reviewer has to rely on his/her expertise for his/her judgment. Also the composition of the sample used for this research (is it similar to the group for which the test is intended, more heterogeneous, or more homogeneous?) and the size of this group must be taken into account.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
9.2.2.2	How old are the normative studies?	
	Not applicable	n/a
	No information given	0
	Inadequate, 20 years or older	1

	Adequate, norms between 15 and 19 years old	2
	Good, norms between 10 and 14 years old	3
	Excellent, norms less than 10 years old	4
9.2.2.3	Practice effects (only relevant for performance tests)	
	No information given though practice effects can be expected	[ ]
	General information given	[ ]
	Norms for second test application after typical test-retest-interval	[ ]
9.3	**Overall adequacy** This overall rating is obtained by using judgment based on the ratings given for items 9.1 – 9.2.2.3. The overall rating for *norm-referenced interpretation* can never be higher than the rating for the sample-size-item, but it can be lower dependent on the other information provided. From this other information especially information about the representativeness and the ageing of norms is relevant. If non-probability norm groups are used the quality of the norms can at most be qualified as 'adequate', but only when the description of the norm group shows that the distribution on relevant variables is similar to the target or referred group. The overall rating should reflect the characteristics of the largest and most meaningful norms rather than 'average' across all published norms. The overall rating for *criterion-referenced interpretation* in case judges are used to determine the critical score can never be higher than the rating for the size of the inter-rater agreement, but it can be lower dependent on the other information provided. From this other information especially the correct application of the method concerned and the quality, the training and the number of judges are important. If the critical score is based on empirical research, the rating can never be higher than the rating for item 9.2.2.1, but it can be lower when the studies are too old.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

Reviewers' comments on the norms: Brief report about the norms and their history, including information on provisions made by the publisher/author for updating norms on a regular basis. Comments pertaining to non-local norms should be made here.

10 Reliability

General guidance on assigning ratings for this section

Reliability refers to the degree to which scores are free from measurement error variance (i.e. a range of expected measurement error). For reliability, the guidelines are based on the need to have a small Standard Error for estimates of reliability. Guideline criteria for reliability are given in relation to two distinct contexts: the use of instruments to make decisions about groups of people (e.g. organizational diagnosis) and their use for making individual assessments. Reliability requirements are higher for the latter than the former. Other factors can also affect reliability requirements, such as the kind of decisions made and whether scales are interpreted on their own, or aggregated with other scales into a composite scale. In the latter case the reliability of the composite should be the focus for rating not the reliabilities of the components.

When an instrument has been translated and/or adapted from a non-local context, one could apply reliability evidence of the original version to support the quality of the translated/adapted version. In this case evidence of equivalence of the measure in a new language to the original should be proposed. Without this it is not possible to generalise findings in one country/language version to another. For internal consistency reliability evidence based on local groups is preferable, however, as this evidence is more accurate and usually easy to get. For some guidelines with respect to establishing equivalence see the introduction of the section on Validity. An aide memoire of critical points for comment when an instrument has been translated and/or adapted from a non-local context is included in the Appendix.

It is difficult to set clear criteria for rating the technical qualities of an instrument. These notes provide some guidance on the values to be associated with inadequate, adequate, good and excellent ratings. However these are intended to act as guides only. The nature of the instrument, its area of application, the quality of the data on which reliability estimates are based, and the types of decisions that it will be used for should all affect the way in which ratings are awarded. Under some conditions a reliability of 0.70 is fine; under others it would be inadequate. For these reasons, summary ratings should be based on your judgment and expertise as a reviewer and not simply derived by averaging sets of ratings.

In order to provide some idea of the range and distribution of values associated with the various scales that make up an instrument, enter the *number of scales* in each section. For example, if an instrument being used for *group-level decisions* had 15 scales of which five had retest reliabilities lower than 0.6, six between 0.60 and 0.70 and the other four in the 0.70 to 0.80 range, the median stability could be judged as 'adequate' (being the category in which the median of the 15 values falls). If more than one study is concerned, first the median value per scale should be computed, taking the sample sizes into account; in some cases results from a meta-analysis may be available, these can be judged in the same way. This would be entered as:

Stability	Number of *scales (if applicable)*	M*
No information given	[-]	0
Inadequate (e.g. $r < 0.60$)	[5]	1
Adequate (e.g. $0.60 \leq r < 0.70$)	[6]	2
Good (e.g. $0.70 \leq r < 0.80$)	[4]	3
Excellent (e.g. $r \geq 0.80$)	[0]	4

* M = median stability

For each of the possible ratings example values are given *for guidance only* - especially the distinctions between 'Adequate', 'Good' and 'Excellent'. For *high stakes decisions*, such as personnel selection, these

example values will be .10 higher. However, it needs to be noted that decisions are often based on aggregate scale scores. Aggregates may have much higher reliabilities than their component primary scales. For example, primary scales in a multi-scale instrument may have reliabilities around 0.70 while Big Five secondary aggregate scales based on these can have reliabilities in the 0.90s. Good test manuals will report the reliabilities of secondary as well as primary scales.

It is realised that it may be impossible to calculate actual median figures in many cases. What is required is your best estimate, given the information provided in the documentation. There is space to add comments. You can note here any concerns you have about the accuracy of your estimates. For example, in some cases, a very high level of internal consistency might be commented on as indicating a 'bloated specific'.

10	**Reliability**	
10.1	Data provided about reliability *(select two if applicable)*	
	No information given	[ ]
	Only one reliability coefficient given (for each scale or subscale)	[ ]
	Only one estimate of standard error of measurement given (for each scale or subscale)	[ ]
	Reliability coefficients for a number of different groups (for each scale or subscale)	[ ]
	Standard error of measurement given for a number of different groups (for each scale or subscale)	[ ]
10.2	**Internal consistency** The use of internal consistency coefficients is not sensible for assessing the reliability of speed tests, heterogeneous scales (also mentioned empirical or criterion-keyed scales; Cronbach, 1970), effect indicators (Nunnally & Bernstein, 1994) and emergent traits (Schneider & Hough, 1995). In these cases all items concerning internal consistency should be marked *'not applicable'*. It is also biased as a method for estimating reliability of ipsative scales. Alternate form or retest measures are more appropriate for these scale types. Internal consistency coefficients give a better estimate of reliability than split-half coefficients corrected with the Spearman-Brown formula. Therefore, the use of split-halves is only justified if, for any reason, information about the answers on individual items is not available. Split-half coefficients can be reported in item 10.7 (Other methods).	
10.2.1	Sample size	
	Not applicable	n/a
	No information given	0
	One inadequate study (e.g. sample size less than 100)	1
	One adequate study (e.g. sample size of 100-200)	2
	One large (e.g. sample size more than 200) or more than one adequate sized study	3
	Good range of adequate to large studies	4

10.2.2	Kind of coefficients reported *(select as many as applicable)*		
	Not applicable		n/a
	Coefficient alpha or KR-20		[ ]
	Lambda-2		[ ]
	Greatest lower bound		[ ]
	Omega (factor analysis)		[ ]
	Theta (factor analysis)		[ ]
	Other, describe:		[ ]
10.2.3	Size of coefficients	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.70$)	[ ]	1
	Adequate (e.g. $0.70 \leq r < 0.80$)	[ ]	2
	Good (e.g. $0.80 \leq r < 0.90$)	[ ]	3
	Excellent (e.g. $r \geq 0.90$)	[ ]	4
10.2.4	Reliability coefficients are reported with samples which .... (*select one*)		
	.... do not match the intended test takers, leading to more favourable coefficients (e.g. inflation by artificial heterogeneity)		[ ]
	.... do not match the intended test takers, but the effect on the size of the coefficients is unclear		[ ]
	.... do not match the intended test takers, leading to less favourable coefficients (e.g. reduction by restriction of range)		[ ]
	.... match the intended test takers		[ ]
	Not applicable		n/a
10.3	**Test retest reliability – temporal stability** Test retest refers to relatively short time intervals, whereas temporal stability refers to longer intervals in which more change is acceptable. Particularly for tests to be used for predictions over longer periods both aspects are relevant. To assess the temporal stability more than one retest may be required. The use of a test retest design is not sensible for assessing the reliability of state measures (actually a high test retest coefficient would invalidate the state character of a test). In this case all items concerning test retest reliability should be marked '*not applicable*'.		

10.3.1	Sample size		
	Not applicable		n/a
	No information given		0
	One inadequate study (e.g. sample size less than 100)		1
	One adequate study (e.g. sample size of 100-200)		2
	One large (e.g. sample size more than 200) or more than one adequate sized study		3
	Good range of adequate to large studies		4
10.3.2	Size of coefficients	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.60$)	[ ]	1
	Adequate (e.g. $0.60 \leq r < 0.70$)	[ ]	2
	Good (e.g. $0.70 \leq r < 0.80$)	[ ]	3
	Excellent (e.g. $r \geq 0.80$)	[ ]	4
10.3.3	Data provided about the test-retest interval (*select or fill in test-retest interval*)		
	Not applicable		n/a
	No information given		[ ]
	The interval is:		.....
10.3.4	Reliability coefficients are reported with samples which .... (*select one*)		
	.... do not match the intended test takers, leading to more favourable coefficients (e.g. inflation by artificial heterogeneity)		[ ]
	.... do not match the intended test takers, but effect on size of coefficients is unclear		[ ]
	.... do not match the intended test takers, leading to less favourable coefficients (e.g. reduction by restriction of range)		[ ]
	.... match the intended test takers		[ ]
	Not applicable		n/a

10.4	**Equivalence reliability (parallel or alternate forms)**		
10.4.1	Sample size		
	Not applicable		n/a
	No information given		0
	One inadequate study (e.g. sample size less than 100)		1
	One adequate study (e.g. sample size of 100-200)		2
	One large (e.g. sample size more than 200) or more than one adequate sized study		3
	Good range of adequate to large studies		4
10.4.2	Are the assumptions for parallelism* met for the different versions of the test for which equivalence reliability is investigated? *Note that tests can be considered to be parallel tests if in the same group the mean scores, variances and correlations with other tests are the same.		
	Not applicable		n/a
	No information given		0
	Inadequate		1
	Adequate		2
	Good		3
	Excellent		4
10.4.3	Size of coefficients	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.70$)	[ ]	1
	Adequate (e.g. $0.70 \leq r < 0.80$)	[ ]	2
	Good (e.g. $0.80 \leq r < 0.90$)	[ ]	3
	Excellent (e.g. $r \geq 0.90$)	[ ]	4
10.4.4	Reliability coefficients are reported with samples which .... (*select one*)		
	.... do not match the intended test takers, leading to more favourable coefficients (e.g. inflation by artificial heterogeneity)		[ ]
	.... do not match the intended test takers, but effect on size of coefficients is unclear		[ ]

	.... do not match the intended test takers, leading to less favourable coefficients (e.g. reduction by restriction of range)	[ ]
	.... match the intended test takers	[ ]
	Not applicable	n/a
10.5	**IRT based method**	
10.5.1	Sample size It is difficult to give uniform guidelines for the adequacy of sample sizes in case IRT methods for the estimation of reliability are used, because the requirements are different in function of the item response format and the item response model used. Dependent on the item response model used minimum values for 'adequate' sample sizes are: 200 for 1-parameter studies, 400 for 2-parameter studies, and 700 for 3-parameter studies (based on Parshall, Davey, Spray, & Kalohn, 2001). These values apply to dichotomous models, but can be of some guidance for the reviewer when polytomous models are used for which the sample sizes may be smaller.	
	Not applicable	n/a
	No information given	0
	One inadequate study	1
	One adequate study	2
	One large or more than one adequate sized study	3
	Good range of adequate to large studies	4
10.5.2	Kind of coefficients reported (*select as many as applicable*) The first method gives the reliability of the estimated latent trait which in IRT replaces the estimated true score, i.e. test score (see Embretson & Reise, 2000). The second method is based on information about the individual items and gives an estimate of the reliability when the requirements typical for IRT are met (Mokken, 1971). The third method gives an estimate of the accuracy of the measurement related to the position on the latent trait.	
	Reliability of the estimated latent trait	[ ]
	Rho	[ ]
	Information function	[ ]
	Others, describe:	[ ]
	Not applicable	n/a

10.5.3	Size of coefficients (based on the final test length) Both guidelines for reliability coefficients (including rho) as for the information function are given. The guidelines for the information function are based on those for reliability coefficients since Information = $1/SE^2$, and given some often made assumptions, $r = 1 - SE^2$. Note that SE and information values are dependent on the value of the latent trait and that each test has a range within which the information value is optimal. The rating should not a priori be based on this optimal value, but on the information value of the score or range of scores that are of specific importance (e.g., critical scores). For these scores the information value may be optimal, but not necessarily so. If there are no such scores, the rating should be based on the mean information value (see also Reise & Havilund, 2005). Because there is not much experience with these rules-of-thumb, we advise raters to use these rules with care.	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.70$; information < 3.33)	[ ]	1
	Adequate (e.g. $0.70 \leq r < 0.80$; $3.33 \leq$ information < 5.00)	[ ]	2
	Good (e.g. $0.80 \leq r < 0.90$; $5.00 \leq$ information < 10.00)	[ ]	3
	Excellent (e.g. $r \geq 0.90$; information ≥ 10.00)	[ ]	4
10.6	**Inter-rater reliability** If the scoring of a test involves no judgmental processes (e.g. simply summing the scores of multiple-choice items), this type of reliability is not required and all items concerning inter-rater reliability should be marked *'not applicable'*. Note that although inter-rater reliability may not apply to the test as a whole, it may apply to one or more subtests (e.g. some subtests of an intelligence test).		
10.6.1	Sample size		
	Not applicable		n/a
	No information given		0
	One inadequate study (e.g. sample size less than 100)		1
	One adequate study (e.g. sample size of 100-200)		2
	One large (e.g. sample size more than 200) or more than one adequate sized study		3
	Good range of adequate to large studies		4

10.6.2	Kind of coefficients reported (*select as many as applicable*)		
	Not applicable		n/a
	Percentage agree		[ ]
	Coefficient Kappa		[ ]
	Intra Class Correlation		[ ]
	Coefficient Iota		[ ]
	Other, describe:		[ ]
10.6.3	Size of coefficients To some methods mentioned in 10.6.2 the guide numbers may not apply as no *r*'s are computed.	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.60$)	[ ]	1
	Adequate (e.g. $0.60 \leq r < 0.70$)	[ ]	2
	Good (e.g. $0.70 \leq r < 0.80$)	[ ]	3
	Excellent (e.g. $r \geq 0.80$)	[ ]	4
10.7	**Other methods of reliability estimation**		
10.7.1	Sample size		
	Not applicable		n/a
	No information given		0
	One inadequate study (e.g. sample size less than 100)		1
	One adequate study (e.g. sample size of 100-200)		2
	One large (e.g. sample size more than 200) or more than one adequate sized study		3
	Good range of adequate to large studies		4
10.7.2	Describe method:		
10.7.3	Results	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0

	Inadequate	[ ]	1
	Adequate	[ ]	2
	Good	[ ]	3
	Excellent	[ ]	4
10.8	**Overall Adequacy** This overall rating is obtained by using judgment based on the ratings given for items 10.1 – 10.7.3. *Do not simply average numbers to obtain an overall rating.* For some instruments, internal consistency may be inappropriate (broad traits or scale aggregates), in which case more emphasis on the retest data should be placed. In other cases (state measures), retest reliabilities would be inappropriate, so emphasis should be placed on internal consistencies. For your final judgment you should also take into account: – whether the test is used for individual assessment or to make decisions on groups of people – the nature of the decision (high-stakes vs. low-stakes) – whether one or more (types of) reliability studies are reported – whether also standard errors of measurement are provided – procedural issues, e.g. group size, number of reliability studies, heterogeneity of the group(s) on which the coefficient are computed, number of raters if inter-rater agreement is computed, length of the test-retest interval, etc. – comprehensiveness of the reporting on the reliability studies.		
	No information given		0
	Inadequate		1
	Adequate		2
	Good		3
	Excellent		4

Reviewers' comments on Reliability: Underline the strong and weak aspects of the evidence of reliability available. Comments pertaining to equivalence/reliability generalisation should also be made here (if applicable).

11 Validity

General guidance on assigning ratings for this section

Validity is the extent to which a test serves its purpose: can one draw the conclusions from the test scores which one has in mind? In the literature many types of validity are differentiated, e.g. Drenth and Sijtsma (2006, p. 334 – 340) mention eight different types. The differentiations may have to do with the purpose of validation or with the process of validation by specific techniques of data analysis. In the last decades of the past century there was a growing consensus that validity should be considered as a unitary concept and that differentiations in types of validity should be considered as different ways of gathering evidence only (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Borsboom, Mellenbergh, and Van Heerden (2004) state that a test is valid for measuring an attribute if variation in the attribute causally produces variation in the measured outcomes. Although this is a different approach, also in the opinion of these authors a differentiation between types a validity is not relevant.

However, whichever approach to validity one prefers, for a standardised judgment it is necessary to structure the concept of validity a bit. For this reason, separate sub-sections on construct and criterion validity are differentiated. Depending on the purpose of the test one of these aspects of validity may be more relevant than the other. However, it is realized that construct validity is the more fundamental concept and that evidence on criterion validity may add to establishing the construct validity of a test.
It is realized also, that a test may have different validities depending on the type of decisions made with the test, the type of samples used, etc. However, inherent in a test review system is that one quality judgment is made about *the* (construct or criterion) validity of a test. This judgment should be a reflection of the quality of the evidence supporting the claim that the test can be used for the interpretations that are stated in the manual. The broader the intended applications, the more validity evidence the author/publisher should deliver. Note that the final rating for construct and criterion validity will be a kind of average of this evidence and that there may be situations or groups for which the test may have higher or lower validities (or for which the validity may not have been studied at all).

When an instrument has been translated and/or adapted from a non-local context, evidence of equivalence of the measure in a new language to the original should be proposed. Without this it is not possible to generalise findings in one country/language version to another. Examples of equivalent evidence:

- Invariance in construct structure – e.g. via factor structure or correlation with standard measures.
- Similar criterion related validity – e.g. similar profile of correlations of a multi-scale instrument with independent external criterion – such as ratings of job competencies.
- Items show similar patterns of scale loadings e.g. items correlate in same pattern with other scales; strongest/weakest loading items are similar in original and new languages.
- Bilingual candidates have similar profiles in two languages (c.f. alternate form reliability).

Validity generalisation needs stronger evidence when translating tests across linguistic families (e.g. from an Indo-European to a Semitic language). In such a situation equivalence is under greater threat because of the differences in language structure and cultural differences. However, validity generalisation might be inferred from evidence of validity invariance in previous translations when a test has been translated into multiple languages. For instance, if a Swedish test has already been translated into French, German and Italian and has been shown to have equivalence in these languages.
In considering the whole issue of equivalence, it may be useful to follow Van de Vijver and Poortinga's (2005) classification:

- Structural / functional equivalence

  - There is evidence that the source and target language versions measure the same psychological constructs across groups. This is generally demonstrated by showing that patterns of correlations between variables are the same across groups.
- Measurement unit equivalence
  - There is evidence that the measurement units are the same, but there are different origins across groups (i.e. individual differences found in group A can be compared with differences found in group B, but the absolute raw scores for A and B are not directly comparable without some form of re-scaling).
- Scalar / Full score equivalence
  - The same measurement unit and the same origin (i.e. raw scores have the same meanings and can be compared across groups).

The benchmarks and the notes in the sub-sections 11.1 and 11.2 provide some guidance on the values to be associated with inadequate, adequate, good and excellent ratings. However these are intended to act as guides only. The nature of the instrument, its area of application, the quality of the data on which validity estimates are based, and the types of decisions that it will be used for should all affect the way in which ratings are awarded. For validity, guidelines on sample sizes are based on power analysis of the sample sizes needed to find moderate sized validities if they exist.

11.1 Construct validity

The purpose of construct validation is to find an answer to the question whether the test actually measures the intended construct or, partly or mainly, something else. Common methods for the investigation of construct validity are exploratory or confirmatory factor analysis, item-test correlations, comparison of mean scores of groups for which score differences may be expected, testing for invariance of factor structure and item-bias (DIF) for different groups, correlations with other instruments which are intended to measure the same (convergent validity) or different constructs (discriminant validity), Multi-Trait-Multi-Method research (MTMM), IRT-methodology and (quasi-)experimental designs.

11.1	**Construct validity**	
11.1.1	Designs used *(select as many as are applicable)*	
	No information is supplied	[ ]
	Exploratory Factor Analysis	[ ]
	Confirmatory Factor Analysis	[ ]
	(Corrected) item-test correlations	[ ]
	Testing for invariance of structure and differential item functioning across groups	[ ]
	Differences between groups	[ ]
	Correlations with other instruments and performance criteria	[ ]
	MTMM correlations	[ ]

	IRT methodology	[ ]
	(Quasi-)Experimental Designs	[ ]
	Other, describe:	[ ]
11.1.2	Do the results of (exploratory or confirmatory) factor analysis support the structure of the test?	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.3	Do the items correlate sufficiently well with the (sub)test score? Note that very high correlations may mean that items are more or less synonymous and that the concept measured may be very narrow (a so-called 'bloated specific')	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.4	Is the factor structure invariant across groups and/or is the test free of item-bias (DIF)? This kind of research can be carried out on basis of models within classical test theory or the IRT framework. If item-bias is found, the effect on the total score should be estimated (small effects are acceptable).	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

11.1.5	Are differences in mean scores between relevant groups as expected? E.g. pupils in group 8 are expected to score higher than pupils in group 6 on a test for numerical proficiency; children with the diagnosis ADHD should score higher on a test for hyperactivity than children not diagnosed with ADHD; salespersons should score higher on a test for commercial knowledge than the average working population. Even though the results are in the expected direction, this kind of research usually is inconclusive with respect to the construct validity of the test. However, the value of this kind of research is that when the expected differences are not shown, this would raise strong doubts about the construct validity of the test.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.6	Median and range of the correlations between the test and tests measuring similar constructs An essential element of the process of construct validation is correlating the test score(s) with scales from similar instruments, the so-called congruent or convergent validity. The guidelines on congruent validity coefficients need to be interpreted flexibly. Where two very similar instruments have been correlated (with data obtained concurrently) we would expect to find correlations of 0.60 or more for 'adequate'. Where the instruments are less similar, or administration sessions are separated by some time interval, lower values may be adequate. When evaluating congruent validity, care should be taken when interpreting very high correlations. When correlations are above 0.90, the likelihood is that the scales in question are measuring exactly the same construct. This is not a problem if the scales in question represent a new scale and an established marker. It would be a problem though, if the scale(s) in question was (were) meant to be adding useful variance to what other scales already measure. The guidelines given concern correlations that are not adjusted for common-method variance or attenuation. Therefore, also the reliabilities of both instruments should be taken into account when judging the congruent validity coefficients. E.g., when both instruments have a reliability of .75, the maximum correlation between the instruments is .56. If reliabilities are higher, higher correlations are to be expected.	
	No information given	0
	Inadequate ($r < 0.55$)	1
	Adequate ($0.55 \leq r < 0.65$)	2
	Good ($0.65 \leq r < 0.75$)	3
	Excellent ($r \geq 0.75$)	4
11.1.7	Do the correlations with other instruments show good discriminant validity with respect to constructs that the test is not supposed to measure?	
	No information given	0
	Inadequate	1

	Adequate	2
	Good	3
	Excellent	4
11.1.8	If a Multi-Trait-Multi-Method design is used, do the results support the construct validity of the test (does it really measure what it is supposed to measure and not something else)? Note that if an MTMM design is used, research as mentioned in 11.1.6 and 11.1.7 may not be required anymore.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.9	Other, e.g. IRT-methodology, (quasi-)experimental designs (describe):	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.10	Sample sizes The guidelines below concern studies within the classical test theory framework. For the estimation of item-parameters within IRT methodology 'adequate' sample sizes are: more than 200 for 1-parameter studies, more than 400 for 2-parameter studies and more than 700 for 3-parameter studies (based on Parshall, Davey, Spray, & Kalohn, 2001).	
	No information given	0
	One inadequate study (e.g. sample size less than 100)	1
	One adequate study (e.g. sample size of 100-200)	2
	One large (e.g. sample size more than 200) or more than one adequate sized study	3
	Good range of adequate to large studies	4
11.1.11	Quality of instruments as criteria or markers	
	No information given	0

	Inadequate quality	1
	Adequate quality	2
	Good quality	3
	Excellent quality with wide range of relevant markers for convergent and divergent validation	4
11.1.12	How old are the validity studies? It is difficult to formulate a general rule for taking the age of the research into account. For tests that intend to measure constructs in an area on which important theoretical developments have taken place, 15 year old research may be almost useless, whereas for other tests 20 year old (or even older) research still may be relevant.	
	Number of years	.....
11.1.13	**Construct validity - Overall adequacy** This overall rating is obtained by using judgment based on the ratings given for items 11.1.1 – 11.1.12. *Do not simply average numbers to obtain an overall rating.* In addition to the outcomes of the construct validity research, for your final judgment you should also take into account whether analysis techniques are used correctly (e.g. is the significance level corrected for correlating the instrument to other instruments without clear hypotheses, so-called 'fishing'), whether the research samples are similar to the group(s) for which the test is intended (e.g., more heterogeneity will inflate correlations, samples of students may give results that cannot be generalized), the size of the research sample(s), the quality of other instruments that are used (e.g. in convergent and discriminant validity research), and the age of the studies.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

11.2 Criterion-related validity

Criterion-related evidence of validity (concurrent and predictive validity) refers to studies where real-world criterion measures (i.e. not other instrument scores) have been correlated with scales. Predictive studies generally refer to situations where assessment was carried out at a 'qualitatively' different point in time to the criterion measurement - e.g. for a work-related selection measure intended to predict job success, the instrument would have been carried out at the time of selection - rather than just being a matter of how long the time interval was between instrument and criterion measurement. Studies can also be 'post-dictive', for example, where scores on a potential selection test are correlated with job incumbents' earlier line manager ratings of performance. Basically, evidence of criterion validity is required for all kinds of tests. However, when it is explicitly stated in the manual that test use does not serve prediction purposes (such as educational tests that measure progress), criterion validity can be considered 'not applicable'.

11.2	**Criterion-related validity**	
11.2.1	Type of criterion study or studies (*select as many as are applicable)*	
	Predictive	[ ]
	Concurrent	[ ]
	Post-dictive	[ ]
11.2.2	Sample sizes	
	No information given	0
	One inadequate study (e.g. sample size less than 100)	1
	One adequate study (e.g. sample size of 100-200)	2
	One large (e.g. sample size more than 200) or more than one adequate sized study	3
	Good range of adequate to large studies	4
11.2.3	Quality of criterion measures	
	No information given	0
	Inadequate quality	1
	Adequate quality	2
	Good quality	3
	Excellent quality with respect to reliability and representation of the criterion construct	4
11.2.4	Strength of the relation between the test and criteria It is difficult to set clear criteria for rating the size of the criterion validity coefficients of an instrument. A criterion-related validity of 0.20 can have considerable utility in some situations, while one of 0.40 might be of little value in others. A coefficient of .30 may be considered good in personnel selection, whereas in educational situations higher coefficients are common. For these reasons, ratings should be based on your judgment and expertise as a reviewer and not simply derived by averaging sets of correlation coefficients. The guidelines given are based on Hemphill (2003; see also Meyer et al., 2001) and concern correlations that are not corrected for attenuation in either the predictor or the criterion. However, coefficients may be corrected for restriction of range.	

	The ranges given below concern validity coefficients, because *correlations* between tests and criteria are the most used way to represent criterion validity. However, particularly for use in clinical situations data on the sensitivity and the specificity of a test may give more useful information on the relation between a test and a criterion. ROC-curves are a popular way of quantifying the sensitivity and specificity. Swets (1988) presents an overview of values of ROC-curves in different areas. For certain types of medical diagnosis the values are between .81 and .97, for lie detection between .70 and .95, and for educational achievement (pass/fail) between .71 and .94. These values may be used as guidelines, but it is left to the expertise of the reviewer to decide to what extent the test can make a useful contribution to the decision concerned. Also when still other indices are reported, such as the positive and negative predictive value of a test, the likelihood ratio, etc.	
	No information given	0
	Inadequate ($r < 0.20$)	1
	Adequate ($0.20 \leq r < 0.35$)	2
	Good ($0.35 \leq r < 0.50$)	3
	Excellent ($r \geq 0.50$)	4
11.2.5	How old are the validity studies? It is difficult to formulate a general rule for taking the age of the research into account. For tests that intend to predict behaviour in rapidly changing environments, 15 year old research may be almost useless, whereas for other tests 20 year old (or even older) research may still be relevant.	
	Number of years	.....
11.2.6	**Criterion-related validity – Overall adequacy** This overall rating is obtained by using judgment based on the ratings given for items 11.2.1 – 11.2.5. *Do not simply average numbers to obtain an overall rating.* Apart from the outcomes of the criterion validity research, for your final judgment you should also take into account whether the right procedures and analysis techniques are used (e.g. is there criterion contamination, correction for attenuation, cross-validation), whether the research samples are similar to the group(s) for which the test is intended (e.g. correction for restriction of range), the size of the research sample(s), the quality of the criterion instruments that are used (e.g. is there criterion deficiency), and the age of the studies.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

11.3 Overall validity

When judging overall validity, it is important to bear in mind the importance placed on construct validity as the best indicator of whether a test measures what it claims to measure. In some cases, the main evidence of this could be in the form of criterion-related studies. Such a test might have an 'adequate' or better rating for criterion-related validity and a less than adequate one for construct validity. In general the rating for Overall Validity will be equal to either the Construct Validity or the Criterion-related Validity, whichever is the greater. However, depending on the purpose of the test, one of these types of evidence may be considered more relevant than the other. The rating for Overall Validity should not be regarded as an average or as the lowest common denominator.

11.3	**Validity – Overall adequacy** This overall rating is obtained by using judgment based on the ratings given for items 11.1.1 – 11.2.6. *Do not simply average numbers to obtain an overall rating.*	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

Reviewers' comments on validity (all the evidence of validity included). Comments pertaining to equivalence/validity generalisation should also be made here (if applicable).

12 Quality of computer generated reports

Judging computer-based reports is made difficult by the fact that many suppliers will, understandably, wish to protect their intellectual property in the algorithms and scoring rules. In practice, sufficient information should be available for review purposes from the technical manual describing the development of the reporting process and its rationale, and through the running of a sample of test cases of score configurations. Ideally the documentation should also describe the procedures that were used to test the report generation for accuracy, consistency and relevance. For the purpose of reviewing at least three reports based on different score profiles including the actual scores should be provided, even if the algorithms for generating the reports are confidential.
For each of the following attributes, some questions are stated that should help you make a judgment, and a definition of an 'excellent' (4) rating is provided.

Items to be rated n/a or 0 to 4, 'benchmarks' are provided for an 'excellent' (4) rating.		
12.1	**Scope or coverage** Reports can be seen as varying in both their breadth and their specificity. Reports may also vary in the range of people for whom they are suitable. In some cases it may be that separate tailored reports are provided for different groups of recipients. • *Does the report cover the range of attributes measured by the instrument?* • *Does it do so at a level of specificity justifiable in terms of the level of detail obtainable from the instrument scores?* • *Can the 'granularity' of the report (i.e. the number of distinct score bands on a scale that are used to map onto different text units used in the report) be justified in terms of the scales measurement errors?* • *Is the report designed for the same populations of people for whom the instrument was developed? (e.g. groups for whom the norm groups are relevant, or for whom there is relevant criterion data etc.).*	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent: Excellent fit between the scope of the instrument and the scope of the report, with the level of specificity in the report being matched to the level of detail measured by the scales. Good use made of all the scores reported from the instrument.	4
12.2	**Reliability** • *How consistent are the reports in their interpretation of similar sets of score data?* • *If report content is varied (e.g. by random selection from equivalent text units), is this done in a satisfactory manner?* • *Is the interpretation of scores and the differences between scores justifiable in terms of the scale measurement errors?*	

	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent: Excellent consistency in interpretation and appropriate warnings provided for statements, interpretation and recommendations regarding their underlying errors of measurement.	4
12.3	**Relevance or validity** The linkage between the instrument and the content of the report may be explained either within the report or be separately documented. Where reports are based on clinical judgment, the process by which the expert(s) produced the content and the rules relating scores to content should be documented. • *How strong is the relationship between the content of the report and the scores on the instrument? To what degree does the report go beyond or diverge from the information provided by the instrument scores?* • *Does the report content relate clearly to the characteristics measured by the instrument?* • *Does it provide reasonable inferences about criteria to which we might expect such characteristics to be related?* • *What empirical evidence is provided to show that these relationships actually exist?* It is relevant to consider both the construct validity of a report (i.e. the extent to which it provides an interpretation that is in line with the definition of the underlying constructs) and criterion-validity (i.e. where statements are made that can be linked back to empirical data).	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent: Relationship between the scales and the report content, with clear justifications provided.	4
12.4	**Fairness, or freedom from systematic bias** • *Is the content of the report and the language used likely to create impressions of inappropriateness for certain groups?* • *Does the report make clear any areas of possible bias in the results of the instrument?* • *Are alternate language forms available? If so, have adequate steps been taken to ensure their equivalence?*	
	No information given	0
	Inadequate	1
	Adequate	2

	Good	3
	Excellent: Clear warnings and explanations of possible bias, available in all relevant user languages.	4
12.5	**Acceptability** This will depend substantially on the complexity of the language used in the report, the complexity of the constructs being described and the purpose for which it is intended. • *Is the form and content of the report likely to be acceptable to the intended recipients?* • *Is the report written in a language that is appropriate for the likely levels of numeracy and literacy of the intended reader?*	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent: Very high acceptability, well-designed and well-suited to the intended audience.	4
12.6	**Length** This is also an aspect of Practicality and should be reflected in the rating given for this, but too long reports may also be an indication of over-interpretation of scores. Therefore the length of reports is rated separately also. Generally reports that on average take more than one page per scale (excluding title pages, copyright notices etc.) may be over long and over-interpreted.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
12.7	**Overall adequacy of computer generated reports** This overall rating is obtained by using judgment based on the ratings given for items 12.1 –12.6. *Do not simply average numbers to obtain an overall rating.*	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

Reviewers' comments on computer generated reports

The evaluation can consider additional matters such as whether the reports take into account any checks of consistency of responding, response bias measures (e.g. measures of central tendency in ratings) and other indicators of the confidence with which the person's scores can be interpreted.

Comments on the complexity of the algorithms can be included, e.g. whether multiple scales are considered simultaneously, how scale profiles are dealt with etc. Such complexity should, of course, be supported by a clear rationale in the manual.

13 Final evaluation

Evaluative report of the test

This section should contain a concise, clearly argued judgment about the test. It should describe its pros and cons, and give some general recommendations about how and when it might be used - together with warnings (where necessary) about when it should not be used.

A summary of any positive or negative points raised in connection with adapted and translated tests should be summarised here. A checklist of the important considerations for such instruments is added in the Appendix as a reminder of the notes in the relevant sections. Only comment on these if this is appropriate.

The evaluation should cover topics such as the appropriateness of the instrument for various assessment functions or areas of application; any special training needs or special skills required; whether training requirements are set at the right level; ease of use; the quality and quantity of information provided by the supplier and whether there is important information which is not supplied to users and where there are issues arising from the instrument being translated or adapted (see Appendix).

Include comments on any research that is known to be under way, and the supplier's plans for future developments and refinements etc.

Conclusions

Recommendations *(select one)* The relevant recommendation, from the list given, should be indicated. Normally this will require some comment, justification or qualification. A short statement should be added relating to the situations and ways in which the instrument might be used, and warnings about possible areas of misuse. **All the characteristics listed below should have ratings of either n/a, 2, 3, or 4 if an instrument is to be 'recommended' for general use (box 4 or 5).** 9 Norms 10 Reliability–overall 11 Validity-overall 12 Computer generated reports If any of these ratings are 0 or 1 the instrument will normally be classified under Recommendation 1, 2, or 3 or it will be classified under 'Other' with a suitable explanation given.	1 Requires further development. Only suitable for use in research, not for use in practice	[ ]
	2 Only suitable for use by an expert user (exceeding EFPA User Qualification Level 2) under carefully controlled conditions or in very limited areas of application	[ ]
	3 Suitable for supervised use in the area(s) of application defined by the distributor by any user with general competence in test use and test administration (exceeding EFPA User Qualification Level 2)	[ ]
	4 Suitable for use in the area(s) of application defined by the distributor, by test users who meet the distributor's specific qualifications requirements (at least EFPA User Qualification Level 2)	[ ]
	5 Suitable for unsupervised self-assessment in the area(s) of application defined by the distributor	[ ]
	6 Other	[ ]

PART 3 BIBLIOGRAPHY

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. *European Journal of Psychological Assessment, 12*, 62–71.

Bartram, D. (2002a). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Evaluation Form.* Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

Bartram, D. (2002b). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Notes for Reviewers.* Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

Bartram, D., & Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the Internet.* Chichester, UK: Wiley and Sons.

Bartram, D., Lindley, P. A., & Foster, J. M. (1990). *A review of psychometric tests for assessment in vocational training.* Sheffield, UK: The Training Agency.

Bartram, D., Lindley, P. A., & Foster, J. M. (1992). *Review of psychometric tests for assessment in vocational training.* BPS Books: Leicester.

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering* [On the use of continuous norming]. Arnhem, The Netherlands: Cito.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the Internet* (pp. 201-217). Chichester, UK: Wiley and Sons.

Brennan, R. L. (Ed.) (2006). *Educational measurement.* Westport, CT: ACE/Praeger.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development.* Hillsdale, NJ: Erlbaum.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (pp. 471-515).Westport, CT: ACE/Praeger.

Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* (4e herziene druk) [Test theory. Introduction in the theory and application of psychological tests (4th revised ed.)]. Houten, The Netherlands: Bohn Stafleu van Loghum.

Embretson, S. E. (Ed.) (2010*). Measuring psychological constructs. Advances in model-based approaches.* Washington, D. C.: American Psychological Association.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.*Mahwah, NJ: Erlbaum.

Evers, A. (2001a). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing, 1*, 137–153.

Evers, A. (2001b). The revised Dutch rating system for test quality. *International Journal of Testing, 1*, 155–182.

Evers, A., Braak, M., Frima, R., & van Vliet-Mulder, J. C. (2009-2012). *Documentatie van Tests en Testresearch in Nederland* [Documentation of Tests and Testresearch in The Netherlands]. Amsterdam: Boom test uitgevers.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de Kwaliteit van Tests (geheel herziene versie; gewijzigde herdruk)* [COTAN Rating system for test quality (completely revised edition; revised reprint)]. Amsterdam: NIP.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., et al. (2012). Testing practices in the 21st Century: Developments and European psychologists' opinions. *European Psychologist*, in press.

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure and results. *International Journal of Testing, 10,* 295-317.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309-334.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*, 355–366.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Erlbaum.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78-80.

International Test Commission. (2005). *International Guidelines on Computer-Based and Internet Delivered Testing.* Bruxelles, Belgium: Author.

Kersting, M. (2008). DIN Screen, Version 2. Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen [DIN Screen, Version 2. Guide line for monitoring and optimizing the quality of instruments and their application in proficiency assessment procedures.]. In M. Kersting. *Qualitätssicherung in der Diagnostik und Personalauswahl - der DIN Ansatz* (S. 141-210) [Guaranteeing quality in diagnostics and personnel selection (p. 141-210)]. Göttingen: Hogrefe.

Lindley, P. A. (2009). *Reviewing translated and adapted tests: Notes and checklist for reviewers:5 May 2009.* Leicester, UK: British Psychological Society. Retrieved from http://www.efpa. eu/professional-development/tests-and-testing.

Lindley, P.A. (2009, July). Using EFPA Criteria as a common standard to review tests and instruments in different countries. In D.Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Lindley, P., Bartram, D., & Kennedy, N. (2004). *EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.3.* Leicester, UK: British Psychological Society (November, 2004).

Lindley, P., Bartram, D., & Kennedy, N. (2005). *EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.41.* Brussels: EFPA Standing Committee on Tests and Testing (August, 2005).

Lindley, P., Bartram, D., & Kennedy, N. (2008). *EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.42.* Brussels: EFPA Standing Committee on Tests and Testing (September, 2008).

Lindley, P. A. (Senior Editor), Cooper, J., Robertson, I., Smith, M., & Waters, S. (Consulting Editors). (2001). *Review of personality assessment instruments (Level B) for use in occupational settings. 2nd Edition.* Leicester, UK: BPS Books.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-165.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Moosbrugger, H., Kelava, A., Hagemeister, C., Kersting, M., Lang, F., Reimann, G., et al. (2009, July). The German Test Review System (TBS-TK) and first experiences. In D. Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2*, 65-72.

Muñiz, J., & Bartram, D. (2007). Improving international tests and testing. *European Psychologist, 12*, 206-219.

Nielsen, S. L. (2009, July). Test certification through DNV in Norway. In D. Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Parshall, C. G., Spray, J. A., Davey, T., & Kalohn, J. (2001). *Practical Considerations in Computer-based Testing.* New York: Springer Verlag.

Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España [A model for the evaluation of test quality in Spain]. *Papeles del Psicólogo*, *77*, 65–71.

Reise, S. P., & Havilund, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Measurement, 84*, 228-238.

Tideman, E. (2007). Psychological tests and testing in Sweden. *Testing International*, *17*(June), 5–7.

Schneider, R. J., & Hough, L. M. (1995). Personality and industrial/organizational psychology. In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology, 10*, 75-129.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7*, 301-317.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

Testkuratorium. (2009). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009 [TBS-TK. Test review system of the board of testing of the Federation of German psychologists' associations]. *Report Psychologie, 34*, 470-478.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Erlbaum.

Van der Linden, W. J., & Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing.* London: Springer.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Ziegler, M., MacCann, C., & Roberts, R. (Eds.) (2011). *New perspectives on faking in personality assessment.* Oxford, UK: Oxford University Press.

APPENDIX

An aide memoire of critical points for comment when an instrument has been translated and/or adapted from a non-local context

Development	
Evidence or discussion of	*Input from native speakers of new language*
	Multiple review by both language and content (of test) experts
	Back translation from new language into original language
Basic psychometric properties	*Item performance*
	Reliability
Norms	
	A local norm is provided
Non-local norm	*Strong evidence of equivalence for both test versions and samples*
International norms	*Larger than the typical requirements of local samples*
The nature of the sample	*Balance of sources of the sample*
	Equivalence of the background of the different parts of the sample
The type of measure	*Little or no verbal content*
The equivalence of the test version	*All the language versions are well translated/adapted*
	Some groups have completed the test in a non-primary language
Similarities of scores in different samples	*Where there are large differences these should be accounted for and the implications in use discussed*
Guidance about generalising the norms	
Equivalence/ Reliability/Validity	
Invariance in construct structure	*Via factor structure, equivalence of correlation matrices or similarity of patterns of correlation with standard measures*
Similar criterion related validity	*Strongest correlation with similar competencies*
Similar patterns of scale loadings	*Items correlate in same pattern with other scales*
	Strongest/weakest loading items are similar in original and new languages
Alternate form reliability	*Bilingual candidates have similar profiles in two languages*
Validity generalisation	
Validity generalisation needs strong evidence	*When translating tests across linguistic families (e.g. from an Indo-European to a Semitic language*
Validity generalisation can be inferred	*Where a test has been translated into multiple languages some validity generalisation can be inferred from evidence of validity invariance in previous translations: Swedish test has already been translated into French, German and Italian and has been shown to have equivalence in these languages*

B

Percentil tabell 1 oppdatert versjon 2015

Percentil Rank	Kartlegging 1. Høst											
	Trinn2			Trinn3			Trinn4			Trinn5		
	K 1.1	K 1.2	K 1.3	K 1.1	K.1.2	K 1.3	K 1.1	K 1.2	K 1.3	K1.1	K 1.2	K 1.3
100	>=121	>=163	>=134	>=185	>=199	>=197	>=195	>=215	>=204	>=193	>=185	>=214
99	120-117	162-127	133-116	184-161	198-176	196-166	194-170	214-195	203-184	193-177	184-169	213-195
98	116-102	126-119	115-99	160-144	175-154	165-155	169-164	194-187	183-175	176-167	168-167	194-187
97	101-95	118-100	98-90	143-137	153-147	154-147	163-162	186-175	174-166	167-165	166-163	186-185
96	94-93	99-97	89-81	136-129	146-141	146-142	160-155	174-170	165-159	164	162	184
95	92-88	96	80-78	128-125	140	141-139	154	169-168	158-157	163	161-157	183-182
94	87-86	95-91	77	124-121	139-135	138-131	153	167-165	156-151	162-159	156-154	181-179
93	85	90-88	76-75	120-119	134-132	130-127	152	164-163	150-148	158-156	153	178
92	84-81	87	74	118-116	131-126	126-123	151	162	147	155-154	152	177-176
91	80-79	86-84	73	115	125	122-120	150-149	161-160	146-145	153	151	175-172
90	78-77	83	72	114-113	124	119-118	148	159	144-141	152	151	171-170
89	76	82-81	71	112-111	123-122	117	147-146	158	140-137	151	150	169-167
88	75	80	70	110	121-120	116	145-142	157	136	150-149	149-148	166
87	74-73	79	69-68	109-108	119	115-113	141-140	156-155	136	148-146	147-146	165

Percentil tabell 1 oppdatert versjon 2015

86	72	78	67	107-106	118	112-110	139-138	154	135	145-144	145-143	164
85	71-70	77	66	105	117-116	109	136-137	153-152	134	143	142-140	164
84	69	76	65	104	115-114	108-107	135	151-150	133-132	142-141	139	163
83	68	75	64	104	113-112	106	134-133	149-145	131	140	138	162
82	67	74	63	103-102	111-110	105-103	132-131	144	130	139-138	137	161-160
81	66	73-72	62	101-99	109	102	130	143	129	137-136	136	159-157
80	65	71-70	62	98-97	108	101-100	129	142	128	135	135	156
79	64	69	61	96	107	99-98	128	141	127	134-133	135	155
78	63	69	60-59	95-94	106	97-96	127	140	126	133-132	134	154-153
77	62	68-67	58	93-92	105	95	126	139	125	131	133	152
76	62	66	57	91-89	104-103	94	125	138	125	130-129	132-131	151
75	61-59	**65**	56	88	102	93-92	125	137	124	128	130-129	150
74	58-56	64-62	55	87	101	91-90	124-123	137	123	127	128	150
73	55	61	54-52	87	100	91-90	122-121	136-135	122	126	127	149-148
72	54	61	51	86	99-98	89	120-119	134	121	125	126	147
71	53	60-57	50	85	67-96	88	118	133	120	124	126	146
70	52	56-54	49	84-83	95-94	88	118	132	119	123	126	145-143

Percentil tabell 1 oppdatert versjon 2015

69	51	53	48-47	82	93-90	87	117	131	118	122	125	142
68	51	52	47-46	81-80	89-88	86-85	116	130	117	122	125	141
67	50	51	45	79-78	87	84-83	115	129-128	116	121	124	140
66	49-47	50	44	77-76	86	82	115	127-126	115-114	120	123	139
65	46-45	49	44	75	85-84	81	114	125	113	119	122-121	138
64	46-45	48-47	43	75	83-82	80	113	124	112	118	120-119	137
63	44	46	42	74	81	79	112	123	111	117	118	136
62	43-42	45	41	73	81	78	111-110	123	110-109	116	117	135
61	41	44	40	72-71	80	77	109-108	122	108	116	116	135
60	40	43	39	70	79-78	76	107	121	107	115	116	134
59	39	42	38	70	77	75-74	107	121	106	114	116	133
58	38	41	37	69	76	73	106	120	106	114	115	132
57	37	40	36	68	75	72	105	120	105	113	114-113	131-130
56	36-35	39	35	67	75	71	104	119	104	112	112	129-127
55	34	38	35-34	66-65	74	70-69	103	118	103	112	112	126
54	33	38	33	66-65	73	68-67	103	117	102	111	111	125
53	32	37	33	65-64	73	66	102	117	101	110	110	125

Percentil tabell 1 oppdatert versjon 2015

52	32	36	32	64-63	72	66	101	116	101	109	109	124
51	31	36	32	64-63	72	65	100	115-114	100	109	109	123-122
50	30	36	31	62-61	71	64	100	113	100	108	108	121
49	29	35	30	60	70	63	99	113	99	107	107	120
48	28	35	30	59	69	62	99	112-111	98	106	106-104	120
47	27	35	30	59	69	61	99	110	98	105	103	119
46	27	34	29	58	68-66	60	98-97	109	97	105	102	119
45	26	34	28	58	65	59	96	109	97	104	102	118-117
44	26	33	27	57	64-63	58	95	108	96	103	102	116
43	25	33	26	56	62	57	94	107	95	103	101	116
42	25	32	25	55	61	56	93	106-105	94-93	103	100	115
41	25	32	25	55	60-59	56	93	104	92	102	99	114
40	24	31	24	54-53	58	55	92	103-101	91	101	98-97	113
39	24	31	24	53	57	54-53	92	100	90	100	96	112
38	23	30	24	52	56-55	52-51	93-91	99-98	89	99	95	112
37	23	29	23	51	54	50-49	90	97	88	99	95	111
36	22	28	23	51	54	48-47	89	96	87	98	94	110-109

Percentil tabell 1 oppdatert versjon 2015

35	22	28	22	50	53	46-45	88	95-94	86	97	94	110-109
34	22	27	21	49	52	45	87	93	85	96	93	108
33	21	26	21	48	51	45	86	92	84-83	95	92	107
32	21	26	20	47	50	45	85	91	82-81	94	91	106
31	20	25	20	46	49	44	84	90	80	94-93	90	105-104
30	20	25	19	45	49	44	84	89	79-78	93	90	103
29	19	24	19	44-43	48	43	83-81	89	77	92	89	102
28	19	24	19	42	47	42	80	88-87	76-75	92	88-85	101
27	18	24	18	41	46	41	79-78	86	74-73	91	84-82	100
26	17	24	18	40	45	40	77-76	85	72-71	91	81	99-97
25	16	23	18	39	44	39	75	84-83	70-69	90	80	96-95
24	16	23	17	38	44	38	74-73	82-81	68	89	79	94
23	15	23	17	38	43-42	38	72	80	67	89	78	93
22	15	23	17	37	41	37	71	79	66-65	88-87	77	92
21	14	22	17	37	40	37	70	78	64-63	86	76-75	91-90
20	14	22	16	36	39	36	69-67	77	64-63	85	74	89-88
19	13	21	16	36	38	36	66	76	62	84	73	87

Percentil tabell 1 oppdatert versjon 2015

18	13	21	16	35	37	35	65	75-74	61	83-82	72	86-85
17	12	21	15	34-33	36	34	64	73-71	60-57	83-82	71	84-83
16	12	20	15	34-33	35-34	33	63-61	70	56	81	70	82
15	12	20	15	32	33	32-31	60	69	55-54	80	69	81-80
14	11	19	15	31	32	30	59-57	68	53-52	80	68	79
13	11	18	14	30-29	31	30	56	67-65	51-48	79	67	78-77
12	10	18	14	30-29	30	29	55-54	64	47	78	66-65	76-75
11	9	17	14	28	29-28	29	53-52	63	46-45	77	64	74-73
10	8	17	13	27-26	28-26	28	51	62-61	44	76	63	72
9	8	16	13	25-24	25	27-26	50-48	60-59	43	75	62-60	71-70
8	7	16	13	23-22	24-23	25-24	47-44	58-57	42	74-73	59-57	69-68
7	6	15	12	21	22	23-22	43-42	56-55	41-37	72-69	56-54	67
6	5	14	12	21	21	21	41-37	54-45	36-34	68-65	53-50	66-63
5	4	13	11	20-19	20-19	20-19	36-35	44-41	33-26	64-61	49	62-60
4	4	13	11	18	19-18	18-16	34-30	40-34	25-24	60-58	48-46	59-54
3	3	12-11	10	17-15	17-16	15-13	29-25	33-27	23-15	57-44	45-42	53-46
2	3	10-6	9-7	14-12	15-10	12-11	24-15	26-23	14-11	43-34	41-27	45-32

Percentil tabell 1 oppdatert versjon 2015

1	<=2	<=5	<=6	<=11	<=9	<=10	<=14	<=22	<=10	<=33	<=26	<=31

Percentil tabell 2 oppdatert versjon 2015

Percentil Rank	Kartlegging 2 Vinter											
	Trinn2			Trinn3			Trinn4			Trinn5		
	K 2.1	K 2.2	K.2.3	K 2.1	K 2.2	K 2.3	K 2.1	K 2.2	K 2.3	K 2.1	K 2.2	K. 2.3
100	>168	>163	>167	>208	>191	>199	>205	>214	>200	>206	>=208	>=227
99	167-136	162-149	166-137	207-158	190-156	198-162	204-187	213-202	199-188	205-185	207-196	226-197
98	135-128	148-126	136-121	157-149	155-154	161-156	186-173	201-195	187-174	184-183	195-191	196-188
97	127-124	125-119	120-119	148-147	153-151	155-152	172-171	194-192	173-169	182-180	190-182	187-182
96	123-117	118-108	118-109	146	150-147	151-144	170-164	191-190	168-162	179-175	181-175	181-179
95	116-115	107	108-105	145-141	146-141	143-142	163-161	189-186	161-160	174-172	174	178
94	114-112	106-104	104-403	140	140-138	141-140	160-159	185-182	159-157	171	173-171	177
93	111-108	103-102	102-101	139-138	137-134	139-138	158-157	181-179	156-155	170-169	170-168	176-174
92	107	101-100	100	137-133	133-129	137-134	156-154	178-177	154-152	168-167	167-166	173-172
91	106-105	99-95	99-98	132-131	128	133-132	153-152	176-175	151-149	166-165	166	171
90	104	94-93	97-96	130-128	127-126	131	151-149	174	148	164	165-163	170
89	103	92	95	127-126	125-122	130-126	148	174	147	163-162	162-161	169-168
88	102-100	91	94	125	121	125-124	147-144	173	146	161-159	160	167-166
87	99	90-89	93-92	124-123	120-119	123-122	143-141	172-171	145-143	158-155	159-158	165-164
86	98-97	88	91-90	122-120	118	121	140	170-169	142-140	154-153	157-156	163-161
85	96-94	87	89	119	117	120-119	139-138	168	139-138	152	155	160
84	93-92	86	88	118	116	118	137	167-164	137-136	151-150	154	159
83	91	85-84	87-85	117-116	115	117	136	163-162	135-134	149	153	158-156
82	90	83-82	84-83	115	114-112	116	136	161	133-132	148	152-151	155
81	89-87	81	82-80	114	111	115	135	160-158	131	148	150	155
80	86	80	79	113	111	114-112	134-133	157-156	130	147	149-146	154-153
79	85-84	79-78	78-77	112	110	111	132	155-154	129-128	146	145-143	152
78	83	77-76	76	111	109	110	131-129	153	127	145-144	142-140	151-150
77	82-81	75-73	75	110	108	109	128-126	152	126	143	139	149-148
76	80-79	72	74-73	109-108	108	108	125-124	151	125	142	138	147-146
75	78	71-69	72	107	107	107-106	123-121	150	124-123	141	137	145

Percentil tabell 2 oppdatert versjon 2015

74	77	68	71	106-105	107	105	120-119	149	122	140-139	136	144
73	76-74	67	70	104	106	104	118	148-146	121	138	136	144
72	73-72	67	69	103	105	103	118	145	120	137	135	143-142
71	71	67	68	102	104	102	118	144	119	137	135	141
70	70	66	67	101	103	101	117	144	118	136	134	140
69	69	65-64	67	100	102	100-99	117	143	117	135	134	139-137
68	68	63	66	100	101-100	98-97	116	142	116	134-133	133	136
67	67	63	65	99	99	96-95	115	141	116	132	133	136
66	66-65	62	64-63	98	98	94	114	140-139	115	132	132	135-134
65	64	61	62	97	97	93	113	138	114	131	132	133
64	63	60	61	96-95	96	92	113	137-136	113-112	130	131	132
63	62	59	60	94	96	92	112	135	111	129	131	131
62	61-60	58	59	94	95	91	111	134	110	128	130	131
61	59-58	57-56	58-57	93-92	94	90	110	133	110	128	130	130
60	57	55	56	91	93	89	109	132	109	127	129	129
59	56	55	56	91	93	88	108	131	109	127	129	129
58	55-54	54	55	90	92	87	107	130	109	126	128	128
57	53	53	54	89	91	86	106	129	108	126	128	128
56	52	52	53	89	90	85	106	128	108	125-124	127	127
55	51	52	52	88	89	84	105	127	107	123	126	127
54	50	51	51	87	88	84	105	126	106-105	122	125-124	126-125
53	49	50	50	87	88	84	104-103	125-124	104	122	123	124-123
52	48	49	49	86	87	83	102	123	104	121	122	122
51	47	48	49	85	86	83	101	123	104	120	121	121
50	46	47-46	48	85	85	82-81	100-99	122	103	119	120	120
49	45	45	47	84	85	80	98	122	103	118	119	119
48	45	45	47	84	84	79	97	121	102	118	118	118
47	44	44	46	83	83-82	78	96	121	101	118	118	117-116
46	43	43	45	82	81-79	77	96	120	101	117	117	115
45	43	43	44	81	78	76	95	120	100	116	117	115
44	42	42	44	81	77	76	94	119	100	115	116	114-113
43	42	42	43	80	76	75-74	94	118-117	99	115	115	112

Percentil tabell 2 oppdatert versjon 2015

42	41	41	43	79	75-74	73	93	116	98	115	114	111
41	41	41	42	78-77	73	72	92-91	115	97	114	113	110-108
40	40	40	41	76	72	71	90	114	97	113	112	107
39	39	39-38	40	75	71	71	89-88	113	96	113	112	106
38	38	37	39	74	71	70-69	87	112-111	95	112	111	106
37	37	36	39	73-72	70	68	87	110-109	94	112	110	105
36	36-34	35	38	71	70	68	86	108	93-92	111	109	104
35	33	34	37	70	69	67	85-83	108	91	110-109	108-107	104
34	32	34	36	69	68	66	82	107	90-89	108	106	103
33	31	34	36	68-67	67	65	81	106-105	88-87	107	106	102
32	31	33	35	66	66	64	81	104	86	106	105	101
31	31	33	34	65	65	64	80	103	85	105	104	100-99
30	30	32	33	64	64	63	79	102-101	84-83	104-103	103-102	98
29	30	31	33	63	63	62	78	100	82	102	101	97
28	29	31	32	62-61	62-61	61-60	78	99	81-80	101	100	96
27	29	30	31	60	60	59	77	99	79	101	100	95
26	28	29	31	59	60	58	77	98	79	100	99	95
25	27	28	30	58	59-58	57	76	98	79	99	98	94
24	27	28	29	57	57-56	56	75	97	78	98	98-97	94
23	26	27	29	56	55	55	74-73	96-95	77	97	97	93
22	26	27	28	55	54-52	54	72	94	77	96	96	92
21	25	26	27	55	51	53	71-69	94	76-75	95-94	95	91
20	24	26	27	54	50-49	52	68	93-91	74	93-92	94	90-89
19	24	25	26	53-52	48	51	67	90-89	73	91	93	88-87
18	23	25	26	51	48	50	67	88	73	90	92-91	86
17	22	24	26	50	47-45	49	66-65	87	72	89	90	85
16	21	24	25-24	49	44	48	64	86-85	71	88	89-88	84-83
15	21	23	23	48-47	43-42	47	63-61	84	70-69	87	87	82
14	20	22	22	46-45	41	46	60-59	83-80	68	86-85	86	81-80
13	20	21	22	44	40-39	46	58-57	79-76	67	84	85	79
12	19	21	21	43	38-37	45	56	75-74	66	83	84	78-76
11	18	20	20	42-41	36	44-43	55	73-71	65-63	82-80	83-80	75

Percentil tabell 2 oppdatert versjon 2015

10	17	20	19	40	35-34	42-41	54-53	70	62	79-76	79-78	74-72
9	17	19	18	39	33-32	40-39	52-51	69-68	61-60	75-74	77-76	71-70
8	16	18	18	39	31-30	38-37	50	67-66	59-56	73-68	75-73	69-65
7	15	17	17	38-37	29	36-35	49	65-62	55-54	67-66	72-71	64-61
6	14	16	16-15	36-32	28-27	34-33	48-46	61-54	53-50	65-64	70-67	60-59
5	13-12	15	14	31-27	26-25	32	45-44	53-49	49-44	63-58	66-61	58
4	11	14-13	13	26-24	24	31-29	43-36	48-43	43-42	57-55	60-57	57-53
3	10	12	12-11	23-22	23-22	28-21	35-34	42-37	41-36	54-52	56-46	52-43
2	9-7	11-10	10	21-20	21-17	20-15	33-27	36-30	35-33	51-44	45-38	42-31
1	<=6	<=9	<=9	<=19	<=16	<=14	<=26	<=29	<=32	<=43	<=37	<=30

Percentil tabell 3 oppdatert versjon 2015

Percentil Rank	Kartlegging 3 Vår - Råskåre											
	Trinn2			Trinn3			Trinn4			Trinn5		
	K 3. 1	K 3. 2	K 3. 3	K 3.1	K 3.2	K 3.3	K 3.1	K 3.2	K 3.3	K 3.1	K.3.2	K.3.3
100	>=170	>=173	>=168	>=221	>=230	>=179	>=222	>=233	>=209	>=228	>200	>194
99	169-154	172-145	167-147	220-185	229-177	178-170	221-208	232-213	208-200	227-192	199-191	193-189
98	153-143	144-137	146-133	184-170	176-162	169-166	207-200	212-203	199-192	191-189	190-187	188-185
97	142-138	16-126	132-128	169-167	161-158	165-159	199-195	202-197	191-187	188-187	186	184-181
96	137-131	125-122	127-125	166-164	157-151	158	194-193	196-194	186	186-180	185-183	180-179
95	130-127	121	124-121	163-161	150-148	157-153	192-187	193-191	185-182	179-178	182	178-176
94	126-123	120-115	120-119	160-158	147-145	152-151	186-185	190-187	181-178	177-175	181-180	175-173
93	122-120	114-113	118	157-155	144	150-149	184	186-184	177-175	174-172	179	172-170
92	119-118	112	117-114	154-153	143	148-147	183-181	183-180	174-172	171	178	169-168
91	117-115	111-110	113-110	152-149	142-140	146	180-178	179-178	171-170	170-169	177-175	167
90	114	109	109-106	148-147	139-137	145-144	177-176	177-175	169-166	168-165	174-173	166
89	113-111	108	105-103	146-144	136-134	143-140	175-174	174-173	165-164	164	172-171	165
88	110-109	107	102-100	143-136	133-132	139-133	173-172	172-171	163-161	163-162	170	164
87	108-107	106	99-97	135-134	131-129	132-130	171	170	160	161	169	163
86	106	105	96	133-132	128-127	129	170-167	169-168	159	160	168-167	163
85	105	104	95-94	131	126-123	128-127	166	167	158-156	159-757	166	162-161
84	104-103	103	93	130	122-121	126-125	165-164	166	155	156	165	160
83	104-103	102-101	92	129	120	124	163-161	165-164	154-152	155	164	159
82	102-101	100-99	91	128	119	123	160	163-162	151	154-151	163	158
81	100	98-97	90	127-126	118-117	122-121	159	161	151	150	162-160	157
80	99	96-93	89	125-123	116	120	158-156	160-158	150-147	149-148	159-158	156
79	98	92	88	122-121	115-114	119	155-154	157	146-145	148-146	157-156	155
78	97-96	91-90	87	120	113-112	119	153-152	156-155	144-143	145-144	155	154
77	95	89	86-85	119	111	118	151	154-153	142	143	154	153
76	94-93	88	84-83	118	110	117	151	152	141	142	153	152
75	92-91	87-86	82	117	109-107	116	150-149	151	140	141	152	151

Percentil tabell 3 oppdatert versjon 2015

74	90	85	81	117	106	115	148	150	139	141	151	150
73	89	84-83	80	116-115	106	115	147-145	150	138-136	140	150	149
72	88-87	82	80	114	105	114	144	149-148	135	139	149-148	149
71	86	82	79	114	104	113	143	147	134-133	138	147	148
70	85	81	78	113	103	112	142	146	132	137	147	147
69	84-83	80	78	112	102-101	112	141	145	131	136	146	146
68	82	79-78	77	111	100	111	141	144	130	135	145	145
67	82	77	76	111	99	110	140	143	129-128	135	144	145
66	81	77	75	110-109	99	110	139	142	127	134	143	144
65	80	76	75	108	98	109	138	141-140	127	133	142	144-143
64	79	75	74	107	97	108-107	137	139	126	133	141	142
63	78	74	73	107	96	106	137	139	125	132	140	142
62	77	73	72	106-105	95	106	136	138	124	132	139	141
61	77	72	71	104-103	94	105	135	138	123	131	138	140
60	76	71	70	102	94	104-102	134	137	122	131	137-136	140
59	75	70	70	101-100	93	102	134	137	121	131	135	139
58	74	69	69-68	99	92	101	133	136	120	130	135	139
57	73	68	67	98	92	100	132	135	119	130	134	138
56	72	68	66	97	91	99	131	134	119	129	133	138
55	71-70	67	66	96	90	99	130	133	118	129	132	137
54	69	66-65	65-64	95	90	98	130	132	117	128	132	137
53	68	64	63	94	89	98	129	131-130	116	128	131	136
52	67-66	63	62	94	88	97	128	130	115	127	130	136
51	65	62	61	93	87	96	127	129	114	127	129	135
50	65	61-60	60	92-91	87	95	126	129	113	126-125	128	134
49	64	59	59	90	86	95	125-124	128	112-111	124	127	133
48	64	59	59	90	86	94	123	128	110	123	126	132
47	63	58-57	58-55	89	85-84	93	122	127	109	123	125	131
46	62	57	54	88	83	92-91	121-120	127	108	122	125	130-129
45	61	56	53-52	87	82	90	119	126	107	121	124	128
44	61	55	51	87	81	90	118-117	125	107	120	123	128
43	60	55	50	86	81	89	116	124	106	120	122	128

Percentil tabell 3 oppdatert versjon 2015

42	60	54	50	86	80	88	115	124	105	119	121	127
41	59	54	49	85	79	88	114	123	104	119	121	126-125
40	58	53-52	49	84	78	87	114	123	103	118	120	124-123
39	57	51	48	84	77	86	113	122	102	117	119	122
38	56	50	47	83	76	85	113	122	101	116	118	121
37	55	49	46	82	75	84	112	121	100	115	117	120-119
36	54	48	46	81-80	74	83	111	120	99-98	115	116	118
35	53	47	45	79	73	83	110	119	97	114-113	115	117
34	52	46	44	78	72	82	109	118	96	112	114	116
33	51	46	43	78	71	81	108	118	95	111	113	115
32	50	45	42	77	70	80	107-106	117	94-93	110	112	115
31	49	44	42	76	69	79	105	117	92	110	111	114
30	48	43	41	75	68	78	105	116	91	109	110	113
29	47	42	41	74	67	77	104	115	90-89	109	109-107	112
28	47	41	40	73	67	76	104	114-113	88-87	108	106-105	111
27	46	40	39	72	66	75-74	103-101	112	86	108	105	110
26	45	39	39	72	66	73	100-99	111	85	107	104	109
25	44	38	38	71	65	72	98	110	84	107	103	108
24	44	38	38	70-69	64	71	97	109	83-82	106-102	102	107
23	43	37	37	68	63	70	96	108	81	101	101	106
22	42	36	36	67	62	69	95	107-105	80	100	100	105-104
21	42	35	35	66	61	68	94	104-101	79	99	99	103
20	41	34	34	66	60	67-66	93-92	100-98	78	98	98	102
19	41	34	33	65-64	60	65-64	91-90	97-95	77	98	97	101
18	40	33	32	63	59-58	63-62	89	94-93	76	97	96	100
17	40	32-31	31	62	57-56	61-60	88-87	92	75-73	96	95	99
16	39	30	30	61	55	59	86	91	72	95	94-93	98
15	38	29	29	60	54	58-57	85-83	90	71	94-93	92-91	97
14	37	28	29	59	53	56-54	82	90	70	92	90-88	96
13	36-35	28	28	58	52	53	81-80	89-87	69	91	87	95
12	34	27-26	28	57	51-50	52-51	79-78	86-85	68	90	86-85	94
11	33	25	27	56-55	49	50-49	77-75	84-81	67	89	84-83	93-91

Percentil tabell 3 oppdatert versjon 2015

10	32	24	27	54-51	48-46	48-47	74	80	66-65	88-84	82-80	90-89
9	31	23-22	26-25	50	45-44	46-45	73-72	79-76	64-63	83-82	79-78	88
8	30-29	21-19	24-23	49-48	43	41-40	71-69	75-74	62-61	81-78	77-74	87
7	28	18	22-21	47	42-39	39-38	68	73-72	60	77-71	73-72	86-83
6	27-26	17	20	46	38-36	37-34	67-59	71-68	59-56	70-69	71-66	82-75
5	25-23	16-15	19	45-40	35-34	33-32	58-55	67-58	55-50	68-65	65-57	74-69
4	22-20	14-13	18-17	39-35	33-30	31-30	54-43	57-55	49-45	64-62	56-52	68-63
3	19-17	14-13	16	34-31	29-22	29-21	42-39	54-46	44-37	61-53	51-47	62-49
2	16-14	12-11	15-14	30-20	21-19	20-12	38-31	45-35	36-27	52-46	46-25	48-33
1	<=13	<=10	<=13	<=19	<=18	<=11	<=30	<=34	<=26	<=45	<=24	<=32

C

Norsk samfunnsvitenskapelig datatjeneste AS
NORWEGIAN SOCIAL SCIENCE DATA SERVICES

Harald Hårfagres gate 29
N-5007 Bergen
Norway
Tel: +47-55 58 21 17
Fax: +47-55 58 96 50
nsd@nsd.uib.no
www.nsd.uib.no
Org.nr: 985 321 884

Anne Arnesen
Institutt for spesialpedagogikk Universitetet i Oslo
Postboks 1140 Blindern
0318 OSLO

Vår dato: 21.11.2014 Vår ref: 40192 / 3 / JSL Deres dato: Deres ref:

TILBAKEMELDING PÅ MELDING OM BEHANDLING AV PERSONOPPLYSNINGER

Vi viser til melding om behandling av personopplysninger, mottatt 07.10.2014. Meldingen gjelder prosjektet:

40192	*Prososial atferd og leseutvikling*
Behandlingsansvarlig	*Universitetet i Oslo, ved institusjonens øverste leder*
Daglig ansvarlig	*Anne Arnesen*

Personvernombudet har vurdert prosjektet, og finner at behandlingen av personopplysninger vil være regulert av § 7-27 i personopplysningsforskriften. Personvernombudet tilrår at prosjektet gjennomføres.

Personvernombudets tilråding forutsetter at prosjektet gjennomføres i tråd med opplysningene gitt i meldeskjemaet, korrespondanse med ombudet, ombudets kommentarer samt personopplysningsloven og helseregisterloven med forskrifter. Behandlingen av personopplysninger kan settes i gang.

Det gjøres oppmerksom på at det skal gis ny melding dersom behandlingen endres i forhold til de opplysninger som ligger til grunn for personvernombudets vurdering. Endringsmeldinger gis via et eget skjema, http://www.nsd.uib.no/personvern/meldeplikt/skjema.html. Det skal også gis melding etter tre år dersom prosjektet fortsatt pågår. Meldinger skal skje skriftlig til ombudet.

Personvernombudet har lagt ut opplysninger om prosjektet i en offentlig database, http://pvo.nsd.no/prosjekt.

Personvernombudet vil ved prosjektets avslutning, 31.12.2016, rette en henvendelse angående status for behandlingen av personopplysninger.

Vennlig hilsen

Katrine Utaaker Segadal

Juni Skjold Lexau

Kontaktperson: Juni Skjold Lexau tlf: 55 58 36 01
Vedlegg: Prosjektvurdering

Dokumentet er elektronisk produsert og godkjent ved NSDs rutiner for elektronisk godkjenning.

Personvernombudet for forskning

Prosjektvurdering - Kommentar

Prosjektnr: 40192

Prosjektet gjennomføres i samarbeid med Atferdssenteret og UniRand. Institutt for spesialpedagogikk ved Universitetet i Oslo er behandlingsansvarlig institusjon. Personvernombudet forutsetter at ansvaret for behandlingen av personopplysninger er avklart mellom institusjonene. Vi anbefaler at det inngås en avtale som omfatter ansvarsfordeling, ansvarsstruktur, hvem som initierer prosjektet, bruk av data og eventuelt eierskap.

Formålet med prosjektet er å validere kartleggingsverktøy for lærervurdering av barnetrinnselevers sosiale atferd i læringssituasjonen som kan ha betydning for læring og utvikling.

Utvalget (foreldrene) informeres skriftlig og muntlig om prosjektet og samtykker til deltakelse. Revidert informasjonsskriv mottatt på e-post 14.11.2014 er godt utformet.

Vi legger til grunn at kontaktlærer som skal fylle ut skjemaene for enkeltelevene, samtykker til denne registreringsoppgaven.

Det behandles sensitive personopplysninger om etnisk bakgrunn eller politisk/filosofisk/religiøs oppfatning, helseforhold.

Det tas høyde for at det behandles enkelte opplysninger om tredjeperson (den andre forelderen som evt ikke er tilstede når det oppgis bakgrunnsinformasjon om foreldrene/familien). Det skal kun registreres opplysninger som er nødvendig for formålet med prosjektet. Opplysningene skal være av mindre omfang og ikke sensitive, og skal anonymiseres i publikasjon. Så fremt personvernulempen for tredjeperson reduseres på denne måten, kan prosjektleder unntas fra informasjonsplikten overfor tredjeperson i de tilfeller det er nødvendig, fordi det anses uforholdsmessig vanskelig å informere.

Personvernombudet legger til grunn at forsker etterfølger Universitetet i Oslo sine interne rutiner for datasikkerhet. Dersom personopplysninger skal sendes elektronisk, bør opplysningene krypteres tilstrekkelig. Vi legger til grunn at UiO godkjenner bruken av e-post med elevnavn til rektor og at den tekniske løsningen for innhenting av spørreskjemadata er god nok. Vi anbefaler at forsker rådfører seg med IT-avdelingen ved UiO før igangsetting og før spørreskjemaleverandør velges.

Det er ikke oppgitt hvorvidt det skal benyttes en databehandler til innhenting av spørreskjema. Hvis det benyttes databehandler, gjør vi oppmerksom på at Universitetet i Oslo skal inngå skriftlig avtale med databehandler om hvordan personopplysninger skal behandles, jf. personopplysningsloven § 15. For råd om hva databehandleravtalen bør inneholde, se Datatilsynets veileder: http://www.datatilsynet.no/Sikkerhet-internkontroll/Databehandleravtale/. Personvernombudet ber om kopi av avtalen for arkivering (sendes: personvernombudet@nsd.uib.no).

Forventet prosjektslutt er 31.12.2016. Ifølge prosjektmeldingen skal innsamlede opplysninger da anonymiseres.

Anonymisering innebærer å bearbeide datamaterialet slik at ingen enkeltpersoner kan gjenkjennes. Det gjøres ved å:

- slette direkte personopplysninger (som navn/koblingsnøkkel)
- slette/omskrive indirekte personopplysninger (identifiserende sammenstilling av bakgrunnsopplysninger som f.eks. bosted/arbeidssted, alder og kjønn)

Vi gjør oppmerksom på at også databehandler må slette personopplysninger tilknyttet prosjektet i sine systemer. Dette inkluderer eventuelle logger og koblinger mellom IP-/epostadresser og besvarelser.

D

Norsk samfunnsvitenskapelig datatjeneste AS
NORWEGIAN SOCIAL SCIENCE DATA SERVICES

Harald Hårfagres gate 29
N-5007 Bergen
Norway
Tel: +47-55 58 21 17
Fax: +47-55 58 96 50
nsd@nsd.uib.no
www.nsd.uib.no
Org.nr. 985 321 884

Terje Ogden
Atferdssenteret - Norsk senter for studier av problematferd og innovativ praksis AS
Postboks 7053 Majorstuen
0306 OSLO

Vår dato: 14.09.2012 Vår ref:31412 / 3 / MAS Deres dato: Deres ref:

TILBAKEMELDING PÅ MELDING OM BEHANDLING AV PERSONOPPLYSNINGER

Vi viser til melding om behandling av personopplysninger, mottatt 10.09.2012. Meldingen gjelder prosjektet:

31412	*Læringsmiljø, prososial atferd og lesing*
Behandlingsansvarlig	*Unirand AS, ved institusjonens øverste leder*
Daglig ansvarlig	*Terje Ogden*

Personvernombudet har vurdert prosjektet, og finner at behandlingen av personopplysninger vil være regulert av § 7-27 i personopplysningsforskriften. Personvernombudet tilrår at prosjektet gjennomføres.

Personvernombudets tilråding forutsetter at prosjektet gjennomføres i tråd med opplysningene gitt i meldeskjemaet, korrespondanse med ombudet, eventuelle kommentarer samt personopplysningsloven og helseregisterloven med forskrifter. Behandlingen av personopplysninger kan settes i gang.

Det gjøres oppmerksom på at det skal gis ny melding dersom behandlingen endres i forhold til de opplysninger som ligger til grunn for personvernombudets vurdering. Endringsmeldinger gis via et eget skjema, http://www.nsd.uib.no/personvern/forsk_stud/skjema.html. Det skal også gis melding etter tre år dersom prosjektet fortsatt pågår. Meldinger skal skje skriftlig til ombudet.

Personvernombudet har lagt ut opplysninger om prosjektet i en offentlig database, http://pvo.nsd.no/prosjekt.

Personvernombudet vil ved prosjektets avslutning, 31.12.2014, rette en henvendelse angående status for behandlingen av personopplysninger.

Vennlig hilsen

Vigdis Namtvedt Kvalheim

Mads Solberg

Kontaktperson: Mads Solberg tlf: 55 58 89 28
Vedlegg: Prosjektvurdering

Avdelingskontorer / District Offices
OSLO: NSD, Universitetet i Oslo, Postboks 1055 Blindern, 0316 Oslo Tel: +47-22 85 52 11 nsd@uio.no
TRONDHEIM: NSD, Norges teknisk-naturvitenskapelige universitet, 7491 Trondheim Tel: +47-73 59 19 07 kyrre.svarva@svt.ntnu.no
TROMSØ: NSD, SVF, Universitetet i Tromsø, 9037 Tromsø Tel: +47 77 64 43 36 nsdmaa@svfat.no

Norsk samfunnsvitenskapelig datatjeneste AS
NORWEGIAN SOCIAL SCIENCE DATA SERVICES

Harald Hårfagres gate 29
N-5007 Bergen
Norway
Tel: +47-55 58 21 17
Fax: +47-55 58 96 50
nsd@nsd.uib.no
www.nsd.uib.no
Org.nr. 985 321 884

Terje Ogden
Atferdssenteret - Norsk senter for studier av problematferd og innovativ praksis AS
Postboks 7053 Majorstuen
0306 OSLO

Vår dato: 14.05.2013 Vår ref: 31412 JSL/RF Deres dato: Deres ref:

ENDRINGSMELDING

Vi viser til endringsmelding mottatt15.04.2013 for prosjektet:

31412 *Læringsmiljø, prososial atferd og lesing*

Vi har registrert at det søkes om at elevenes nasjonale prøver i leseferdigheter skal utleveres, slik at disse kan sammenholdes med tilsvarende data som allerede er innhentet fra elevene i prosjektet.

Det skal gis ny informasjon og innhentes nye samtykker fra foreldre til dette. Vi finner revidert informasjonsskriv og samtykkeerklæring tilfredsstillende, slik det foreligger i e-post mottatt 13.05.2013.

Ta gjerne kontakt dersom noe er uklart.

Vennlig hilsen

Vigdis Namtvedt Kvalheim

Juni Skjold Lexau

Kontaktperson: Juni Skjold Lexau tlf: 55 58 36 01

Avdelingskontorer / District Offices:
OSLO: NSD, Universitetet i Oslo, Postboks 1055 Blindern, 0316 Oslo. Tel: +47-22 85 52 11. nsd@uio.no
TRONDHEIM: NSD, Norges teknisk-naturvitenskapelige universitet, 7491 Trondheim. Tel: +47-73 59 19 07. kyrre.svarva@svt.ntnu.no
TROMSØ: NSD, HSL, Universitetet i Tromsø, 9037 Tromsø. Tel: +47-77 64 43 36. martin-arne.andersen@uit.no

E

Kartleggingsverktøy for elevenes leseutvikling og sosiale atferd _Versjon2 kopi

Side 1

Institutt for spesialpedagogikk ved Universitetet i Oslo gjennomfører en studie om tilgjengelige kartleggingsverktøy i skolen for vurdering av elevenes leseutvikling og sosiale atferd, og skolens bruk av dem.

Studien vil fremskaffe en oversikt over relevante verktøy, samt kvaliteteten og anvendbarheten av dem i skolens arbeid. Din skoles svar på de følgende spørsmålene vil være et viktig bidrag i studien. Vi håper derfor at du/dere kan avsette ca. 5 - 8 minutter til å besvare dem.

På forhånd takk for hjelpen!

Vennlig hilsen,

Anne Arnesen og Monica Melby-Lervåg
stipendiat professor

Skole: *

Skole:

Velg ...

Spørreskjemaet er besvart av *

- O Skolens rektor/ledelse
- O Skolens ledelse/rektor i samarbeid med lærerne
- O Skolens lærere

Skolens trinninndeling *

- ☐ 1-4
- ☐ 5-7
- ☐ 8-10
- ☐ 1-10

Andre trinninndelinger, vennligst spesifiser:

Side 2

Sideskift

Kartlegging av elevenes leseferdigheter

Elevenes leseferdigheter (ordavkoding, nøyaktighet, leseflyt/hastighet og leseforståelse) blir vurdert ved bruk av egne kartleggingsverktøy/-prøver. *

O Ja O Nei

Dette elementet vises dersom et av følgende alternativer er valgt på spørsmål «Elevenes leseferdigheter (ordavkoding, nøyaktighet, leseflyt/hastighet og leseforståelse) blir vurdert ved bruk av egne kartleggingsverktøy/-prøver.»: Ja

Dersom "Ja", vennligst angi alle de ulike kartleggingsverktøy/-prøver som skolen anvender på de ulike klassetrinn (f.eks. Carlsten lesestest 2.-5.trinn; STAS 2.-7.trinn): *

Skolen bruker egne utviklede kartleggingsverktøy eller vurderingsmåter for elevenes leseutvikling. Vennligst beskriv disse kort:

Elevenes leseutvikling kartlegges og vurderes: *

- O Oftere enn 3 ganger per skoleår
- O 3 ganger per skoleår
- O 2 ganger per skoleår
- O 1 gang per skoleår
- O Sjeldnere enn 1 gang per skoleår

Kartleggingsresultater og/eller vurderingsinformasjon fører til konkret handling i arbeidet for den enkelte elev og oppfølging av elevens leseutvikling. *

- O Ja
- O Nei
- O Vet ikke

Side 3

Sideskift

Kartlegging av elevenes sosiale atferd

Elevenes sosiale atferd blir vurdert ved bruk av egne kartleggingsverktøy. *

O Ja O Nei

Dette elementet vises dersom et av følgende alternativer er valgt på spørsmål «Elevenes sosiale atferd blir vurdert ved bruk av egne kartleggingsverktøy.»: Ja

Dersom "Ja", vennligst angi alle kartleggingsverktøy som skolen anvender for å vurdere elevenes sosiale atferd på de ulike klassetrinn (f.eks. «7 spørsmål om relasjoner til andre» - 1.-6. trinn): *

Skolen bruker egne utviklede kartleggingsverktøy eller vurderingsmåter for elevenes sosiale atferd i skolesituasjonen. Vennligst beskriv disse kort:

Utvikling av elevenes sosiale atferd blir kartlagt og vurdert. *

- O Oftere enn 3 ganger per skoleår
- O 3 ganger per skoleår
- O 2 ganger per skoleår
- O 1 gang per skoleår
- O Sjeldnere enn 1 gang per skoleår

Kartleggingsresultater og/eller vurderingsinformasjon fører til konkret handling i arbeidet for den enkelte elev og oppfølging av elevens sosiale atferd. *

- O Ja
- O Nei
- O Vet ikke

Takk for hjelpen!

Nettskjema v100.6

F

Assessment of Students' Social Functioning and Reading Proficiency:

A systematic review and psychometric evaluation of assessment instruments used in Norwegian Elementary School

Protocol systematic literature review

1. The review questions

The overall question to be addressed in the review is:

What are the empirically qualities (evidence) of assessment instruments for students' social behavior and reading proficiency used in Norwegian elementary schools?

In order to achieve all the aims of the review the further questions to be addressed are:

- *What are the instruments' properties in terms of predicting social skills difficulties and reading difficulties in early elementary school age/grades?*
- *If any, how are the instruments designed to identify and progress monitor students at risk for difficulties in social skills and reading in early elementary school age/grades?*
- *What are the psychometric properties in terms of validity and reliability?*
- *How do the instrument properties vary with*
    - *characteristics of assessments (screening, tests, etc)*
    - *characteristics of students*
    - *characteristics of assessors or raters*
- *What can be learned with regard to feasibility in terms of time consuming, administering, scoring, cost efficiency, etc.?*
- *How is performance assessment used for assessment for learning?*
- *What are the implications for assessment policy and practice of these findings?*

2. Definitions and conceptual issues

Social function refers to how students behave in a school setting "by relying on social skills and interacting with others." (Beauchamp & Anderson, 2010, p.40). Social skills refer to cognitive and interpersonal abilities that are required for appropriate specific behaviors associated with social relations, emotional and academic engagement, and motivation to be successful in school (Beauchamp & Anderson, 2010; Cordier et al., 2015; Gresham, 2007).

Reading proficiency refers to students' skills in reading based on achievement scores and descriptions of what students are expected to know and do in reading at each grade level (Child Trends Databank, 2015). In other words, reading proficiency is commonly defined as an active and complex process that involves interpreting, understanding and using the meaning of varied texts (Gough & Tunmer, 1986; Hoover & Gough, 1990; The National Assessment of Educational Progress, 2013).

Assessment instruments are authentic instruments for early screening and progress monitoring children's social functioning and reading proficiency that are relevant for learning and development of social and reading skills in elementary school.

Psychometric properties are the construct and the validity of assessment instruments to be used in elementary schools: evaluation of the quality in terms of reliabelity, validity, norms.

3. Search strategy

3.1. Languages: English, Norwegian, Swedish, Danish, Finish

3.2. Time frame: No limits (?)

3.3. Sources

3.3.1. Electronic databases

- ERIC
- PsycInfo

- Norart
- Oria

3.3.2. Google and Google Scholar

3.3.3. Hand searching in manuals of measurements listed by schools participating in the survey that is the base of the review (see lists in 3.5 below)

3.3.4. Hand searching in journals:

- Assessment in Education: Principles, policy & practice
- Educational Measurement: Issues and Practice
- Studies in Educational Evaluation
- Spesialpedagogikk
- International Journal of Selection and Assessment

3.3.5. Handsearching databases existing reviews:

- Campbell: http://www.campbellcollaboration.org/
- Cochrane: http://www.nelh.nhs.uk/cochrane.asp
- Institute of Education Sciences: http://ies.ed.gov/ncee/wwc/
- Eppi Center: http://eppi.ioe.ac.uk/

3.3.6. Handsearching databases Conference Proceedings:

- Index of Conference Proceedings
- Index to Social Sciences and Humanities Proceedings
- http://www.tandfonline.com (Educational Research Abstracts)

3.3.7. Hand searching reference lists of key authors of manuals and papers

3.3.8. Citation searches of key authors

3.3.9. Hand searching Grey Literature:

- Web of Science
- NORA

- BASE (Bielefeld Academic Search Engine)
- Pro Quest Dissertations and Theses
- OpenGrey.eu

3.4. Inclusion and exclusion criteria

Searching and selection of studies is guided by the following inclusion criteria:

- Assessment instruments designed to predict difficulties in social functioning and reading
- Assessment instruments designed to identify children's difficulties in social functioning and reading
- Measurements designed to be used in Norwegian elementary school (Gr. 1-6)
- Assessment instruments designed for age 6-12
- Assessment instruments designed to progress monitor students development of social functioning and reading (trajectories/change/response to intervention)
- Measurements with described psychometric properties per ce
- Assessment instruments with defined norms/ cut-off scores classified (specificity/sensitivity)
- Assessment instruments with documented evidence

Language of the report: Studies /manuals included are written in English, Norwegian, Danish, Swedish (or Finnish?)

Types of assessment instruments: Studies are included which deal with performance instruments as defined above. In addition studies that report on instruments that are designed to predict difficulties in social functioning and reading, and to identify students at risk for these difficulties.

Study population and setting: Studies are included that report on performance assessment instruments in elementary/primary school. In addition studies that report on performance assessment in learning & development or early school age are included as well.

Study type and study design: Studies are included if they report quantitative or qualitative evidence of the psychometric properties of performance assessment in terms of validity and/or reliability.

Studies meeting some of the above inclusion criteria will be excluded for the following reasons and labelled accordingly:

- Studies which content one or more of the inclusion criteria, but are irrelevant as regards assessments of social functioning and reading as defined
- Studies that focus intervention/training programs without inclusion of any descriptions of measurement instruments

Furthermore searching and selection of studies is guided by the following exclusion criteria:

- Not performance assessment (excluded if the performance assessment is too 'shallow')
- Not related to assessment in elementary/primary school contexts for learning and development settings.
- Not reporting validity or reliability of performance assessment, but just *describing* the performance assessment.
- No research (excluded if not empirical study; also excluded if not satisfied pre-determined methodological criteria, namely: the assessment approaches were described in adequate detail, any instruments used were accessible and there was information about the process of administering data collection procedures; the raw data were illustrated adequately to create a sense of the actual evidence used to

support claims; the structure, content and timeline of any interventions were described in adequate detail).

- Assessment instruments focusing only on disabled students or diagnostic
- Assessment instruments which is not designed for use by teachers/schools

3.5 Lists of measurements used in Norwegian Elementary schools

3.5.1 Reading

Measurements used in Norwegian Elementary Schools:	
1.	AMO (Automatisert Mest høyfrekvente ord)
2.	arbeidmedord.no (kartleggingsmateriell: «Leseklar» og kartlegging av leseferdighet)
3.	Arbeidsprøven (Dynamisk kartlegging – Bredtvet kompetansesenter)
4.	Aski Raski (treningsprogram + kartlegging) utviklet av Ingrid Ask
5.	Aston Index (screening og diagnostisk/observasjon og vurdering av lese/skrive/språkvansker)> Gallefoss, 1996
6.	Bokstavtesten (Lesepedagogen; Greta Storm Ofteland)
7.	Carlsten leseprøve/test
8.	DAMMS leseunivers kartleggingsskjema
9.	God leseutvikling – kartleggingsmateriell Cappelen Akademiske; Herrlin, Lundberg, 2008)
10.	Haugland-Gilje: Leseprøver
11.	HOA lesetest (H. O. Aker, 1994: Lærerhjelp, sandefjors))
12.	IL-basis (Frost et al..)
13.	Kartleggeren.no
14.	KOAS (Kartlegging av ordavkodingsstrategier)
15.	KTI kontrollert tegneiakttagelse (Vurdere førsteklassingers forutsetninger for å motta og bearbeide språklige budskap) Dansk opplegg Jensen og Krogh
16.	Kåre Johnsen diagnostisk lese- og skriveprøve
17.	Leselos (Lesesenteret, Universitetet i Stavanger)
18.	Lesesenterets staveprøve (Astrid Skaathun)
19.	LOGOS (håndbok) (Logometrica)
20.	LUS (LeseUtviklingsSkjema)
21.	Nasjonale kartleggingsprøver leseferdighet (Utd.direktoratet)
22.	NSL (Norsk Som Læringsspråk; Heller, 2014)
23.	Ordkjedetest (Logometrica/Høien & Tønnesen)
24.	OL64; OL120; MiniSL1; MiniSL2
25.	OS400
26.	Osloprøven i lesing (6.trinn) basert på formatet som nasjonale prøver - lesetekster
27.	PDK-testen (nedlagt 31.12.15) Lesesenteret - vox
28.	Ringeriksmaterialet (kartlegging barns språklige bevissthet i alderen 5-7 år; Lyster, 2002)
29.	S40/SL40 /SL 60 setningsleseprøve (Lesesenteret, Universitetet i Stavanger)
30.	SOL (Systematisk Observasjon av Lesing)
31.	Språk 6-16 (Ottem & Frost; Bredtvet kompetansesenter)
32.	STAS (Standardisert Test i Avkoding og Staving): Klinkenberg & Skaar, 2001; Ringerike PPT
33.	Tempolex
34.	20 Spørsmål (Ottem)

Key words Reading				
read*	measur*	psychometr*	"elementary school"	"at risk"
Oral reading (psycINFO)	evaluat*	valid*/ test validity (PsycInfo)	"early school age"	At risk student*
	assess*	reliab*	"primary school"	Risk*
	screen* / (PsycINFO: screening tests/or screening)	analy*		Risk factor* (PsycINFO)
	exam*	norm*		
	map*	"cut* score*"		
	scal*	propert*		
	test*	"test length"		
	predict* ("predictive measurement")	standard*		
	"progress monitor*"			
	identif*			
	rat* (rating; rate)			
	"formative evaluation"			

3.5.2 Social Functioning

Kartleggingsverktøy – tester for vurdering av barns sosiale fungering som et utvalg av norske barneskoler (n=234) oppgir at de bruker
Verktøy/Test:
1. Snakk med meg! SMM-metoden (mobbing-psykososial): Veiledningshefte om å snakke med barn i barnevernet: Barne- og likestillingsdepartementet 2009
2. trivselsleder.no
3. 5-15 Nordisk skjema for utredning av barns utvikling og atferd (Kadesjö et al.)
4. ADDES (Attention Deficit Disorder Evaluation Scale): McCarney & Arthaud
5. Eyberg Child Behavior Inventory – Kartlegging av atferdsvansker ECBI ble oversatt til norsk i 1999 ved Willy-Tore Mørch, Universitetet i Tromsø – Et foreldreskjema?? I såfall ekskluder
6. Elevenes selvrapport fra PPT: Systematisk kartlegging av elevens subjektive forståelse av egen livssituasjon (Duna, K. E. & Frost, J. (1999) Elevens selvrapport. Jaren: PP-tjenestens materiellservice.)
7. Ingen utenfor (5-7): Redd Barna?
8. Innblikk 5-7: et sosial-analytisk verktøy for forebygging og avdekking av skjult mobbing. Utviklet av Tove Flack, 2010
9. Kartlegging av klassens sosiale miljø: Utdanningsdirektoratet
10. Klassetrivsel.no - sosiogram
11. Mobbeundersøkelsen: Læringsmiljøsenteret Stavanger – del av Zeroprogrammet
12. Olweus spørreundersøkelse – Elevundersøkelsen 5. - 7 trinn - Utdanningsdirektoratet
13. Psykologisk førstehjelp; Solfrid Raknes
14. Respekt Elevundersøkelsen: Læringsmiljøsenteret Stavanger ??
15. Sosiogram (dansk opplegg)
16. TRF (Teacher Report Form): Liste over barns atferd i alderen 6-18 år - Lærerskjema Amerikanske normer 6-18 år. Den norske normeringen baserer seg på lærere i 1.-7. klasse og 858 barn
17. Trivselsundersøkelser - selvlagde
18. Zippys venner (1-2trinn): intervensjonsprogram. Voksne for Barn - effektevalueringsstudie Holen – ekskluder dersom ikke inneholder kartleggingselement

Key words Social Functioning				
Social*	measur*	psychometr*	"elementary school"	"at risk"
Social skill*	evaluat*	valid*	"early school age"	At risk student*
Social performance*	assess*	reliab*	"primary school"	Risk*
Social proficienc*	screen*	analys*		Risk factor* (PsycINFO)
Social difficult*	exam*	norm* (normert)		

Social develop*	map*	propert* (properties)		
Social competenc*	scale*	standard* (standardized)		
Social problem*	test*	construct*		
Social Behav*	predict*			
	"progress monitor*"			
	identif*			
	rat* (rating; rate)			
	"respons to intervention"			

G

Sosiale ferdigheter 1.-6.trinn - Lærerversjon_Wave1-2 kopi

Side 1

Lærerens vurderingsskjema utfylles to ganger med åtte ukers mellomrom for hver elev. Vennligst marker om dette er første eller andre gangs utfylling for eleven du nå skal vurdere ved å velge et av alternativene under:

- O 1.gang
- O 2.gang

Skolens Id-nr (2 sifre - sendt deg per e-post): *

Det fylles ut et nytt skjema for hver elev. Du får tilgang til nye skjema ved å klikke om igjen i lenken som ga deg tilgang til dette skjemaet. (Gjentas for hver elev du skal vurdere.)

Vurderingsskjemaet har tre sider:

A) Opplysninger om eleven og læreren

B) Vurderingsskjema *Sosiale ferdigheter - "Elementary Social Behavioral Assessment": ESBA*

C) Vurderingsskjema *Sosiale ferdigheter - "Social Skills Rating System": SSRS*

A.Opplysninger om eleven og læreren

Elevens Id-nr (4 sifre - sendt deg per e-post): *

Klassetrinn: *

O 1.trinn O 2.trinn O 3.trinn O 4.trinn O 5.trinn O 6.trinn

Elevens fødselsdato (oppgis med 6 sifre for ddmmåå): *

Kjønn: *

O Jente O Gutt

Eleven har enkeltvedtak om spesialundervisning *

O Ja O Nei

Eleven har individuell opplæringsplan (IOP) *

O Ja O Nei

Lærerens Id-nr (3 sifre - sendt deg per e-post): *

Kjønn: *

O Kvinne O Mann

Hva er din rolle i forhold til eleven: *

O Kontaktlærer O Time/faglærer O Spesialpedagog O Skoleleder O Annet

Dette elementet vises dersom et av følgende alternativer er valgt på spørsmål «Hva er din rolle i forhold til eleven:»: Annet.

Dersom annet, hva:

Side 2

Sideskift

B. Vurderingsskjema Sosiale ferdigheter - "Elementary Social Behavioral Assessment": ESBA

Norsk tilpasning fra Marquez, Marquez, Vincent, Pennefather, Sprague, Smolkowksi & Yeaton

Dette skjemaet består av 12 utsagn og er laget for å vurdere elevenes prososiale atferd som kan ha betydning i læringssituasjonen i 1.-6.trinn.

Vennligst vurder hvert utsagn nøye og tenk på elevens atferd "nå om dagen". Bestem deg for **hvor ofte** du mener utsagnene stemmer utfra ditt kjennskap til eleven ved å angi med "knappen" til høyre for utsagnet.

- Hvis eleven nesten alltid gjør som beskrevet, markeres ***"Nesten alltid"***
- Hvis eleven bare av og til gjør som beskrevet, markeres ***"Av og til"***
- Hvis du sjelden eller ikke har sett at eleven gjør som beskrevet, markeres ***"Sjelden/aldri"***

Det er ingen rette eller gale svar. Dersom du er usikker, bruk ditt beste skjønn.

	Nesten alltid	Av og til	Sjelden/aldri
1. Hører godt etter når læreren snakker og gir beskjeder (eks.: vender seg mot deg når du snakker; ser på deg og hører etter; er oppmerksom og får med seg det du sier, etc.) *	O	O	O
2. Følger lærerens beskjed (eks.: tar frem nødvendig materiale; går fort i gang med arbeidsoppgavene; gjør det han/hun er bedt om uten å somle, etc.) *	O	O	O
3. Viser god arbeidsinnsats (eks.: gjør sitt beste; er engasjert; holder seg til oppgaven; gjør en ting om gangen, etc.) *	O	O	O
4. Sitter ved plassen sin og jobber når det er forventet (eks.: fullfører oppgavene; arbeider konsentrert, etc.) *	O	O	O
5. Ber om hjelp på en hensiktsmessig måte (eks. rekker opp hånden eller viser annet tegn til å trenge hjelp; oppsøker deg; venter på tur, etc.) *	O	O	O
6. Oppfører seg som forventet i klasse- og undervisningsrommet *	O	O	O
7. Følger regler selv om jevnaldrende oppmuntrer til å bryte dem *	O	O	O
8. Oppfører seg som forventet utenfor klasse- og undervisningsrommet *	O	O	O
9. Håndterer/regulerer sine følelser på en hensiktsmessig måte (inkluderer både inagerende og utagerende atferd) *	O	O	O
10. Kommuniserer greit med jevnaldrende uten å provosere eller være negativt innstilt (eks. tåler korreksjon; reagerer hensiktsmessig; forholder seg rolig, etc.) *	O	O	O
11. Går godt overens med jevnaldrende (eks. er vennlig; roser og anerkjenner andre; tar andre med på lek/aktiviteter; etc.) *	O	O	O
12. Løser konflikter med jevnaldrende uten hjelp fra voksne (eks; snakker rolig; godtar et unnskyld fra andre; beklager når han/hun gjør feil; inngår kompromisser, etc.) *	O	O	O

Side 3

Sideskift

C. Vurderingsskjema Sosiale ferdigheter - "Social Skills Rating System": SSRS

Norsk tilpasning fra Gresham og Elliot

Lærerutgave 1.-7.trinn

Dette skjemaet har 57 utsagn og er laget for å vurdere elevenes sosiale ferdigheter i skolesituasjonen. Vurdering av problematferd og funksjonsnivå i skolen inngår også.

Les hvert av utsagnene nedenfor (1 –57) og tenk på elevens atferd nå om dagen. Bestem deg for **hvor ofte** du mener eleven gjør det som står beskrevet.

- Hvis eleven aldri gjør som beskrevet, markeres **"Aldri"**
- Hvis eleven av og til gjør som beskrevet, markeres **"Av og til"**
- Hvis eleven ofte gjør som beskrevet, markeres **"Ofte"**
- Hvis eleven svært ofte gjør som beskrevet, markeres **"Svært ofte"**

Det er ingen rette eller gale svar. Dersom du er usikker, bruk ditt beste skjønn.

SOSIALE FERDIGHETER

30 utsagn (1-30)

	Aldri	Av og til	Ofte	Svært ofte
1. Styrer sinnet sitt i konflikter med andre barn *	O	O	O	O
2. Presenterer seg uoppfordret når han/hun møter nye mennesker *	O	O	O	O
3. Protesterer mot regler som kan virke urimelige *	O	O	O	O
4. Endrer egne meninger i konflikter for å oppnå enighet *	O	O	O	O
5. Reagerer hensiktsmessig på gruppepress fra jevnaldrende *	O	O	O	O
6. Omtaler seg selv i positive vendinger når det er forventet *	O	O	O	O
7. Inviterer andre til å bli med på aktiviteter *	O	O	O	O
8. Bruker ledig tid (fritid) på en konstruktiv/positiv måte *	O	O	O	O
9. Fullfører arbeidsoppgaver i timene i tide *	O	O	O	O
10. Får lett venner *	O	O	O	O

	Aldri	Av og til	Ofte	Svært ofte
	Aldri	Av og til	Ofte	Svært ofte
11. Reagerer hensiktsmessig på erting fra jevnaldrende *	O	O	O	O
12. Kontrollerer sinnet sitt i konflikter med voksne *	O	O	O	O
13. Tåler å få kritikk *	O	O	O	O
14. Tar initiativ til samtaler med jevnaldrende *	O	O	O	O
15. Bruker tiden konstruktivt mens han/hun venter på hjelp *	O	O	O	O
16. Gjør skolearbeidet riktig/korrekt *	O	O	O	O
17. Sier i fra når han/hun mener at du har behandlet ham/henne urettferdig *	O	O	O	O
18. Godtar medelevers/kameraters forslag til aktiviteter *	O	O	O	O
19. Gir ros (komplimenter) til jevnaldrende *	O	O	O	O
20. Følger dine beskjeder (instrukser) *	O	O	O	O
	Aldri	Av og til	Ofte	Svært ofte
21. Rydder arbeidsmateriell og utstyr etter seg *	O	O	O	O
22. Samarbeider uoppfordret med andre elever *	O	O	O	O
23. Tilbyr seg å hjelpe medelever *	O	O	O	O
24. Blir uoppfordret med i aktiviteter eller i elevgruppa *	O	O	O	O
25. Reagerer hensiktsmessig hvis han/hun blir dyttet eller slått av andre elever *	O	O	O	O
26. Lar seg ikke distrahere av andre elever når han/hun arbeider i timene *	O	O	O	O
27. Holder uoppfordret puten/arbeidsplassen ryddig *	O	O	O	O
28. Følger med når du snakker til klassen *	O	O	O	O
29. Skifter lett fra en aktivitet til en annen i klassen/timene *	O	O	O	O
30. Kommer godt overens med personer som er annerledes *	O	O	O	O

PROBLEMATFERD

18 utsagn (31 - 48)

	Aldri	Av og til	Ofte	Svært ofte
31. Slåss med andre *	O	O	O	O
32. Mangler selvtillit *	O	O	O	O
33. Truer eller plager andre *	O	O	O	O
34. Virker ensom *	O	O	O	O
35. Blir lett distrahert *	O	O	O	O
36. Bryter inn i andres samtaler *	O	O	O	O
37. Forstyrrer pågående aktiviteter *	O	O	O	O
38. Er engstlig for å være i gruppe med andre elever *	O	O	O	O
39. Blir lett forlegen/flau *	O	O	O	O
40. Lytter ikke til hva andre sier *	O	O	O	O
	Aldri	Av og til	Ofte	Svært ofte
41. Krangler med andre *	O	O	O	O
42. Svarer tilbake når voksne irettesetter ham/henne *	O	O	O	O
43. Blir lett sint *	O	O	O	O
44. Har sinneutbrudd *	O	O	O	O
45. Liker å være for seg selv *	O	O	O	O
46. Virker trist eller deprimert *	O	O	O	O
47. Handler impulsivt *	O	O	O	O
48. Rastløs og urolig, i stadig bevegelse *	O	O	O	O

FUNKSJONSNIVÅ I SKOLEN

9 utsagn (49-57)

Utsagnene vurderes i forhold til elevens medelever i klassen.

	Laveste 10%	Nest laveste 20%	Middels 40%	Nest høyeste 20%	Høyeste 10%
49. Hvordan er denne elevens skolefaglige prestasjoner sammenlignet med de andre elevene i klassen? *	O	O	O	O	O
50. I lesning, hvordan er denne eleven sammenlignet med de andre elevene? *	O	O	O	O	O
51. I matematikk, hvordan er denne eleven sammenlignet med de andre elevene? *	O	O	O	O	O
52. I forhold til forventningene på klassetrinnet er denne elevens ferdigheter i lesning: *	O	O	O	O	O
53. I forhold til forventningene på klassetrinnet er denne elevens ferdigheter i matematikk: *	O	O	O	O	O
54. Denne elevens motivasjon for å lykkes på skolen er: *	O	O	O	O	O
55. Foreldrenes støtte og oppmuntring til eleven for a lykkes i skolen er: *	O	O	O	O	O
56. Sammenlignet med de andre elevene i klassen så er denne elevens intellektuelle funksjonsnivå: *	O	O	O	O	O
57. Sammenlignet med de andre elevene i klassen så er denne elevens klasseromsatferd: *	O	O	O	O	O

Gjenta elevens Id-nr (4 sifre tilsendt per e-post): *

Du er nå ved slutten av vurderingsskjemaet for denne eleven. Vi takker for at du har tatt deg tid til å gjennomføre det. For å åpne et nytt skjema for den neste eleven du skal vurdere, vennligst klikk på lenken du fikk tilsendt.

H

Norwegian adapted version of Oral Reading Fluency (ORF) Grades 2 to 5

Sett deg godt inn i skåringsreglene. De er de samme for alle tre kartleggingstidspunktene

Gi eleven følgende instruksjon:

Les denne historien høyt for meg!

Hvis du stopper opp fordi du ikke kan et ord, kommer jeg til å si ordet høyt slik at du kan fortsette lesingen. Les til jeg sier "stopp"!

Det kan hende jeg ber deg fortelle om hva du har lest!

Gjør ditt beste!

Legg elevarket foran eleven!

Denne historien heter (eller handler om) (si tittelen/overskriften på teksten høyt samtidig som du peker på den).

Pek på det første ordet i teksten. ***Start her! Begynn!***

Tid	1 minutt. Start stoppeklokken din når eleven leser det første ordet i teksten. Sett en strek ( ]) og si stopp etter 1. minutt.
Vent	Vent 3 sekunder på elevens respons. Hvis eleven nøler eller strever i mer enn 3 sekunder, marker ordet med en strek (/) og les ordet høyt for eleven. Hvis det er nødvendig, pek på neste ord i teksten for å indikere at eleven skal lese videre.
Avbrytelse	Avbryt kartleggingen hvis eleven ikke leser noen av ordene i den første linjen i den første kartleggingsteksten. Noter en skåre 0 på totalt antall ord i skåringsboksen. Hvis eleven leser færre enn 10 ord korrekt på den første kartleggings-teksten, ikke administrer kartleggingstekst 2 og 3. Ikke be eleven gjenfortelle teksten. Hvis eleven leser færre enn 40 ord korrekt på noen av kartlegging-tekstene, vurder nøye om du skal be eleven gjenfortelle teksten.
Påminnelse	Hvis eleven stopper å lese (og det ikke er knyttet til nøling eller streving) eller begynner å snakke om noe annet si: ***Fortsett.*** Gjenta dette så mange ganger som nødvendig. Hvis eleven mister oversikten, pek. Gjenta dette så mange ganger som nødvendig.

(Retell and qualitative assessment of reading behavior is not included in the following)

Versjon 4 –2014
Utarbeidet av Atferdssenteret – Norsk senter for studier av problematferd og innovativ praksis

Basert på DIBELS™ Data System med tillatelse fra Center on Teaching and Learning – University of Oregon

Grade 2 – Fall 1.1

Kjære dagbok

0	I går var jeg hos tannlegen. Jeg er litt redd når jeg skal til	14
14	tannlegen og håper at det ikke er vondt. Jeg hadde ikke noen	26
26	hull. Tannlegen sa at jeg var flink til å pusse tennene. Det er	39
39	fint. Nå skal jeg ikke gå til tannlegen på et helt år.	51
51	Da jeg kom hjem fra skolen, gikk jeg en liten tur med hunden	64
64	vår. Den heter Stella. Den er liten og søt.	73
73	Så gjorde jeg leksene mine. Jeg liker best å lese og tegne. Vi	86
86	hadde om troll. Vi skulle tegne et troll. Jeg tegnet et med to	99
99	hoder.	100
100	I dag skal jeg være sammen med Line og Eli. Det er snart jul,	114
114	og vi skal bake kaker.	119
119	Jeg gleder meg til jul. Da skal mamma og jeg dra til mormor	132
132	og morfar. Vi må ta fly.	138
138	Pappa bor ikke sammen med oss. Han bor i en annen by. Jeg	151
151	liker å være sammen med pappa og savner han mye. Jeg har	163
163	fått en ny bror. Han er liten og heter Emil. Jeg leker med	176
176	Emil, og da ler han.	181
181	Nå kan jeg ikke skrive mer i dag. Jeg kan skrive mer i	194
194	morgen og fortelle om da vi bakte kaker.	202

Totalt antall ord: ____
Antall ord feil (inkl. utelatte ord): ____
Antall ord korrekt: ____

Grade 2 – Fall 1.2

Syk

0	Mor, jeg er så syk. Kan jeg få noe å drikke? Kan jeg få litt brus?	16
16	Når du er syk, må du være inne mens de andre er ute og leker. Det	32
32	er fint å slippe skolen, men etter noen dager savner du skolen og	45
45	vennene dine.	47
47	Når du er syk, har du vondt i hodet og får lyst til å kaste opp. Da	64
64	er det best å ligge under dyna og ha det helt mørkt i rommet. Litt	79
79	feber er ikke farlig. Da blir du bare litt slapp.	89
89	Alle i familien synes synd på deg. Mor eller far vil gjerne gjøre det	103
103	beste for deg. Er det noe du vil ha, spør de. Da kan du ønske deg	119
119	mange ting. Med svak stemme sier du at du vil ha brus, is og et	134
134	blad. Hvis du ønsker deg en ny sykkel, sier de nei. Det får du ikke.	149
149	Så syk er du ikke.	154
154	Det er ikke lurt å stå rett opp fra senga og gå på skolen med en	170
170	gang når du har vært syk. Du trenger litt tid for å bli helt bra, og da	187
187	er det best å være hjemme en dag til. Men så er du frisk og kan	203
203	møte vennene dine igjen.	207

Totalt antall ord: ____

Antall ord feil (inkl. utelatte ord): ____

Antall ord korrekt: ____

Grade 2 – Fall 1. 3

Fadi – en brevvenn

0	Jeg heter Fadi. Jeg har en venn som jeg skriver til. Han bor i	14
14	et land langt borte. Der er det kaldt. Her er det varmt. Han er	28
28	lys, og jeg er mørk. Vi går begge på skolen og begge liker	41
41	musikk.	42
42	Vi skriver til hverandre en gang i uka. Nå skrev jeg om den	55
55	løse tanna mi. En dag da jeg satt og spiste, falt tanna ut. Hos	69
69	oss kaster vi tanna opp på taket av huset. Vi vil at en fugl skal	84
84	komme med ei ny tann. Jeg fortalte min venn om alt dette.	96
96	Han skrev tilbake og sa at når han hadde ei tann som var løs,	110
110	festet han en tråd i tanna. Så festet han tråden i ei dør. Så	124
124	ventet han til noen kom og åpnet døra. Da for tanna ut. Om	137
137	kvelden la han den i et glass med vann. Da han sto opp, var	151
151	det ingen tann som lå der lenger, men en mynt.	161
161	Neste gang jeg mister ei tann, skal jeg også legge den i et	174
174	glass med vann og håpe på at det skjer noe med tanna i løpet	188
188	av natta.	190

Totalt antall ord: ____

Antall ord feil (inkl. utelatte ord): ____

Antall ord korrekt: ____

Grade 2 – Winter 2.1

Liten og søt – Bolla pinnsvin

0	Bolla pinnsvin skal på skolen. Hun får mat av mor. Så går	12
12	hun. Når klokka slår, sitter hun på plassen sin sammen med	23
23	de andre dyra i skogen.	28
28	Om natta leter Bolla etter mat. Så går hun hjem for å legge	41
41	seg. Huset er under en hekk. Når det blir vinter, legger Bolla	53
53	seg til å sove. Hun står ikke opp før det er vår.	65
65	Hvis du har et Bolla pinnsvin i hagen, skal du ikke gi det	78
78	melk eller brød. Da kan Bolla bli syk. Bolla kan få vann og	91
91	eple.	92
92	Det er fint å ha et pinnsvin i hagen. Det spiser mus og mark.	106
106	Hvis det ikke bor et pinnsvin i hagen din, kan du legge ut litt	120
120	mat som du vet det liker. Da kan du få en ny venn. Men pass	135
135	på at ikke rotter og mus tar maten.	143
143	Pinnsvinet bryr seg ikke mye om oss. Det kan rusle rundt	154
154	beina våre og snuse etter noe å spise. Du skal ikke ta på et	168
168	pinnsvin. Da kan du bli syk.	174
174	Vi må ta vare på pinnsvinet. Det er fredet. Mange blir drept	186
186	når de går over en vei om natta. De går så seint, og bilene ser	201
201	dem ikke.	203

Totalt antall ord: ____

Antall ord feil (inkl. utelatte ord): ____

Antall ord korrekt: ____

Grade 2 – Winter 2.2

På hytta

0	Stine gleder seg til å reise på hytta sammen med mor og far.	13
13	De drar av sted i bilen og etter to timer kommer de frem.	26
26	Hytta ligger på fjellet. De må gå en time. Stine må bære en	39
39	sekk. Hun blir trett, men vet at det ikke er langt igjen. Far har	53
53	gått foran. Når mor og Stine kommer frem, har han tent opp i	66
66	ovnen.	67
67	Stine liker seg om kvelden. Mor har tent mange lys. Det er	79
79	varmt og godt. Nå har mor og far god tid. De spiller spill.	92
92	Stine får også være oppe så lenge hun vil.	101
101	Neste dag går de på tur. I dag er det fint vær, men Stine har	116
116	også vært her når det er mye vind. Da kan hun nesten ikke	129
129	stå. De finner seg en fin plass og tenner opp et bål. Mor og	143
143	far drikker kaffe, mens Stine får en kopp varm kakao. Det er	155
155	en fin utsikt fra der de sitter. Langt der nede kan de se hytta.	169
169	Den er så liten.	173
173	Så går de tilbake til hytta og lager en god middag. Dagen	185
185	etter går de en liten tur før de drar tilbake til byen.	197

Totalt antall ord: ____

Antall ord feil (inkl. utelatte ord): ____

Antall ord korrekt: ____

Grade 2 – Winter 2. 3

Skoleveien

0	Hvordan kommer du deg til skolen? Bor du i en by kan	12
12	du enten gå eller sykle til skolen. De som bor ute på	24
24	landet og har lang vei til skolen, må ta bussen. Noen	35
35	steder er det høye fjell. Når det regner mye, kan jord og	47
47	stein falle ned på veien. Om vinteren kan det komme	57
57	mye snø. Da må alle være hjemme.	64
64	Andre bor på små øyer ute ved havet. Det er ikke skole	76
76	på øya, og de må reise med båt for å komme på skolen.	89
89	Noen dager er det store bølger. Men elevene er vant til	100
100	det. Når de blir eldre, må de gå på skole et annet sted. Da	114
114	kommer de hjem bare hver helg.	120
120	Rundt om i verden kommer barn seg til skolen på mange	131
131	måter. Noen må gå langt. Og de som er eldst, må passe	143
143	på de små. Noen barn reiser rundt sammen med	152
152	familiene sine. Da må lærerne komme til dem.	160
160	Andre steder er det ikke skole der de bor. Da må de ha	173
173	skole over radioen. Læreren sitter et helt annet sted og	183
183	snakker til dem. Mor og far passer på at de gjør leksene	195
195	sine til neste gang det er time på radioen.	204

Totalt antall ord: ____

Antall ord feil (inkl. utelatte ord): ____

Antall ord korrekt: ____

Grade 2 – Spring 3. 1

Redd for mørket

0	Jeg er redd for mørket. Det jeg er mest redd for, er at et dyr	15
15	skal komme og bite meg. Jeg er også redd for at noen skal	28
28	følge etter meg. En gang vi var på hytta skulle jeg gå hjem fra	42
42	en venn. Det blåste og var mørkt. Veien var våt. Da jeg	54
54	begynte å gå, hørte jeg at det sa "plopp", "plopp" bak meg.	66
66	Jeg ble veldig redd og løp så fort jeg kunne. Hele tiden hørte	79
79	jeg den samme lyden. Da jeg kom frem til hytta, snudde jeg	91
91	meg. Det var ingen der. Det var jeg selv som hadde laget	103
103	lyden med skoene mine da jeg løp.	110
110	Det var fint å komme inn i den varme hytta. Far så på meg og	125
125	spurte om jeg hadde løpt. Litt, svarte jeg, men jeg sa ikke	137
137	hvor redd jeg hadde vært.	142
142	Den kvelden var det godt å legge seg. Mor leste et eventyr for	155
155	meg, og jeg sovnet fort. Om natten drømte jeg at noen kom	167
167	etter meg. Da de skulle gripe tak i meg, våknet jeg. Da var	180
180	det fint å vite at jeg lå på rommet mitt, og at det bare var en	196
196	vond drøm. Nå er jeg ikke så redd for mørket lenger.	207

Totalt antall ord: ____

Antall ord feil (inkl. utelatte ord): ____

Antall ord korrekt: ____

Grade 2 – Spring 3.2

Vi blir alle sinte

0	Alle kan bli sinte. Det er mange ting som kan få oss sint. De	14
14	voksne blir ofte sinte når de sitter i bil. Det går for seint. Eller	28
28	når de skal ut å fly og må stå i en lang kø. De kan også bli	45
45	sinte hvis barna ikke vil legge seg om kvelden. Ingen barn	56
56	liker at voksne snakker med sint stemme.	63
63	Hva er det som gjør deg sint? Er det når noen sier stygge ting	77
77	og erter deg? Blir du sint hvis du får skylda for noe som du	91
91	ikke har gjort? Eller blir du sint på deg selv når du ikke får til	106
106	stykkene i matte?	109
109	Hvordan vet vi at vi blir sinte? Jo, vi blir rød og varm.	122
122	Stemmen kan også bli høy og rar. Av og til kan vi bli så sint	137
137	at vi får lyst til å slå. Da må vi tenke oss om. Vi kan gå vekk	154
154	og telle til ti. Noen er sinte ei lang stund og tenker mye på det	169
169	som har hendt. Andre blir fort sinte, men det går snart over.	181
181	Alt er glemt etter ei kort stund. Hva er du? Og hva gjør du når	196
196	du blir sint?	199

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 2 – Spring 3. 3

Livet ved elva

0	Jeg heter Lin og bor i en liten by i Kina. Det er ikke så mange	16
16	hus her, og alle kjenner alle. Byen ligger ved ei stor elv. Vi	29
29	vasker oss i elva. Når vi skal dra et annet sted, tar vi båten.	43
43	Når sola står opp, vekker mor meg. Jeg lager mat. Når vi har	56
56	spist, må vi se etter dyrene våre. Vi har ei ku og en gris.	70
70	Hos oss må barna hjelpe mor og far fra de er små. Jeg hjelper	84
84	til med å holde huset i orden og å lage mat. Når jeg blir lei,	99
99	drar jeg ned til elva og leker med vennene mine. Vi hopper	111
111	fra et tre og har det moro. Vi lærer å svømme fra vi er små.	126
126	Om kvelden sitter vi sammen ved bålet. Snart går sola ned.	137
137	Da forteller de gamle om hvordan det var da de var barn. Det	150
150	er lenge siden. Det liker jeg å høre på. De forteller om	162
162	hvordan elva fikk sin farge.	167
167	Jeg liker skolen. Jeg kan lese og skrive. Når jeg blir stor, skal	180
180	jeg flytte til en stor by og gå på skole der. Jeg vil bli lærer. Da	196
196	skal jeg fortelle alle barna om livet ved elva.	205

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 3 - Fall 1.1

Fotballturnering i Danmark

0	Jeg spiller på jentelaget i fotball. I år har vi samlet inn penger for å dra	16
16	til Danmark.	18
18	Turen dit var lang. Først kjørte vi buss i mange timer. Fra Oslo tok vi	33
33	båten til Danmark og tog videre. Vi var slitne da vi kom frem til	47
47	skolen hvor vi skulle bo.	52
52	Den første kampen var mot et dansk lag. De var litt bedre enn oss og	67
67	vant med to mål. Vi fikk trøst av treneren vår som sa at vi trengte litt	83
83	tid for å komme i gang. Han hadde rett. Vi vant de to neste kampene	98
98	og ble nummer to i vår gruppe. Fra nå av var det cup. Tapte vi, var vi	115
115	ute av spillet.	118
118	Vi spilte først mot et svensk lag. Det sto null-null helt til siste minutt.	132
132	Jeg sendte i vei et langt spark. Line kastet seg frem og fikk tåa på	147
147	ballen. Vi vant. Treneren vår hoppet opp og ned. Han skrek av glede.	160
160	Det neste laget vi skulle møte var fra Tyskland. Vi hadde sett dem	173
173	spille og visste at vi ikke ville vinne. Vi prøvde så godt vi kunne, men	188
188	de var alt for sterke for oss. Snipp, snapp, snute, vi var ute.	201
201	Det gjorde ikke så mye. Nå fikk vi tid til å kose oss. Jeg brukte alt for	218
218	mange penger på klær og sko.	224
224	Vi har bestemt oss for å være med neste år også. Det betyr at vi må	240
240	jobbe hardt for å samle inn penger. Det skal vi greie.	251

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 3 - Fall 1.2

På biblioteket

0	Det er sikkert et bibliotek nær der du bor. Vet du hvor det er? Har du	16
16	vært der i det siste? Dersom du liker å lese bøker, er biblioteket stedet	30
30	du skal gå til. Der er det mange bøker som passer for deg. Og der	45
45	finner du alle bøkene som er kommet ut det siste året. Hvis du vet hva	60
60	slags bok du skal ha, er det lett å finne den i hylla. Dersom boken er	76
76	lånt ut, kan du stå på en liste og få låne boka når den kommer inn. Du	93
93	kan ha en bok i fire uker før du må levere den tilbake. Du får god hjelp	110
110	fra dem som jobber der. De kan mye om alle slags bøker.	122
122	Du kan låne cd`er av den musikken du liker, eller du kan låne en film	137
137	som du vil se. På noen bibliotek kan du også låne pc-spill. Du kan	151
151	også låne lydbøker. De er gode å ha når dere skal kjøre langt med bil.	166
166	Da går turen mye fortere.	171
171	Det er hyggelig på biblioteket. Inne er det ganske stille. Du kan sette	184
184	deg i en krok og lese bøker, blader eller tegneserier. Hvis du går dit	198
198	ofte, vil du se at det er mange som sitter der hver dag. Noen kommer	213
213	bare for å lese avisa.	218
218	Noen ganger er en forfatter på besøk og forteller om boken han har	231
231	skrevet. For små barn er det eventyrstund.	238
238	Dersom du blir glad i biblioteket, vil du helt sikkert gå dit ofte, også	252
252	når du blir voksen.	256

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 3 - Fall 1.3

På leirskole

0	Klassen skulle på leirskole. Vi møtte opp på skolen klokka åtte med	12
12	hver vår sekk med masse klær i. Så kom en stor buss og kjørte i	27
27	mange timer oppover en lang dal. Vi bråkte en del, og lærerne ba oss	41
41	være stille.	43
43	Langt oppe på fjellet lå leirskolen. Det var plass til fire på hvert rom,	57
57	og vi fikk velge hvem vi ville være sammen med.	67
67	I stallen var det mange hester. Noen av jentene var flinke til å ri fordi	82
82	de red nesten hver dag etter skolen. Vi andre som aldri hadde ridd før,	96
96	syntes det var litt vanskelig til å begynne med. Noen av hestene ville	109
109	ikke gå. De var lei av å ha unger på ryggen hele dagen.	122
122	Stedet lå ved et lite vann. Vi fikk også lære å padle kano. Det var ikke	138
138	lett å padle rett frem. Kanoen hadde lett for å svinge til siden. Av og til	154
154	svingte den helt rundt. Men vi ble flinkere etter hvert. Da padlet vi om	168
168	kapp.	169
169	Vi lærte mye på leirskolen. Vi gikk tur i fjellet og lærte om planter og	184
184	dyr. En dag så vi en ørn som svingte seg langt oppe i luften. En annen	200
200	dag så vi en rev som lusket rundt på jakt etter mat. Vi så også en flokk	217
217	med rein.	219
219	Om kvelden skulle det være ro klokka ti. Da gikk lærerne rundt,	231
231	slukket lyset og sa god natt. Men vi ble liggende lenge å prate. Noen	245
245	sprang rundt i gangene. Guttene ville besøke jentene. Da kom lærerne	256
256	og jaget dem i seng. Vi var ganske trøtte om morgenen, men vi måtte	270
270	opp for å spise frokost klokka åtte.	277
277	Dagene gikk veldig fort, og ingen lengtet hjem. Så kom bussen og	289
289	hentet oss. Nesten alle sov hele veien hjem til byen.	299

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 3 – Winter 2.1

Hvorfor leker ikke gutter med jenteting - og jenter med gutteting?

0	Er jenter og gutter like? Både ja og nei. Jentene liker å leke med	14
14	dukker. De kler av og på dem, bader dem og legger dem til sengs	28
28	mens de synger for dem. Guttene liker best å leke med biler og fly. De	43
43	bygger tårn av klosser som de river ned igjen. Rosa er en farge som	57
57	jentene liker. Få gutter liker rosa, de vil heller ha blå eller en annen	71
71	mørk farge.	73
73	Både jenter og gutter liker å leke med lego, spill og PC. Når de blir	88
88	store, vil jentene bli små prinsesser og guttene vil bli politi eller	100
100	brannmann. Jentene skal være små og søte, mens guttene skal være	111
111	store og tøffe.	114
114	Hvorfor er det slik? Er det fordi vi er født sånn? Eller er det de som	130
130	lager og selger leker som vil ha det slik? Er det mor eller far som	145
145	bestemmer? Mor forteller jentene om hva hun gjorde da hun var liten.	157
157	Far forteller guttene hva han gjorde. Og så tar jenter og gutter etter	170
170	mor og far.	173
173	Det er lettere for ei jente å leke med det som guttene liker enn det er	189
189	for en gutt å leke med det som jentene liker. En gutt som leker med	204
204	dukker kan bli ertet, mens ei jente som leker med fly får være i fred.	219
219	Hvorfor er det slik?	223
223	Det ville være fint om både gutter og jenter fikk leke med det de	237
237	hadde lyst til, og at ingen ertet dem for det. Da kunne de også leke	252
252	mer sammen.	254

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 3 – Winter 2.2

Bestefar og Ole på fisketur

0	Ole skulle være med bestefar for å dra opp garnet han hadde satt	13
14	kvelden før. Det var kaldt ute på havet. I båten var det varmt. Ole	27
27	skjønte ikke hvordan bestefar greide å finne frem i mørket. Men	38
38	bestefar var godt kjent. Han hadde fisket her i mange år. Bestefar	50
50	drakk kaffe, og Ole drakk varm kakao som mor hadde laget før han	63
63	dro. Det var godt.	67
67	Etter en time kom de frem. Nå var det blitt lyst. Bestefar stoppet	80
80	båten, og de begynte å dra opp garnet. Det var tungt. Her er det noe	95
95	stort på, sa bestefar. Hva kan det være, spurte Ole. Jeg tror det er ei	110
110	kveite, svarte bestefar. Noe stort og brunt dukket opp. Kveita plasket	121
121	vilt med halen og ville slett ikke la seg fange. Den er for tung til å få	138
138	opp i båten, sa bestefar. Vi må slepe den etter oss. Båten gikk innover	152
152	med sakte fart. Ole styrte, og bestefar passet på at de ikke fikk garnet i	167
167	propellen.	168
168	Da de kom til land, fikk de hjelp av far og noen andre folk. De festet	184
184	et tau i kveita og heiste den opp på brygga. Vekta viste nesten hundre	197
197	kilo. Det er den største kveita jeg noen gang har fått, sa bestefar. Det	212
212	var bra du var med, sa han til Ole. Uten din hjelp ville vi aldri ha greid	228
228	det. Ole ble stolt og bestemte seg for at han skulle bli fisker når han	244
244	ble stor. Han gledet seg til å fortelle mor om alt han hadde opplevd.	258

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 3 – Winter 2.3

Ville dyr på besøk

0	Ole bodde i et hus på landet. Ole likte naturfag på skolen, men mest av	15
15	alt likte han å se ville dyr ute i naturen. Av og til fikk de besøk av rev	33
33	og rådyr. Ole kunne stå i vinduet og se på dem. Særlig reven var ivrig.	48
48	Den luktet at det var høner i hønsehuset og ville gjerne ha en til	62
62	middag.	63
63	En kveld hørte Ole skritt på terrassen. Gjennom vinduet så han et stort	75
75	hode. Han ble litt redd og ble sittende helt stille. Hva var dette? Hodet	89
89	snudde seg sakte rundt, og da så han hva det var - en stor elg. Elgen	105
105	sto helt rolig og stirret på seg selv i vinduet. Den virret litt på hodet.	120
120	Ole ble redd for at den trodde den så en annen elg. Ole visste at elger	136
136	kunne bli veldig sinte og stange mot hverandre med full kraft.	148
148	Akkurat da kom Kari inn i stuen. Hun er Oles lillesøster og bare tre år	162
162	gammel. Hva er det, spurte hun da hun fikk se det store hodet. Hun	177
177	løp bort til vinduet, mens hun lo høyt. Da elgen fikk se Kari, trakk den	192
192	hodet til seg. Den hoppet ned fra terrassen, rev med seg to stoler og et	207
207	bord og forsvant i full fart. Da Ole fortalte mor og far om elgen, ble de	223
223	litt redde. Far sa at han aldri hadde hørt om at elgen stanget inn ruter.	238
238	Noen dager senere kunne de lese i avisen at en elg hadde knust en stor	253
253	rute og løpt inn i en butikk som solgte bøker. Den ødela mye før den	268
268	fant døren og løp ut igjen i skogen. Kanskje det var den samme elgen?	282

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 3 – Spring 3.1

Fremmed i Norge

0	Jeg heter Ali og er fra Iran. Jeg kom til Norge sammen med moren	14
14	min og lillebroren min for ett år siden. Vi har ikke hørt fra pappa på en	30
30	lang stund, men håper alt er bra med ham.	39
39	Jeg bor på et sted sammen med mange andre barn og familier. Det er	53
53	ikke lett. Vi venter hele tiden. Norge er et venteland. Jeg vet ikke om	67
67	vi kommer til å bo i Norge eller ikke. Det er verst for mamma. Hun	82
82	gråter mye. En familie som kom etter oss, har fått bli i landet. Det	96
96	tenker jeg mye på og håper at vi også er like heldige.	108
108	Det er ikke lett på skolen. Jeg kan litt norsk, men ikke godt nok til å	124
124	snakke med de andre i klassen.	130
130	Jeg tenker mye. Jeg vil ikke fortelle om alt det vonde jeg har opplevd.	144
144	Jeg vil helst glemme alt. Det går best om dagen når jeg har ting å	159
159	gjøre. Om natta våkner jeg flere ganger av vonde drømmer. Det er	171
171	også mange andre her som er våkne om natta. Noen går frem og	184
184	tilbake, og du kan også høre at noen har mareritt.	194
194	En dag var vi sammen med en gruppe som jobbet for Redd Barna. De	208
208	hentet oss i buss, og vi ble kjørt til en skøytebane. Der fikk vi låne	223
223	skøyter. Det var ikke lett å gå på skøyter, men etter en stund greide jeg	238
238	det bedre. Vi fikk pølser og brus. Den dagen lo jeg mer enn jeg hadde	253
253	gjort på lenge. Håper de finner på andre ting også.	263

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 3 – Spring 3.2

Ei øy blir født

0	For mange år siden var en mann fra Island ute og fisket. Da så han at	16
16	havet begynte å koke og boble. Mannen trodde først at det kom fra en	30
30	båt som sto i brann. Han tok feil. Det var ei øy som var i ferd med å	48
48	bli født. En vulkan hadde våknet. Lava steg opp. Snart var det en stor	62
61	sky av røyk over stedet.	67
67	Vulkanen fortsatte å spy ut lava i over tre år, og øya ble større og	82
82	større. Snart slo fugler seg ned. I lavaen ble det grønne flekker. Frø fra	96
96	planter kom med vinden. Etter en stund begynte trær, blomster og	107
107	gress å vokse. Ingen folk har bodd på øya. Bare de som ville vite mer	122
122	om fuglene og blomstene, fikk bo i ei lita hytte for en kort tid.	136
136	Havet og vinden har nå slitt på øya. Den er blitt mindre og mindre.	150
150	Om noen år vil det bare være en liten klippe igjen med gress på toppen	165
165	som stikker opp av havet.	170
170	På Island er det mange vulkaner og varme kilder. I noen av kildene er	184
184	det fint å bade. Selv om det er vinter og kaldt, er vannet varmt. Mange	199
199	drar til Island for å bade i kildene og se på den fine naturen.	213
213	En vulkan kan sove i mange år, og det er ingen som vet når den	228
228	våkner igjen. Når det kommer et utbrudd, strømmer lava og aske ut.	240
240	Lavaen renner nedover fjellsiden. Asken kan drive langt av sted med	251
251	vinden.	252

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 3 – Spring 3.3

Flytte til byen

0	Lene hadde flyttet til byen sammen med familien sin. Det var flere	12
12	uker til skolen skulle begynne. Hun savnet de gamle vennene sine.	23
23	Hun var vant til å være ute og leke, men her var gata tom for folk. Det	40
40	sto bare biler der.	44
44	En dag møtte hun to andre jenter utenfor huset. Da de fikk se Lene,	58
58	stoppet de. En av dem spurte om hun var ny her.	69
69	Ja, svarte Lene. Vil du være sammen med oss, spurte jentene som sa at	83
83	de het Kari og Trine. Vi skal dra til stranda og bade.	95
95	Lene måtte spørre mor først, som sa at hun fikk lov. Men husk å	109
109	være forsiktig, sa mor. Lene sukket. Det hadde hun hørt mange ganger	121
121	før.	122
122	De måtte ta bussen et lite stykke før de kom til stranda. Jentene kledde	136
136	av seg og hoppet ut i det varme vannet. Etterpå satte de seg på teppene	151
151	sine og så på de andre. Lene fortalte at hun kom fra et lite sted på	167
167	landet, og at hun hadde gått i en klasse med bare seks andre elever.	181
181	Hun grudde seg til å begynne i en stor klasse. Hun var ikke vant til å	197
197	ha så mange andre rundt seg.	203
203	Da fortalte Kari at hun også hadde flyttet til byen for noen år siden og	218
218	visste akkurat hvordan Lene hadde det. Nå kan vi alle være venninner,	230
230	sa hun. Lene kjente at hun ble varm av glede. Det var ikke noe å grue	246
246	seg for lenger.	249

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 4 - Fall 1.1

Besøk i klassen av en forfatter

0	I en time fikk klassen besøk av en kjent forfatter som skrev spennende bøker for	15
15	barn og unge. Elevene hadde lest flere av bøkene hans. De skulle finne frem til	30
30	spørsmål som de kunne stille til forfatteren.	37
37	Da han kom, satte han seg foran klassen. Han var mye eldre enn elevene hadde	52
52	trodd. De ble litt skuffet. Elevene kunne se at han var litt nervøs. Han sa at han	69
69	ikke var så vant til å sitte foran en hel klasse. Men han ville gjøre så godt han	87
87	kunne.	88
88	Har dere lest noen av bøkene, spurte han. Tone, som alltid rakte opp hånda i	103
103	timene, var først ute. Jeg har lest to av bøkene dine, svarte hun. Jeg likte den som	120
120	het "Det spøker i natten" best.	126
126	Forfatteren nikket. Han fortalte at det var mange andre også som hadde likt den	140
140	boka godt. Så begynte han å lese fra den; der det var mest spennende. Læreren	155
155	trakk for gardinene og slo av lyset.	162
162	Forfatteren leste sakte og med lav stemme. Elevene måtte sitte helt stille for å	176
176	høre hva han sa. Da han kom til stedet i boken der spøkelset kom frem og viste	193
193	seg for familien i huset, senket han stemmen enda mer. Alle bøyde seg frem. Så	208
208	klappet han boka sammen med et høyt smell, og alle skvatt høyt.	220
220	Nå var det tid for spørsmål. Guttene i klassen som likte best å spille data, hadde	236
236	også lest bøkene og de var like ivrige etter å stille spørsmål. Forfatteren svarte så	251
251	godt han kunne og leste også fra en annen bok han hadde skrevet.	264
264	Før han gikk, sa han at han likte å være i klassen og var glad for at så mange	283
283	leste bøkene hans. Og ekstra glad var han for at guttene også leste. Fortsett med	298
298	det, gutter, sa han. Så skal jeg fortsette å skrive for dere.	310

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 4 - Fall 1.2

Flytte til London

0	En dag kom mor hjem og fortalte at hun hadde fått ny jobb. Hun ville at familien	17
17	skulle flytte til London og bo der ett år. Knut nektet først å være med. Han ville	34
34	ikke dra fra skolen og vennene sine. Men etter å ha tenkt seg om, skjønte han at	51
51	dette var en fin måte å bli kjent med et annet land på. Han ville også bli god i	70
70	engelsk.	71
71	I London var det mye trafikk. Alle drosjene var svarte. Alle bussene var røde og	86
86	hadde to etasjer. Knut likte å kjøre buss for da kunne han se ned på alle folkene	103
103	som var ute og gikk.	108
108	Knut likte veldig godt å gå på fotballkamper sammen med far. Knut holdt mest	122
122	med Arsenal. Han kjøpte seg en rød og hvit trøye som han alltid hadde på seg når	139
139	Arsenal spilte hjemme. Far sa at han så ut som om han hadde bodd i London hele	156
156	sitt liv. Han var blitt en innfødt.	163
163	Knut gikk på den norske skolen i London. I engelsk hadde de en streng lærer fra	179
179	London. Hun hadde grått hår og briller helt ute på nesen. Knut var litt redd	194
194	henne. Han syntes skolen var vanskelig til å begynne med. Mange av elevene	207
207	hadde vært i England i flere år og kunne språket godt. Knut likte ikke å snakke	223
223	høyt i timen, men etter hvert ble han flinkere og var snart like ivrig som de	239
239	andre. Nå likte han også sin engelske lærer bedre.	248
248	Året gikk veldig fort, og snart kom dagen da de skulle dra hjem. Knut nektet	263
263	også denne gangen, men han visste det ikke nyttet. Han begynte å glede seg til å	279
279	møte de gamle vennene sine igjen. Nå var han helt sikker på at han var den beste	296
296	i klassen til å skrive og snakke engelsk.	304

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 4 - Fall 1.3

Å padle kajakk

0	Noen kajakker er smale og lette. Andre er lengre og bredere. De kalles havkajakk	14
14	fordi de tåler mye vind og sjø. Når vi skal på tur på fjorden eller ute på havet, er	33
33	det best å bruke en havkajakk. Da kan vi også ha med telt, sovepose og annet	49
49	utstyr.	50
50	Dersom du vil begynne å padle kajakk, er det lurt å melde deg inn i en klubb. På	68
68	kurs lærer du hvordan du skal sette deg ned i kajakken, og hvordan du skal	83
83	komme deg opp av den på en enkel måte. Du lærer også hvordan du kan redde	99
99	kameraten din hvis en av dere velter. På noen kurs kan du lære hvordan du kan	115
115	rulle helt rundt i kajakken og komme på rett kjøl ved egen hjelp. Det kalles	130
130	eskimorulle. Den er ikke lett.	135
135	Det er finest å padle tidlig om morgenen. Da er det lite vind og ingen bølger. Du	152
152	glir av sted uten en lyd. Du padler i trange sund og rundt små øyer. Kanskje ser du	170
170	en ørn som svever over deg eller en oter som plasker i sjøen foran deg.	185
185	Det er også spennende å bruke kajakken når det blåser og bølgene er store. Da	200
200	lærer du deg hvordan du skal padle når bølgene kommer mot deg og når du har	216
216	dem bak deg. Det er lettest å velte når de kommer fra siden. Da gjelder det å ha	234
234	god balanse og ikke bli redd.	240
240	Mange båter tar ikke hensyn når de møter en kajakk. De slakker ikke på farten	255
255	selv om de vet at båten deres lager store bølger. Av og til kan du bli så sint at du	275
275	roper etter dem. De hører deg aldri.	282
282	Det er alltid tryggest å padle sammen med noen. Da vet du at dere kan hjelpe	298
298	hverandre hvis en av dere velter.	304

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 4 - Winter 2.1

Sykkeltur i Frankrike

0	Da vi kom frem til hotellet om kvelden, var vi slitne etter den lange reisen med fly	17
17	og tog. Vi var i Frankrike på en ukes sykkeltur. Vi skulle leie sykler og sykle fra	34
34	hotell til hotell rundt omkring i området.	41
41	Neste morgen fikk vi utlevert syklene og kart over ruten. Bagasjen vår ville bli	55
55	fraktet til det neste hotellet. Ut av byen var det en del trafikk, men snart svingte vi	72
72	inn på en liten vei langs jorder og høye trær. Til å begynne med var det flatt, men	90
90	snart kom vi til en lang bakke. Oppe på toppen var det en liten landsby. Husene lå	107
107	tett i tett rundt et lite torg med flere små butikker. Vi kjøpte mat og drikke og satte	125
125	oss ved et lite bord nede ved elven som fløt gjennom byen. Været var fint. Folk	141
141	som gikk forbi, hilste og snakket til oss. Både mor og far kunne litt fransk og	157
157	svarte tilbake så godt de kunne. Søsteren min Tone og jeg forsto ikke så mye.	172
172	Vi fortsatte turen. Nå gikk det nedover igjen, og vi fikk stor fart. Nede på sletta	188
188	kom vi forbi jorder hvor det sto lange rekker med små trær. Det er vinmarker,	203
203	forklarte far. Her dyrker de druer som blir til den beste franske vinen. Jeg tror	218
218	både mor og far gledet seg til å komme frem til hotellet og smake på den.	234
234	Ut på ettermiddagen kom vi frem til hotellet. Det lå ved en kanal og var veldig	250
250	koselig. Etter å ha dusjet, gikk vi ned i spisesalen. Maten var god. Vi fikk flere	266
266	retter og til slutt kom de med et bord med mange oster på.	279
279	Slik fortsatte turen. Vi syklet gjennom mange små byer. Folkene vi traff var	292
292	hyggelige. Dette gjør vi også neste år, sa mor. Vi var helt enig med henne.	307

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 4 - Winter 2.2

En hund i huset

0	Det er mange barn og unge som synes valper er små og søte og gjerne ønsker seg	17
17	en. Men tenker de godt nok på hvor mye arbeid det er med en hund i huset? Og	35
35	hva slags hund skal familien kjøpe?	41
41	Den første tiden tisser den lille valpen på gulvet, og det er som oftest mor og far	58
58	som må tørke opp etter den.	64
64	Hver morgen må den luftes. Hvem skal gjøre det? Kanskje er det regn og vind	79
79	ute, og det er mye bedre å ligge under dyna og vente på at noen roper at nå er det	99
99	frokost.	100
100	På ettermiddagen må hunden ha en lang tur. Den må få løpe fritt, snuse på	115
115	forskjellige ting og hilse på andre hunder.	122
122	En hund må også lære seg å gå pent i bånd, sitte pent og å komme tilbake når vi	141
141	roper på den. Det tar mye tid å lære hunden alt dette. Hunden lærer best når den	158
158	får ros. Enten at man klapper den på hodet, eller at den får små godbiter.	173
173	Det er forskjellige typer hunder. Noen er store og sterke, andre er så små at de får	190
190	plass i en skoeske. Jakthunder må ha lange turer hver dag der de kan løpe mye.	206
206	Andre hunder trenger ikke så mye mosjon, men de må også luftes.	218
218	Det er også forskjell på en hann og ei tispe. En hann som tror den er sjefen over	236
236	alle andre hunder, kan lett begynne å slåss for å vise hvem som er sterkest.	251
251	Det er derfor lurt å tenke seg om flere ganger før dere kjøper hund. Den blir fort	268
268	stor, og hele familien må være enig om at de må dele på arbeidet. Men har dere	285
285	først bestemt dere, er det mye glede og hygge med en ny venn i huset.	300

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 4 - Winter 2.3

En rask og stor katt - geparden

0	Geparden er det raskeste dyret i verden. Den finnes for det meste i Afrika. Når	15
15	den jakter, kan den løpe over 100 kilometer i timen.	25
25	Geparden er en stor katt. Den brøler ikke slik som løven, men maler og hveser	40
40	som katten. Hodet er lite. En svart flekk går ned fra øyet til munnen. Den ser litt	57
57	trist ut, nesten som den gråter. Ørene er små. Pelsen er gul med svarte flekker og	73
73	gjør den nesten usynlig i gresset. Geparden får vanligvis tre til fem unger. De	87
87	fleste av ungene dør i løpet av de første månedene. De blir et lett bytte for løver.	104
104	Geparden jakter på antiloper og andre dyr. Skjult av det høye gresset smyger den	118
118	seg forsiktig frem mot byttet. Den stopper opp noen ganger for så å krype enda	133
133	nærmere. Når den er kommet nær nok, setter den opp farten ved hjelp av de store	149
149	bakbena. Det er da den løper like fort som en bil på motorveien i Norge. Når den	166
166	kommer helt opp til byttet, slår den til med forlabben slik at dyret snubler. Så	181
181	kaster den seg over byttet, biter det i halsen og kveler det.	193
193	Noen ganger jakter flere geparder sammen. Da kan de ta større dyr som gnu og	208
208	sebraer.	209
209	Mange geparder blir skutt av jegere fordi de tar husdyrene deres. Det er lett for	224
224	en gepard å ta en liten kalv som nesten ikke kan løpe. Nå er det ikke så mange	242
242	geparder igjen. Det fører til innavl. Det betyr at geparden får unger med en annen	257
257	gepard som den er i nær slekt med. Mange av ungene dør og de som vokser opp,	274
274	blir lett svake og syke. Arten er truet. Det er synd fordi geparden hører til i	290
290	naturen vår.	292
292	Det er mulig å temme geparden. Den ble holdt som kjæledyr for rike folk.	306

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 4 - Spring 3.1

Brønnen

0	Lars spurte om å få 500 kroner av sine foreldre. Han hadde hørt på skolen at det	17
17	kostet så mye å bygge en brønn i en landsby i Afrika. Mange barn i Afrika blir	34
34	syke fordi de mangler rent vann. Selv om Lars bare gikk i fjerde klasse, var han	50
50	helt bestemt på å hjelpe barna.	56
56	Foreldrene sa at han måtte arbeide for pengene. Han hjalp til i huset, slo gresset,	71
71	malte gjerdet og til slutt hadde han tjent nok penger.	81
81	Lars ga pengene til Redd Barna som skulle sende dem videre til Afrika. De sa at	97
97	pengene bare var nok til en pumpe for å få vannet opp av brønnen. Først måtte de	114
114	borre langt ned i jorda for å finne vann. Det kostet mye mer - hele ti tusen	131
131	kroner.	132
132	Men Lars gav seg ikke. Han var fast bestemt på å fortsette. Han ville at brønnen	148
148	skulle bli ferdig. Da de andre i klassen fikk høre om dette, ville de også være	164
164	med. På et foreldremøte ble det bestemt at de skulle lage et loppemarked for å	179
179	samle inn penger til brønnen.	184
184	På loppemarkedet fikk de inn nesten alle pengene. Nå hadde de nok til å betale	199
199	for både boringen og brønnen.	204
204	Brønnen ble bygd nær en skole. Etter at brønnen var ferdig, sendte elevene på	218
218	skolen brev til Norge med bilde av seg selv foran brønnen. De takket for	232
232	pengene. Alle var glade. Nå hadde de rent vann og kunne drikke så mye de ville	248
248	uten å bli syke.	252
252	Elevene i landsbyen og i klassen til Lars fortsatte å skrive til hverandre. De ble	267
267	gode venner. En dag håper Lars at han kan reise til Afrika, besøke skolen og se	283
283	brønnen med sine egne øyne. Men det må bli når han blir stor og tjener penger	298
298	selv.	300

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 4 - Spring 3.2

Emil og Eilert

0	To brødre bodde i en liten by i et land i Afrika. De hadde mistet både mor og far på	20
20	grunn av sykdom. En familie i landsbyen hadde tatt vare på dem, men de var fattige og	37
37	hadde lite mat. De kunne ikke ha dem lenger, og de to guttene hadde derfor liten sjanse	54
54	til å klare seg videre i livet.	61
61	En norsk mann og kone jobbet for Redd Barna i området. De besøkte landsbyen og fikk	77
77	høre historien om guttene. De snakket lenge sammen. De kunne ikke redde alle barna	91
91	der, men bestemte seg til slutt for å ta med de to guttene hjem til Norge. De bodde på en	111
111	gård på landet og hadde fire barn fra før. De var alle eldre enn guttene og hadde flyttet	129
129	fra gården.	131
131	De to guttene ble kalt Emil og Eilert. De skilte seg ut fra de andre barna. Det var ikke lett	151
151	for guttene til å begynne med. Men de hadde hverandre.	161
161	Emil og Eilert ble tatt godt imot av folket i bygda. Snart lærte de seg å snakke norsk.	179
179	Da var det ingen som lenger tenkte på at de kom fra et annet land. Emil og Eilert var	198
198	gode til å spille fotball. Alle ville ha dem med på laget. De var også gode til å løpe og	218
218	kaste ball.	220
220	I dag er Emil og Eilert voksne. Emil er baker og Eilert er elektriker. De har fått kjærester	238
238	og bor i hver sin by. De kommer hjem til gården så ofte de kan for å besøke foreldrene	257
257	som nå er blitt gamle.	262
262	Det er rart å tenke på at hadde ikke det norske paret kommet på besøk til den lille	280
280	landsbyen den dagen for 25 år siden så hadde kanskje ikke Emil og Eilert levd i dag.	297

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 4 - Spring 3.3

Kairo

0	Kairo er hovedstaden i Egypt. Det er en stor og spennende by, og det er mange ting å se	19
19	der. Utenfor byen ligger pyramidene. Den største og mest kjente heter Keops og er 140	34
34	meter høy. Den er graven til en egyptisk konge. Det tok over 20 år å bygge den, men	52
52	ingen vet hvordan de greide det. Noen av steinene veier hele 15 tonn.	65
65	I Kairo er det mange basarer. Der får man kjøpt smykker av gull og sølv, vaser, tepper	83
83	og mange andre ting. Gatene er trange, og det er vanskelig å finne frem hvis du ikke er	101
101	kjent. I en gate finner du bare smykker, i en annen finner du bare ting som er laget av	119
119	lær. På denne måten er det enkelt å sammenligne varer som du ønsker å kjøpe. Mange	135
135	steder får du også bli med bak i butikken og se på hvordan de lager varene.	151
151	Når du skal betale, sier selgeren en pris. Hvis selgeren ser at du har fine klær eller en dyr	170
170	klokke, setter han prisen høyt. Da tror han at du er rik. Så er det din tur. Du sier at du	191
191	bare vil betale halvparten av det han vil ha. Da tar selgeren seg til hodet og sier noe du	210
210	ikke forstår. Kanskje sier han at han har syv barn, og at han trenger pengene for at de	228
228	skal få gå på skolen; eller at kona hans er syk og han må ha penger til å betale legen.	248
248	Til slutt blir dere enige om en pris. Du vet ikke helt om du har betalt for mye eller om du	269
269	har gjort et godt kjøp. Men det spiller ingen rolle fordi du har opplevd noe nytt og	286
286	annerledes. Så får du en kopp varm te av selgeren som ønsker deg et langt liv og	303
303	velkommen tilbake.	305

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 5 – Fall 1.1

På fotballkamp i London

0	Far og onkel Ole dro hvert år over til England for å se fotball. I år skulle Mads	18
18	og Stein få være med. De gledet seg vilt. Det var første gangen for de to guttene.	35
35	De dro avgårde torsdag ettermiddag med fly til London. Kampen ble spilt først	48
48	på lørdag, så de hadde hele fredagen til å se seg om i London. De vandret rundt i	66
66	byen, handlet litt og så på alle folkene som hastet av sted. De spiste "fish and	82
82	chips" på en pub. Det er en tradisjon, sa far.	92
92	Tottenham skulle spille mot Arsenal. Det var en av toppkampene denne	103
103	sesongen. Utenfor stadion var det mange politifolk med hjelmer og køller. Flere	115
115	satt også på hester. Folk sang og skrålte. De viftet med skjerf og luer. Det var litt	132
132	av et liv.	135
135	De fant plassene sine høyt oppe på tribunen. I hver ende av stadion var det to	151
151	store skjermer. Allerede før kampen startet, var det fullt trøkk på tribunen.	163
163	Tilhengerne ropte til hverandre hele tiden og sang den ene sangen etter den	176
176	andre. Da lagene kom på banen, var lyden så høy at Mads måtte holde seg for	192
192	ørene.	193
193	Kampen bølget frem og tilbake. Mads og Stein holdt med Tottenham. De ble	206
206	revet med av alle brølene og stemningen. Når dommeren dømte mot Tottenham,	218
218	reise de seg fra stolene sine og skrek høyt. Dette var gøy. Det var noe helt annet	235
235	enn å sitte på tribunene hjemme.	241
241	Arsenal fikk det første målet. Like etter fikk Tottenham straffe. Det var helt stille	255
255	på tribunen da spilleren løp mot ballen. Pang! Rett i krysset. Senere fikk lagene	269
269	enda ett mål hver og kampen endte uavgjort. Nå hva synes dere, spurte far da	284
284	kampen var over. Vil dere være med neste år også? Mads og Stein så på	299
299	hverandre. Klart, sa de i kor. Nå var de frelst.	309

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 5 – Fall 1.2

Vikinger

0	Vikingtiden begynner på slutten av 700-tallet. Vikingene kom fra både Sverige,	11
11	Danmark og Norge. De norske vikingene dro på tokt til England, Irland og	24
24	Skottland. De fleste kjenner til historiene hvor vikingene plyndret og drepte folk	36
36	for å skaffe seg gods og gull. De var dyktige krigere og skapte sjokk og redsel	52
52	når de dukket opp. De tok med seg sølv, gull og andre verdifulle saker fra kirker.	68
68	De stjal også husdyr. Noen ganger røvet de vakre kvinner. De drepte de fleste	82
82	mennene som kjempet mot dem, men noen ble tatt med som slaver. Før de dro,	97
97	ble husene satt i brann. Like fort som de hadde kommet, forsvant de.	110
110	Skipene var bygd for å tåle store bølger, men de kunne også seile langt opp i	126
126	elvene. Derfor kom de frem til steder hvor folk følte seg trygge. Hvert skip	140
140	kunne ha 100 mann om bord og hadde 40 par årer. Både foran og bak på skipet	157
157	var det skåret ut dragehoder.	162
162	Men vikingene var ikke bare krigere. De var også dyktige sjøfolk som dro langt	176
176	av sted. Helt til Amerika seilte de - eller Vinland som de selv kalte det nye	192
192	landet. Når de seilte langs kysten, satte de navn på øyer, sund og viker og skrev	208
208	ned navnene i runeskrift på kartet.	214
214	For å finne veien, brukte de sola som kompass. Når de ikke visste hvor de skulle,	230
230	tok de bare ei loppe som de alltid hadde mange av på kroppen og satte den ned	247
247	på toften. Lopper kryper alltid mot nord, og da visste de hvor de skulle seile	262
262	videre.	263
263	Noen bosatte seg også i landene de kom til og giftet seg. De var flinke	278
278	håndverkere og handelsmenn. I Irland grunnla de Dublin.	286
286	De mest kjente gudene fra vikingtida heter Odin og Tor. Odin var konge over	300
300	alle guder.	302

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt:

Grade 5 – Fall 1.3

Redd Barna

0	Redd Barna arbeider for å hjelpe barn i hele verden. Flere hundre millioner barn	14
14	blir hvert år utsatt for sult og nød. Da gjelder det for Redd Barna å være hurtig på	32
32	plass slik at barna og deres familier får mat, vann og medisiner. Noen barn blir	47
47	også vitne til vonde hendelser under en krig. Eldre barn må noen ganger ta vare	62
62	på sine mindre søsken hvis far og mor er forsvunnet.	72
72	Dersom Redd Barna allerede arbeider i landet, er hjelpen ikke langt unna. Redd	85
85	Barna har også en gruppe mennesker i Norge som på kort tid kan reise ut og	101
101	hjelpe til der det trengs. Det aller viktigste er at barna så fort som mulig opplever	117
117	trygghet rundt seg. De må få gå på skolen, få tid til å leke og være sammen med	135
135	venner.	136
136	Redd Barna bryr seg også om hvordan barn i Norge har det. Mange familier har	151
151	lite penger. De har ikke råd til å gi barna leker og utstyr. De har heller ikke råd til	170
170	å reise på ferie sammen. Noen barn blir utsatt for vold og overgrep hjemme. Da	185
185	gjelder det at voksne som kjenner til dette, bryr seg og sier fra. Redd Barna vil at	202
202	de voksne skal ta mer ansvar for at barn får hjelp.	213
213	Redd Barna jobber ikke bare for barn, men også sammen med dem. Barna	226
226	kjenner best til det livet de selv lever. Derfor er det viktig at barna selv kan	242
242	fortelle hvordan det er å være barn der hvor de bor. Barna må også få være med å	260
260	bestemme. Hva skal til for at de føler seg trygge på skoleveien? Hvordan skal	274
274	barnehagen og skolen være? Redd Barna vil at barn skal bli hørt, og at de voksne	290
290	må ta dem på alvor.	295
295	Hvis du vil støtte Redd Barna eller en annen organisasjon som arbeider for barn,	309
309	kan du gjøre det ved å gi en fast sum penger hvert år.	322

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 5 – Winter 2.1

Veiviseren

0	En av de beste filmene som er laget i Norge er Veiviseren. Den bygger på en	16
16	gammel samisk myte eller et sagn som er nesten 1000 år gammelt.	28
28	Det er vinter, og den 16 år gamle samiske gutten Aigin har vært på jakt og er på	46
46	vei hjem til teltene der foreldrene og lillesøsteren bor. Da får han øye på noen	61
61	fremmede folk i svarte klær. De kommer fra Russland og er kjent for å drepe og	77
77	plyndre alle som kommer i deres vei. Han ser at foreldrene og lillesøsteren blir	91
91	drept.	92
92	Aigin flykter og kommer frem til en annen sameleir. Han forteller om det som	106
106	har skjedd og ber dem flykte mot kysten. Sammen med tre andre jegere legger	120
120	Aigin seg i bakhold for å angripe de fremmede. Alle blir drept unntatt Aigin,	134
134	som blir tatt til fange. De fremmede vil plyndre mer, og Aigin blir deres veiviser.	149
149	Han fører dem opp i fjellene. Det er så bratt at alle må gå i tau. Aigin leder dem	168
168	frem til en isete fjellside. Han kommer seg ut av tauet og stikker av. I jakten på	185
185	Aigin styrter alle de andre utenfor stupet og dør.	194
194	Veiviseren er en film om samenes nære forhold til naturen. Den handler også om	208
208	kampen mellom det onde og det gode. Det er et tema som også går igjen i filmen	225
225	om Ringenes Herre. Der kjemper de to små karene Frodo og Sam seg gjennom	239
239	vill og farlig natur for å beseire de onde kreftene.	249
249	Veiviseren er laget av Nils Gaup, som selv er same. Dersom du liker spennende	263
263	filmer, vil du helt sikkert også like denne. Filmen er tatt opp i Finnmark. Den	278
278	første dagen de laget filmen var det nesten 50 kuldegrader ute. Like kaldt som da	293
293	Aigin førte de fremmede på ville veier oppe i fjellene.	303

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 5 – Winter 2.2

Vannet er kilden til alt liv

0	Det er vann over alt; i jorda, i lufta, i havet, i innsjøer og i elver. Rundt begge	18
18	polene er det store masser med is. Vannet går i et kretsløp fra jord til himmel.	34
34	Det renner nedover elver og bekker og ut i sjøen. Når sola skinner og varmer opp	49
49	vannet, fordamper det. Oppe i lufta avkjøles vanndampen og blir til skyer.	62
62	Vinden blåser skyene mot land. Når de treffer fjellene, faller det nedbør som	75
75	regn eller snø.	78
78	Menneskene har lært å utnytte vannet. De første maskinene ble drevet av	90
90	vanndamp. De ble brukt i industrien, i tog og i båter. Vannet ble varmet opp av	106
106	kull, og dampen fikk stempelet i maskinen til å gå frem og tilbake. Senere lærte	121
121	vi oss å bruke fossene til å lage energi. Det kalles fornybar energi fordi vi kan	137
131	bruke vannet på nytt og på nytt. Nå må vi lære oss å utnytte bølgekraften bedre.	153
153	På en meter lager bølger nok energi til å lyse opp 30 norske hjem. Vi har også	170
170	begynt å bygge vindmøller.	174
174	Vi er heldige i Norge som har så mye vann rundt oss. Vi har en lang kyst, og i	193
193	innlandet er det mange elver og store innsjøer. Oppe i fjellene ligger det	206
206	fremdeles breer av is. Noe av vannet må vi bruke til drikkevann. Det lagres i	221
221	store dammer og renses før det sendes videre i rør til husene våre. Det er like rent	238
238	som det vannet vi kjøper på flasker. Mange land mangler vann. Det kan være	252
252	forurenset, og folk blir syke. Når jorda tørker ut, blir også avlingene mindre.	265
265	Nesten to tredjedeler av kroppen vår består av vann eller væsker. Når vi svetter,	279
279	må vi etterfylle for å skape balanse. Vi må drikke omtrent like mye som	293
293	forsvinner. Vannet rengjør også kroppen og skyller ut avfallet gjennom nyrene.	304

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 5 – Winter 2.3

Kroppen lyver ikke

0	Det er to måter å snakke med hverandre på. Den ene er gjennom språket; altså	15
15	det vi sier til hverandre. Den andre er gjennom kroppen; det vi utrykker ved	29
29	blikk, grimaser og håndbevegelser.	33
33	Kroppsspråket avslører våre følelser. Barn som lyver, vil med en gang dekke	45
45	over munnen med hendene. Når vi blir eldre, tar vi oss gjerne til nesen og øynene	61
61	flakker. Det er derfor pokerspillere bruker solbriller. De vil ikke at de andre skal	75
75	se på øynene deres at de bløffer.	82
82	Når vi er ordentlig glade, viser vi det med hele kroppen. Til og med mannlige	97
97	fotballspillere kysser og klemmer hverandre når de lager mål. Det er verre hvis vi	111
111	bare skal late som vi er glade. Vi sier en ting, men kroppen forteller noe helt	127
127	annet. Da blir det kluss.	132
132	Når vi blir sinte, blir vi stive i kroppen og truende i blikket. Se på læreren neste	149
149	gang han eller hun er sint. Er han rød i toppen? Snakker han alt for høyt? Når vi	167
167	er nervøse, lekker kroppen. Selv om vi holder overkroppen rolig, vipper vi urolig	180
180	med foten frem og tilbake under bordet.	187
187	Kroppsspråket er ikke likt over alt i verden. Kineserne pleier å rape etter et godt	202
202	måltid. Hvis en greker sier nei, nikker han på hodet. Når han sier ja, rister han på	219
219	hodet. Det er også mange måter å hilse på. I Russland kysser og klemmer både	234
234	menn og kvinner hverandre, mens i Japan bukker man formelt når man møtes.	247
247	Folk i USA er hyggelige og spør om hvordan du har det uten å vente på svar. I	265
265	Norge sier vi hei, mens i Tyskland ville man ha tatt hverandre i hånden.	279
279	Skal man reise til et land og være der en stund, er det lurt å lese litt om skikkene	298
298	på forhånd.	300

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 5 – Spring 3.1

På familietur til Malta

0	Bestefar og bestemor feiret at de hadde vært gift i 50 år. De inviterte barn og	16
16	barnebarn på en ukes tur til Malta. Til sammen var vi fire familier som dro av	32
32	sted.	33
33	Malta er ei øy i Middelhavet. Den er ganske liten, bare 27 kilometer lang. Vi	48
48	landet på flyplassen i Valletta, som er hovedstaden på øya. Vi bodde på et hotell	63
63	utenfor byen. Været var tørt og varmt. Når vi skulle dra rundt på øya, tok vi en	80
80	av de mange bussene som går på kryss og tvers. De er gamle, malt i alle farger	97
97	og bråker mye.	100
100	En dag dro vi til ei øy som heter Gozo. Der hadde bestefar og bestemor ordnet	116
116	med en konkurranse mellom familiene. Hver familie fikk sin egen åpne bil som	129
129	kalles jeep, en konvolutt med en rebus og et kart over øya. Så skulle vi finne	145
145	frem til det første stedet hvor en person skulle møte oss og gi oss neste rebus.	161
161	Det var gøy. Far kjørte, mens vi barna og mamma prøvde å løse rebusen. Til å	177
177	begynne med fulgte vi etter de andre bilene, men snart var vi alene.	190
190	Veiene var smale og humpete. Vi hadde mer enn nok med å holde oss fast bak i	207
207	bilen. Et sted kom vi til en idyllisk liten by helt nede ved vannet. Her skulle vi	224
224	møte en fisker som holdt til ute på en molo. Vi fant ham, og han fortalte at vi var	243
243	den siste familien. Det likte ikke far så han satte opp farten. Mor ropte til ham at	260
260	han skulle ta det rolig. Vi barna syntes bare det var moro.	272
272	Uka gikk fryktelig fort. Vi badet og koste oss. Hver kveld gikk vi ut og spiste. En	288
288	av hovedrettene på øya er kanin. Jeg holdt meg til pizza og pasta. Jeg skulle	304
304	ønske at vi alle sammen kunne dra på lignende turer i fremtiden. Det er fint når	320
320	alle er samlet.	323

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 5 – Spring 3.2

I klatreveggen

0	Det var første gang Tuva var i klatrehallen. Hun var sammen med Liv som	14
14	ønsket en venninne å klatre sammen med. De måtte først gå på et kurs og lære alt	31
31	om sikkerhet, utstyr og teknikk. Da de var ferdige med kurset, fikk de lov til å	47
47	klatre uten at en instruktør var til stede og passet på dem.	59
59	Klatreveggen er nesten 15 meter høy og delt i flere ruter. Noen er for	73
73	nybegynnere som Tuva og Liv. Andre ruter er mye vanskeligere, og du må ha	87
87	klatret mye for å prøve deg på en av dem. Tuva og Liv klatret hver sin gang. Den	105
105	ene sto nede på gulvet og sikret, mens den andre klatret. Så byttet de på. Til å	122
122	begynne med syntes Tuva det var vanskelig. Hun fikk ikke ordentlig tak, og	135
135	fingrene var for svake. Hun manglet dessuten styrke i beina og var også litt redd	150
150	for høyden. Liv var lettere i kroppen, og hun kom høyere opp i veggen. Men	165
165	Tuva bet tennene sammen selv om hun av og til følte seg dum sammenlignet	179
179	med alle de andre.	183
183	Litt om senn ble hun sterkere. Hun kunne holdet taket lenger, og det ble lettere å	199
199	heise seg opp. Da ble det også morsommere, og hun kunne prøve seg på	213
213	vanskeligere ruter.	215
215	Da sommeren kom, ble de med noen andre i klubben på buldring. Det er klatring	230
230	på store steiner ute i naturen. På bakken la de en myk matte som de kunne falle	247
247	på. Tuva og Liv syntes buldring var gøy. De kunne være en gjeng sammen, og de	263
263	lo mye av hverandre når de falt ned på matten. Det var også god trening for å	280
280	lære seg teknikk og kroppsbeherskelse.	285
285	Da skolen de gikk på skulle bygges om, gikk de til rektor og spurte om det var	302
302	mulig å sette opp en klatrevegg i den nye gymsalen. Rektor likte idéen. Da den	317
317	nye gymsalen sto ferdig, var også klatreveggen på plass.	326

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Grade 5 – Spring 3.3

Delfiner

0	En delfin har en veldig kraftig hale som hjelper den når den skal svømme og	15
15	hoppe. Halen går opp og ned som en propell og driver delfinen forover. Den kan	30
30	svømme like fort som en hai og hoppe like høyt som et hus.	43
43	Delfinene lever ofte i store flokker – helt opp til 500 stykker. De kan være gode	58
58	venner med andre delfiner og bruker mye tid sammen med dem. I en stor flokk	73
73	kan det være mange mindre flokker som holder sammen hele livet.	84
84	Delfinene har spisse tenner og spiser fisk, krabbe og reker. De jakter på stimer av	99
99	fisk og følger etter stimen over lange avstander.	107
107	En av fiendene til delfinen er hai. Når delfinen ser en hai, setter den opp farten	123
123	for å komme seg unna. Da svømmer den like fort som en hurtigbåt.	136
136	Delfinene snakker sammen. Ingen vet hva lydene betyr. Den kan plystre og	148
148	klikke. Når lyden treffer en fiskestim, kommer den tilbake til delfinen som et	161
161	ekko. Det kalles for sonar. Da vet delfinen hvor fiskestimen er og kan svømme	175
175	etter den. Den kan også varsle andre delfiner om at en hai nærmer seg.	189
189	Av og til kommer delfiner inn i norske fjorder. De er på evig jakt etter mat og	206
206	kan ha fulgt en sildestim inn til kysten. Så forsvinner de igjen like fort som de	222
222	kom.	223
223	Delfinen kan lære triks som å hoppe høyt og balansere en ball på snuten. Den	238
238	kan også trenes opp til å bringe verktøy ned til dykkere som arbeider på store	253
253	dyp og til å feste bomber på skip under krig.	263
263	Det er mange historier om delfiner som har reddet mennesker fra å drukne.	276
276	Delfinen blir regnet som ett av de klokeste dyrene vi kjenner. Men mange	289
289	delfiner dør av giftstoffer i havet eller de blir fanget i fiskegarn.	301

Totalt antall ord: _____

Antall ord feil (inkl. utelatte ord): _____

Antall ord korrekt: _____

Errata

Place	Original text	Corrected text	Type of correction
Page 15, Paragraph 2, Lines 12-13	... is twice as high in Grades 8 to 10 (10.2%) as in Grades 8 to 10 (10.2%) than in Grades 1 to 4 (5.1%).	... is twice as high in Grades 8 to 10 (10.2%) as in Grades 1 to 4 (5.1%).	Deleted the words *in Grades 8 to 10 (10.2%) than*
Page 17, Paragraph 2, Line 10	... (e.g., Chapter 2.3.3), ...	... (e.g., Chapters 2.2.3, 2.3.3, 2.5), ...	Updated chapter numbers
Page 18, Paragraph 3, Line 2	... include national and tests ...	... include national and other tests ...	Inserted the missing word *other*
Page 20, Paragraph 2, Line 2	... proficiency and identify student difficulties, ...	... proficiency for identifying students' difficulties, ...	Deleted the words *and identify student.* Inserted *for identifying students'*
Page 22, Paragraph 2, Line 3	... those at-risk is, is a highly ...	...those at-risk, is a highly ...	Deleted the word *is*
Page 24, Paragraph 2, Line 11	If Cartwright' assertions ...	If Cartwright's assertions ...	Corrected *Cartwright's*
Page 27, Paragraph 2, Line 7	... was examines in ...	... was examined in ...	Corrected the word *examined*
Page 27, Paragraph 3, Line 3	... the assessments was ...	... the assessment was ...	Corrected the word *assessment*
Page 34, Paragraph 1, Line 6	... National Tests of Reading of Proficiency ...	... national tests and compulsory assessments of reading proficiency ...	Inserted *and compulsory assessments*
Page 36, Paragraph 3, Line 7	The evaluation of the instruments is, In addition to ...	In addition to ...	Deleted *The evaluation of the instruments is,*
Page 39, Paragraph 2, Line 2	... properties, we used multi-statistical ...	... properties, multi-statistical ...	Deleted the words *we used*
Page 43, Paragraph 2, Line 1	... determine how the it fits the ...	... determine how these fit the ...	Deleted *the it fits.* Inserted *these fit*
Page 46, Line 1	... (Appendix B).	... (Appendices C and D).	Deleted *Appendix B.* Inserted *Appendices C and D*
Page 54, Paragraph 2, Line 7	... the scores are based the number ...	... the scores are on based the number ...	Inserted the word *on*
Page 57, Paragraph 3, Line 5	... research and awaits a study with a larger sample.	... research, studies with larger representative samples are awaiting.	Deleted *and awaits.* Inserted *, studies with larger representative samples are awaiting.*

Note. Paper 1 has been published in the Scandinavian Journal of Educational Research. It has the following reference:

Arnesen, A., Braeken, J., Ogden, T., & Melby-Lervåg, M. (2018). Assessing Children's Social Functioning and Reading Proficiency: A Systematic Review of the Quality of Educational Assessment Instruments Used in Norwegian Elementary Schools. *Scandinavian Journal of Educational Research*. DOI: 10.1080/00313831.2017.1420685