

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Ketonen; Laura; Nieminen; Juuso Henrik

Title: Assessment guides, restricts, supports and strangles : Tensions in teachers' conceptions of assessment during an assessment reform

Year: 2024

Version: Published version

Copyright: © 2024 the Authors

Rights: CC BY-NC 4.0

Rights url: <https://creativecommons.org/licenses/by-nc/4.0/>

Please cite the original version:

Ketonen, Laura, Nieminen, Juuso Henrik. (2024). Assessment guides, restricts, supports and strangles : Tensions in teachers' conceptions of assessment during an assessment reform.

Teaching and Teacher Education, 150, Article 104737.

<https://doi.org/10.1016/j.tate.2024.104737>



Research paper

Assessment guides, restricts, supports and strangles: Tensions in teachers' conceptions of assessment during an assessment reform

Laura Ketonen^{a,*}, Juuso Henrik Nieminen^{b,c}

^a Department of Teacher Education, University of Jyväskylä, Jyväskylä, Finland

^b Faculty of Education, University of Hong Kong, Hong Kong Special Administrative Region

^c Mitch and Leslie Frazer Faculty of Education, Ontario Tech University, Oshawa, Canada

ARTICLE INFO

Keywords:

Teacher conceptions
Assessment reform
Tensions
Assessment culture
Low-stakes assessment

ABSTRACT

This study examines tensions in teachers' conceptions of assessment following an assessment reform in Finland, which has traditionally been a low-stakes assessment culture. There are no national examinations in comprehensive education; instead, teachers assign students' final grades. Recently, the country has introduced more detailed criteria for the final assessment to improve the comparability of grades. The analysis of 28 teacher interviews reveals that this relatively minor reform has created four significant tensions in teachers' conceptions of assessment. We call for assessment reforms to be culturally sensitive, carefully considering the context in which they are implemented and paying attention to possible tensions.

1. Introduction

Educational assessment has a double duty in comprehensive education: it should promote student learning and, at the same time, qualify and rank students for societal selection (Boud, 2000). The latter purpose has increasingly been emphasised amid global testing cultures (Smith, 2016). Countries and regions vary in their approaches to organising assessment. Every developed educational system must have ways of sorting and ranking students for the various tiers of society. This should clearly be done in a credible way. In *low-stakes* assessment contexts, testing and comparison do not characterise schooling as strongly as in *high-stakes* contexts. The more high-stakes the assessment culture, the more attention is paid to testing and the more central the objectivity of assessment becomes. Therefore, high-stakes contexts typically rely on national testing as an objective way to handle student selection. In contrast, low-stakes contexts often build on teachers' ability to assess their students truthfully.

Even in countries with low-stakes and teacher-based assessment, there is a struggle between the summative and formative purposes of assessment (Harris & Brown, 2009; Remesal, 2011). Teachers' dual role as *educators* and *assessors* (Harris & Brown, 2009) may be even more tense than in high-stakes countries in which external operators carry out assessment. As teachers are responsible for planning and conducting assessment, the question of *what assessment is and should be* is negotiated

amid these potential tensions that teachers face. Therefore, how teachers balance their understanding of assessment in low-stakes contexts is a crucial question for research.

Finland is an apt example of a low-stakes assessment culture. Students receive final grades at the end of basic education (Grade 9, when students are typically 15 years old), and the grades are used to facilitate selection for upper secondary education. Nonetheless, Finnish society pays comparatively little attention to grades, which is rather unique with respect to international comparisons (Autio, 2021). That said, the international trend towards increased interest in and measurement of student learning outcomes (Teltemann & Jude, 2019) can be recognised in Finnish public discussions. In Finland, teachers have traditionally been able to assess students with the learning purpose of the assessment as the main priority. In Finnish basic education, there are no national high-stakes examinations. Grades have relatively little influence on one's future, and schools are comparatively equal (Kumpulainen & Lankinen, 2012). Moreover, final grades are solely based on teachers' classroom assessment.

In the new millennium, Finnish researchers and policymakers have raised concerns about the inequality of these teacher-formulated final grades (Lappalainen, 2004; Ouakrim-Soivio, 2013). The Finnish National Agency of Education (FNAE, n.d) implemented an assessment reform introducing more detailed assessment criteria to improve the situation (FNAE, 2020a; FNAE, 2020b). These criteria are designed to be

* Corresponding author. PO Box 35, FI-40014, University of Jyväskylä, Finland.

E-mail addresses: laura.ketonen@jyu.fi (L. Ketonen), juuso@hku.hk (J.H. Nieminen).

<https://doi.org/10.1016/j.tate.2024.104737>

Received 18 February 2024; Received in revised form 30 July 2024; Accepted 2 August 2024

Available online 8 August 2024

0742-051X/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

used by teachers as guiding tools, while teachers' final grades remain unmonitored. Despite the reform, Finland firmly remains a country with low-stakes assessment. However, the reform has brought the country's assessment system one – albeit small – step closer to the international norm of precise measurement, even though such measurement is still conducted by autonomous teachers rather than via standardised tests. This study aims to unpack the potential tensions that arise when teachers are required to assign final grades based on new assessment criteria while still maintaining their vast autonomy as assessors.

Assessment reforms are implemented for two main reasons: to drive broader curriculum reforms (Barnes et al., 2000; Torrance, 2011) and to improve the quality of assessment itself. The Finnish basic education reform falls into the latter category. In the 21st century, such reforms have typically aimed to strengthen the role of formative assessment (Darmody et al., 2020; Hopfenbeck et al., 2015; Rosin et al., 2022; Yan & Brown, 2021). The recent Finnish reform is a rare example of this age of reforms, as it places greater emphasis on summative assessment. Although there are lessons to be learned from all assessment reforms, this article focuses on more recent ones. This is because the educational landscape has changed in ways that may contribute to teachers' conceptions. In particular, we suggest that two phenomena have shaped assessment cultures around the world. The first is the growing understanding of formative assessment, which exploded around the turn of the new millennium (Birenbaum et al., 2015), influencing countries' assessment policies (Volante et al., 2024). The second is the increased interest in assessment and accountability and the increasing competitiveness in education, which are encouraged by international large-scale student assessments, such as Programme for International Student Assessment (PISA) (Teltemann & Jude, 2019).

Implementing assessment reforms is not easy. Such reforms aim to change teachers' assessment practices, but the practices will only change if teachers' conceptions of assessment change (Fives & Buehl, 2016). Indeed, teachers' actions mediate the success of any assessment reform (Arrafii, 2022; Darmody et al., 2020), especially in decentralised assessment contexts such as Finland. Therefore, we consider it particularly important to focus on teachers' perspectives, exploring their conceptions of assessment in reform situations (Harris & Brown, 2009).

Teachers' conceptions of assessment reflect both teacher characteristics and environmental characteristics, such as assessment policies (Darmody et al., 2020; Yan & Brown, 2021; Yates & Johnston, 2018). Teachers' conceptions are influenced by the contexts in which they operate, and these contexts contribute to which conceptions are activated (Fives & Buehl, 2012). Conceptions of assessment are influenced by classroom, school and national environments (Fulmer et al., 2015). Besides national traditions and values, international trends also play a role, such as the increased interest in and teacher accountability for student test results (Fulmer et al., 2015; Teltemann & Jude, 2019). Changes in the environment, particularly changes in assessment policies, create tensions in teachers' conceptions and practices of assessment (Arrafii, 2022; Darmody et al., 2020; Yan & Brown, 2021), at least for a while, before a new equilibrium is established (Xu & Brown, 2016). Therefore, exploring teachers' conceptions of assessment in the aftermath of reform is productive for understanding the consequences of reform.

In the present study, our research objective is to analyse teachers' conceptions of assessment shortly after the assessment reform. We are particularly interested in understanding the potential tensions that may arise in teachers' conceptions of assessment when a low-stakes assessment culture introduces an emphasis on students' accurate measurement. As we show, this creates tensions when the seemingly universal and acontextual ideal of measurement meets teachers' daily work, which is deeply embedded in their school contexts (see Smith, 2016). We discuss the broader implications of such tensions for educational systems.

1.1. Teachers' conceptions of assessment

Our study focuses on teachers' perspectives on the recent assessment reform in Finland. In doing so, we take part in the long tradition of scholarly work that has unpacked teachers' viewpoints on assessment by studying their perceptions, beliefs, conceptions, attitudes and so forth (Alonzo et al., 2021; Barnes et al., 2014; Bonner, 2016; Pajares, 1992). It is widely known that teachers' assessment practices do not change following reforms unless their beliefs about the purposes and practices of assessment also change (Fives & Buehl, 2012, 2016; Harris & Brown, 2009). This makes it particularly important to study teachers' beliefs, opinions and conceptions of assessment in the context of reforms. While definitions and operationalisations differ, all these ideas provide a window into the lived world of teachers, enabling us to understand how teachers themselves make sense of the prevailing assessment policies, practices and discourses. This study focuses on teachers' *conceptions of assessment*, broadly conceptualised as “the belief systems that teachers have about the nature and purposes of assessment” (Xu & Brown, 2016, p. 156).

However, it is not only the definitions that differ but also the overall research paradigms behind them. Most commonly, research on teachers' perspectives on assessment has treated conceptions, beliefs and perceptions as *psychological constructs*. In other words, these constructs are seen to reside within the individual, albeit in vivid interaction with the surrounding school community and society (Alonzo et al., 2021; Barnes et al., 2017; Gebril & Brown, 2014; Postareff et al., 2012). The most notable example of this is the quantitative Teacher Conceptions of Assessment (TCoA) instrument, which has been used extensively in various research contexts to uncover teachers' conceptions of assessment (Brown & Gao, 2015; Brown & Remesal, 2012; Darmody et al., 2020; Yates & Johnston, 2018). Moreover, many current models of teacher assessment literacy – namely, teachers' capacity to design and implement assessment – include the idea that teachers should be aware of their own conceptions of assessment (DeLuca et al., 2016). An apt example is Xu and Brown's (2016) model of teacher assessment literacy, which includes the dimension of *conceptions of assessment* as an *individualistic* construct consisting of cognitive and affective dimensions.

By contrast, we take a *collectivist* approach to teachers' conceptions of assessment by noting that while they are “influenced by policy priorities in specific socio-cultural and institutional contexts” (Xu & Brown, 2016, p. 156), they are formed through broader public and societal discourses about assessment and are “subject to an individual's personal and educational experiences” (Xu & Brown, 2016, p. 156). As such, teachers' conceptions of assessment are constituted and shaped by teachers' collective stories about assessment (e.g., Harris & Brown, 2009).

Our study is particularly interested in the *tensions* within teachers' conceptions of assessment. It is widely recognised that teachers experience tensions as they balance the accreditation and learning purposes of assessment (Harlen, 2005; Harris & Brown, 2009; Whittaker & Young, 2002; Winstone & Carless, 2021). Bonner (2016) further highlights the tensions, stating that “the story of teacher perceptions about assessment is no longer a story of confusion; it is a story of tension and conflict” (p. 21). As Chan and Tan (2022) argue, tensions in teachers' understandings of assessment are not always problematic or undesirable (p. 448). Their study shows that tensions do not always hinder teachers' formative assessment practices but can indeed support them. Similarly, Govaerts et al. (2019) call for moving beyond *either-or* thinking in assessment. This type of thinking might include, for example, emphasising *either* the learning purpose *or* the accreditation purpose of assessment. Instead, the researchers urge teachers to embrace dilemmas and tensions in assessment by adopting a *both-and* approach to assessment design.

In this study, we rely on a qualitative and interpretivist approach to teachers' beliefs to capture teachers' stories about potential tensions in assessment amid an assessment reform. More specifically, we use the specific framework of Remesal (2011) to understand teachers'

conceptions of assessment. Many other studies have identified tensions, but usually with a focus on tensions *between* individuals or *between* groups (see Bonner, 2016). Remesal's (2011) framework is particularly compelling due to its capacity to unpack internal tensions within the same teacher group (Barnes et al., 2014). Moreover, it connects assessment to both school and society, which is relevant in the case of a reform. According to Remesal's (2011) framework, teachers' conceptions of assessment consist of four categories: learning, teaching, certification and accountability. Within these categories, assessment beliefs can be positioned on a bipolar continuum. On one end of the continuum, assessment is viewed as having a pedagogical function. On the other end, assessment is viewed as having a societal function. Teachers' conceptions of assessment are composites of their beliefs across the four categories. This framework is further demonstrated in this article's method section through concrete examples of individual beliefs (readers may also refer to Remesal, 2011).

1.2. High-stakes vs low-stakes assessment cultures

The terms *low-stakes* and *high-stakes* assessment are often used but seldom adequately defined. We explain our understanding of the concepts before describing the Finnish assessment culture, which we claim to be low-stakes.

A test is considered a high-stakes test if its results have a direct and typically long-lasting impact on students, teachers or schools. The primary use of high-stakes assessment is student selection, placement and certification, but other common uses include accountability and setting standards (Stobart & Eggen, 2014). Low-stakes tests do not have such direct effects.

The concept of a high-stakes assessment culture is more complex than that of testing because it includes structural and cultural features. Assessment cultures cannot be divided into high-stakes and low-stakes cultures; rather, these ideas should be understood as a continuum (Chen & Teo, 2020). The structural features relate to equity between schools and the organisation of student selection. The stakes of assessment increase when there are significant differences between schools or classes or between later educational institutions whose selection is based on prior assessment results. Moreover, the stakes increase when educational pathways contain bottlenecks or dead-ends that students must avoid.

An assessment culture becomes more *high-stakes* the more socially accepted assessment is and the more assessments and their results are perceived as legitimate and relevant (Moses & Nanna, 2007). In addition, the more objectively assessment results are considered to reflect students' learning and learning potential, the more high-stakes the assessment culture. In high-stakes assessment cultures, assessment results are seen as reflecting the quality of not only students but also teaching, teachers and schools, and they are used for accountability purposes.

Classroom assessment, or teacher-based assessment, refers to assessment that is planned, implemented and analysed by teachers and used for both summative and formative purposes (Black & William, 2018; McMillan, 2013). In high-stakes contexts, teacher-based summative assessment tends to be questioned due to the potential inconsistency of teachers' judgements (Black & William, 2018). However, while there is more human influence in teacher-based assessment than in standardised tests, teacher-based grades can be a good measure of student achievement (Aléman et al., 2024). They have certain strengths compared to standardised tests. For example, teacher-based grades consist of multiple inferences and can consider a broader range of aspects than standardised tests (Stanley et al., 2009).

1.3. The assessment reform in Finland

Compared to most countries, Finland has a low-stakes assessment culture. In Finnish basic education, assessment is based entirely on

teachers' classroom assessment. The Finnish National Core Curriculum (FNCC) outlines the learning and assessment objectives, but it is up to teachers to decide how to assess them. Teachers have a high degree of autonomy in organising teaching and assessment, and there are no high-stakes examinations (Kumpulainen & Lankinen, 2012). Assessment can be based on various forms of evidence, such as observations, project work, portfolios, examinations, texts and audio and video documents. Examinations are not mandatory, but they are a common practice, as they have historically been (Atjonen et al., 2019; Pulkkinen et al., 2024). Broadly speaking, Finnish teachers have used their autonomy to maintain their assessment practices rather than to transform them. However, according to the FNCC (FNAE, 2020a), teachers must use various forms of evidence to assess students. Furthermore, it is stated that the assessment of all the objectives in any school subject cannot be based on only one type of evidence (e.g. examinations).

According to the FNCC (FNAE, 2020a), students must receive grades, starting from Grade 4 at the latest. Regional variations exist, and grades may be assigned earlier but not later. Grades are assigned in all disciplines except guidance counselling, and the grading scale ranges from 4 (failed) to 10 (excellent). Selection for upper secondary education after basic education (Grade 9) is based on students' final grades, which are determined by teachers' classroom assessment. At this point, classroom assessment has tangible consequences for students' lives.

Admission to upper secondary schools is based on students' grade point average for theoretical subjects. For vocational schools, it is based on the grade point average for all subjects. However, for contextual reasons, the stakes of grades are low compared to other countries. The Finnish educational system is relatively non-competitive (Autio, 2021). It mainly consists of one public system, with only a few private schools. In basic education, students typically go to the assigned local school. General upper secondary schools are equal in terms of the quality of teaching. If students' grades are not high enough for the education of their choice, be it vocational or upper secondary, they can attend preparatory education and reapply. What further reduces the stakes of final grades is the recent extension of compulsory education (Ministry of Education and Culture, 2021), which guarantees every student a secondary qualification. Not having bottlenecks in education is in line with equity, which is a central value of Finnish basic education (Niemi, 2012) and, hence, of assessment.

Previous research has shown that final grades are not comparable (e.g. Ouakrim-Soivio, 2013), and these findings have driven the recent assessment reforms. Students with the same level of skills and knowledge have been shown to receive grades that differ by up to two grades (Lappalainen, 2004; Ouakrim-Soivio, 2013). Before and during the reform, there was broad discussion in the media about the incomparability of grades, and the Finnish National Agency of Education (FNAE) emphasised comparable grading in its communication concerning the reform.

One reason for this discrepancy can be identified in the two previous curricula (FNAE 2004; FNAE, 2014), which describe the assessment criterion for good skills only (Grade 8). With just one criterion, it is not even theoretically possible to provide reliable grades across the whole grading scale. The reliability of assessment during the previous curriculum (FNAE, 2004) was further challenged by the weak link between learning objectives and final assessment criteria. This did not guide teachers to include all learning objectives in their assessment. The FNCC 2014 tackled this issue by presenting a definitive list of learning objectives for each school subject and describing Grade-8 criteria for each learning objective. Overall, the FNCC 2014 paid more attention to the implementation of assessment. The chapter describing the organisation of assessment (FNAE, 2014) was longer than for previous curricula (FNAE, 2004), providing more detailed information on the implementation of formative and summative assessment. However, despite its length, the chapter on assessment led to some misinterpretations among school leaders and teachers. It was thus rewritten soon after its implementation. In 2020, the FNAE introduced a new chapter on assessment

(FNAE, 2020a).

The core of the reform, and the focus of this article, was the introduction of more detailed assessment criteria. At the end of 2020, criteria for Grades 5, 7 and 9 were released to supplement the criteria for Grade 8 (FNAE, 2020b). The purpose of the new criteria was to improve the comparability of final grades (FNAE, nd). Teachers have been reported to appreciate the reform at a general level but to struggle with the details (Ketonen, 2024; Pulkkinen et al., 2024). The criteria were not referred to as a major change but rather as a tool for calibrating assessment. These criteria were used for the first time in the final assessment at the end of the 2021–2022 school year in May 2022. At the time of the interviews, the criteria had been used once.

1.4. Study aim

The purpose of this study is to investigate potential tensions in teachers' conceptions of assessment following the recent assessment reform in Finland. We propose the following research question: As the assessment system moved one step closer to the precise and transparent measurement of student learning, what kind of tensions have arisen between teachers' conceptions of assessment?

2. Methods

The data for this study were collected as part of the nationwide research project PARVI from December 2022 to April 2023. The Finnish Ministry of Culture and Education ordered the research project to provide information on reforming the final assessment in basic education. The research team submitted a report to the ministry in January 2024 (Pulkkinen et al., 2024), and the data were subsequently made available for further study.

2.1. Participants and data

The participants were Finnish lower secondary school mathematics ($n = 15$) and physics ($n = 13$) teachers. Most of them taught both physics and mathematics, but they were interviewed from the perspective of only one discipline. We intentionally chose these disciplines because these teacher groups share similar conceptions of knowledge and teaching, which was deemed appropriate for studying their conceptions. In the PARVI project, a nationally representative sample of schools was selected to answer a questionnaire, and all teachers were given an opportunity to participate in the interviews. Later, the principals of Finnish lower secondary schools were asked to inform their teachers about the possibility of participating in the interviews to increase participation. The final participants were teachers from various regions of the country with diverse career paths and teaching experience.

The data consisted of semi-structured interviews that were transcribed before the data analysis. In the first part of the interview, teachers were asked about their general conceptions of assessment and the reform in general, their experiences with implementing the reform at their school, and the impact of the reform on their work. After the general part, discipline-specific questions were asked, using the assessment criteria of the discipline to support the discussion. Teachers were asked about the use and applicability of the criteria and the interpretation of certain principles, such as weighting of versatile learning objectives and assessment of working skills. They were also given the opportunity to provide feedback and comments on the criteria. Finally, they were asked to explain how they would like to see assessment policies and practices be developed. The interviews were conducted through online video meetings. The recordings began after a brief introduction to the purpose of the study and assurances that teachers had given their consent. The recordings varied between 25 and 87 min, averaging 51 min.

2.2. Analysis

The analysis was theory driven. We used Remesal's (2011) framework to define the concepts of assessment conception and assessment belief. The first researcher began the analysis by carefully reading the transcriptions and identifying the units of analysis in which teachers expressed assessment beliefs. Each unit of analysis contained one meaningful and coherent idea about assessment. The researcher had conducted the interviews and was familiar with the data. The identified extracts were typically long (average length 1193 characters), as contextual information was needed to understand the beliefs. After identifying all the units of analysis, the first researcher assigned them to different categories of Remesal's (2011) framework based on which dimension(s) of assessment the belief related to (1–4; see Table 1) and whether it concerned pedagogical aspects of assessment, societal aspects of assessment, or both (A, B, AB).

Table 1 contains examples of categorisations from Remesal's (2011) study, along with additional data-based examples in italics from the specific context of the Finnish assessment reform. The categorisations of extracts were discussed with a colleague from the research project who had categorised a subset of the dataset in order to test and extend the initial understanding of teachers' thinking. In comparing the categorisations and discussing the differences, we noticed that teachers simultaneously held multiple, often contradictory beliefs. This was not surprising based on previous research (Brown, 2008; Fives & Buehl, 2016), but it made it impossible to code an individual extract as a single category.

Next, the first researcher summarised each unit of analysis – consisting of raw data – into a short description, following the guidelines of in vivo coding in that each unit of analysis was assigned a summary statement by using teachers' own phrases whenever appropriate. She then pasted these descriptions onto digital post-it notes and began to form groups of similar beliefs. This grouping process was data-driven and iterative, informed by Remesal's (2011) framework (see Table 1). As a result, each group reflected a particular category of the framework. Although the individual units of analysis often contained conflicting beliefs, the groups provided a more coherent perspective that could be assigned to a particular category of Remesal's framework. After grouping, each group was named.

Finally, the first researcher identified the tensions *within* and *between* the created groupings. She started by looking for tensions within each unit of analysis and marked these post-it notes with an asterisk. An example of a unit of analysis with an internal tension was "*There will always be human reasoning involved in assessment, and there should be. It [assessment] should not be only about comparability. It is about learning first.*" The author then systematically reviewed all the groups created and identified any tensions between them. Whenever she identified a tension between the groups, she marked it with an arrow. An example of groups in conflict was "*Objectives are appropriate to the nature of the discipline*" vs "*Current objectives are difficult to assess.*" After identifying all the tensions, the researcher collected all the conflicting post-it notes and group names with tensions on a new digital board and formed groups based on the tensions. As a result, four central tensions were constructed between teachers' various conceptions of assessment. Three of the tensions were between two poles of the same assessment dimension, and one was between two different dimensions.

3. Results

In this study, we examined the tensions in teachers' conceptions of assessment following the recent reform that emphasised the measuring purpose of assessment by introducing more detailed final assessment criteria. In our analysis, we found four tensions between eight conceptions of assessment. It is worth noting that all the tensions were between societal and pedagogic functions of assessment. The conceptions and tensions are presented in Fig. 1.

Table 1
Remesal's (2011) framework of teacher conceptions, modified for our analysis. The italics are data-based additions to Remesal's framework in the particular context of the Finnish assessment reform.

	A) Pedagogic regulation	B) Societal accreditation
1. Learning (Assessment has consequences for learning)	The student must be active in the design and development processes concerning assessment. Assessment can provide opportunities for learning new content. Assessment provides opportunities to improve learning-to-learn skills. Students learn by identifying and correcting their own mistakes. <i>The student must be an active participant in assessment. Students learn by using the criteria. The criteria advance students' understanding of the quality of their work.</i>	Assessment informs students about the level of their achievement at the end of learning. Assessment motivates students and makes them increase their study efforts. Knowledge about one's level of learning increases one's study effort. Assessment does not promote learning but has a negative influence on it.
2. Teaching (Assessment has consequences for teaching)	Assessment needs a substantial change in order to meet students' diverse educational needs. Assessment can be used to adjust teaching (immediately, short term and long term). Assessment can be used to adjust teaching in the school (beyond the single classroom). Assessment can be used to make methodological changes only in the long term or for future courses. <i>Assessment must not guide the teacher's work too vigorously. Assessment must be aligned with the nature of the discipline. Assessment can be used for differentiation of learning.</i>	Assessment disturbs the teaching rhythm (e.g. teachers need to slow down when students do not follow or speed up for examinations). <i>Teachers must change the rhythm of teaching because of assessment (this is considered neutral, not harmful).</i> If students' achievement is unsatisfactory, the teaching practices must be repeated by reviewing the learning content a second time. Only school-level changes are sensible (e.g. grouping students and supporting teachers). Assessment does not influence teaching, or the influence is negative (e.g. loss of time and stress). <i>National and local assessment guidelines guide teaching. Assessment has contradictory influences on teaching.</i>
3. Certification (Assessment has consequences for the certification of learning)	Assessment hardly serves as a selective instrument because it does not allow comparison of results. <i>Assessment is not a selective instrument because students' true learning outcomes are challenging to assess.</i> The student's individual progression must be taken as an essential reference criterion (or considered one factor) for accreditation purposes and for undertaking changes in teaching. Assessment must not be norm-based. <i>Students should not be</i>	Quantitative grading of achievement is more important than qualitative assessment of the learning process. <i>More systematic ways are needed to arrange student selection (e.g. national tests).</i> The accreditation (final assessment) must be norm-based. Results of assessment performance must be expressed in numerical or categorical grades. <i>Comparing students to each other is helpful.</i> If a student fails, the teacher should lower the

Table 1 (continued)

	A) Pedagogic regulation	B) Societal accreditation
	<i>compared to each other in assessment. Letting students' self-assessment influence the grades would be a good idea. Qualitative learning advances are more important than quantitative advances.</i>	demands or repeat the testing, forgetting about the first unsatisfactory results. Assessment is an indispensable (or useful) tool for social control and student selection, for differentiation and for determining the accepted performance level. External evaluation is accepted for comparison purposes or agreed upon, but without further interest. Communication with different audiences must be unidirectional if the teacher is to keep their face. <i>The teacher knows best about assessment. Families must be informed about their children's results, but merely as passive recipients of information. The teacher's assessment decisions (especially grades) must be justifiable in front of different audiences (students, colleagues, and school leadership) Grades must be predictable: they should not surprise the student or the family.</i>
4. Accountability (Assessment has consequences for the accountability of teaching)	Assessment enables engagement in a whole systemic change beyond the individual classroom and school. Families must be informed about their children's learning to promote collaboration between the school and the family. <i>Collaboration and two-way communication with families in the assessment of a student's learning is essential.</i> Communication about assessment must be multi-directional with different audiences, including students themselves. External comparative evaluation studies must be rejected if there is no connection with the particular context. <i>Teachers are not accountable for upper secondary schools and their teachers.</i>	

3.1. Tension 1: assessment supports and obstructs learning

The first tension in teachers' conceptions of assessment concerned the dimension of learning. Teachers saw the pedagogical and societal purposes of assessment as competing. They felt that the measurement of student learning had intensified not only as a result of the increased emphasis on equality in assessment but also due to several other features of the curriculum: long lists of learning objectives to be assessed, the requirement to collect diverse evidence of student performance and the requirement for transparency in assessment. These were experienced as leading to continuous evidence gathering, forcing teachers to constantly inform students of the purpose and practices of assessment as well as students' progress. Teachers expressed a desire to focus more on teaching and less on assessment. Some felt that the intensified measurement of learning undermined students' persistence and joy of learning. They saw that this was particularly true for low-achieving students for whom it was difficult to provide both encouraging and realistic assessments. Teachers reported that the societal purpose of assessment led students to focus on grades rather than learning. This was seen to lead to undesirable consequences, such as rote learning, cheating and reduced student well-being, as exemplified by the following extract:

"I think that assessment increases stress. Students now see school as a more stressful place. They no longer go to school because they enjoy learning. The assessment criteria are related to this because most teachers believe that they have to document meticulously what the grade consists of. And this means that they now have to use various assessment tasks, at least one task for each learning objective. This

	A. Pedagogic		B. Societal
1. Learning	Assessment for learning is important.	Tension 1 	Assessment of learning is necessary.
2. Teaching	Teacher autonomy improves the quality of teaching and assessment. Quality of teaching matters when selecting learning objectives.	Tension 2 	Strict assessment guidelines improve the quality of teaching and assessment.
3. Certification		Tension 3 	Comparability of assessment matters when selecting learning objectives.
4. Accountability	The transparency of assessment is fair and productive.	Tension 4 	Despite transparent criteria, the summative assessment is teachers' business.

Fig. 1. The tensions in teachers' conceptions of assessment are categorised according to Remesal's (2011) framework.

means an increased number of assessment tasks for students in all subjects." (T24, 3 years' experience, physics¹)

At the same time, teachers understood the need to measure student learning, even though they were less enthusiastic about it. The following extract illustrates the tensions between the societal and pedagogical purposes of assessment relating to learning:

"I do not see [summative] assessment as the most meaningful part of the teacher's job. I would rather teach and interact with students. I understand what it is for, but I do not see it as the most important task of basic education. Assessment is particularly challenging with low-performing students. I feel that assessment is unfair and unworkable for them. But of course, I understand why we have the final assessment, because everyone needs to be measured somehow." (T18, 8 years' experience, mathematics)

The teacher describes how summative assessment feels like an unmotivating and sometimes even unfair practice, but how they also understand it as justified and necessary for societal purposes.

Teachers generally described the conflict between formative and summative assessment as a permanent one to be endured. Many expressed that basic education had intrinsic value and that providing skills for further education or grades for student selection was secondary. Some teachers hoped for solutions at the administrative level and went so far as arguing that basic education should reject the societal task of sorting students:

"If I could ask one thing of society, it would be to stop striving for equality of assessment. It does not necessarily have to be equal. The goals of basic education are different. In fact, the law does not mention the comparability of assessment. I think that the instructions of the Agency for Education contradict the law. The law mentions three purposes for assessment, and comparability is not one of them. These three purposes are encouragement, guidance and tools for self-assessment. Assessment should be developed from the perspective of the law. How people are selected for schools is a business of upper secondary education, and basic education should not accept it." (T10, 12 years' experience, physics)

The teacher refers to the Basic Education Act to argue that the

societal purpose of assessment is not the business of basic education and suggests that the task could be transferred to upper secondary education.

3.2. Tension 2: assessment guides and strangles teaching

The second tension concerned teaching and its pedagogic and societal aspects. The tension existed between two main conceptions: appreciation of teachers' autonomy and appreciation of students' equal treatment. By equal treatment, teachers meant not only equal grading but also equal teaching. They felt that the new assessment guidelines drove pedagogy and, therefore, standardised teaching. According to this conception, the change was positive and rational. The following are examples of teachers' comments:

"I guess you asked how I see assessment. I see it as a good professional counsellor." (T7, 12 years' experience, physics)

"It [the reform] has changed my assessment practices so that I now pay attention to some learning objectives in the curriculum that I had not noticed before. I realised that I had not taught some of them at all. There were others that I had not assessed. I try to direct my teaching so that such objectives are included in my teaching and maybe even in assessment. That has been the principal change. I think the reform has diversified my teaching." (T9, 15 years' experience, physics)

In both of the above quotes, the teacher views the guidance provided by the assessment guidelines as a positive thing. In the latter quote, the teacher explains that the assessment reform has led them to diversify both teaching and assessment to better align with the curriculum.

At the same time, many teachers emphasised the negative aspects of the new guidelines. They highly valued teachers' autonomy in teaching and saw the increased guidance through assessment as harmful. One such conception is shown in the following quote:

"Personally, I did not like the criteria reform. I thought it was wrong! I liked the previous model with only the Level-Eight criteria. Back then, teachers were given more freedom. Now, we are moving towards a more patronising direction. That's how I see it." (T21, 15 years' experience, mathematics)

The quote illustrates that negative views were not only related to concrete issues, but also to the sense of unnecessary control. However, concrete problems were also mentioned. Some teachers commented that consideration of differentiation in terms of students' individual learning

¹ Discipline refers to the perspective of the interview, not the qualification of the teacher.

needs was challenged because of the requirement for equal assessment. One commented as follows:

“I think it [the reform] has added to teacher’s workload because everything has to be transparent and marked in Wilma [the learning management system used in Finland] so that teachers can justify any grade at any time. Teachers work too hard with their exams. Teaching is guided way too much by assessment. ... It [the reform] limits my autonomy as a professional.” (T25, over 20 years’ experience, mathematics)

The teacher discusses the workload and control resulting from the need to assess and report more than before. They perceive this as limiting their ability to act autonomously and use their professional judgement.

Many teachers held these conflicting conceptions simultaneously. They saw the value of equal treatment of students and the freedom of teachers to adapt their teaching to the circumstances:

“For the assessment to be equal, meaning that it is the same for everyone, the criteria should be more tangible. But on the other hand, it is good to have the freedom to adapt. It’s a contradictory case.” (T4, 7 years’ experience, physics)

In the quote, the teacher is balancing two values: teacher autonomy and equal treatment of students in assessment. They see that both cannot be achieved at the same time.

As an extra twist, teachers recognised that despite the new guidelines, their autonomy remained high; therefore, the current assessment guidelines did not guide all teachers:

“No matter what the criteria are, a fogley can remain a fogley and say, ‘this is how I have always done it, and this is how I will always do it.’ On the other hand, I appreciate autonomy in my work so that no one tells me how to do it. I am not ready to give that up. One solution would be force, such as high-stakes tests that force everyone to teach the same stuff. But I don’t want that either. It’s a tough question.” (T4, 7 years’ experience, physics)

The teacher points out that the criteria do not always provide guidance because teachers can decide not to follow them. Although the teacher sees this as a problem, they are hesitant to increase control because they value teacher autonomy.

3.3. Tension 3: diverse learning objectives support the validity of teaching but conflict with the reliability of assessment

The third tension in teachers’ conceptions of assessment was between the pedagogic pole of teaching and the societal pole of certification. The current assessment reform had introduced numbered lists of learning objectives in each discipline and linked them more closely to assessment by introducing assessment criteria for each objective. In terms of teaching, teachers saw the current learning objectives as consistent with the nature of the discipline. Traditionally, assessment has focused on *content*, but the recent reform made it explicit that *competences* must also be targeted in assessment. Teachers expressed their appreciation of the alignment of assessment with the nature of the discipline. In the following extract, one teacher described a change in mindset brought about by the reform:

“I do not think that assessment should focus more on content. It took me a while to internalise the new curriculum... like what the guiding idea was. The point is precisely that we teach, for example, the skills and knowledge of science by covering certain content. The content is a tool for learning these wider skills. So, no. But I can understand the criticism. If you have the old way of thinking that you must focus on content... You need to twist your brain to get it.” (T3, 16 years’ experience, physics)

The conflict arose from the requirement for comparable grades,

which was given greater emphasis in the reform. Teachers did not believe it was possible to reliably assess diverse and complex learning objectives that endorsed the nature of the discipline. The following extract showed that within this tension too, the conflicting concepts were held simultaneously:

“When planning the [national assessment] criteria, they should have thought very carefully about which objectives are even possible to assess comparably. Or they should have accepted that it is not possible. Personally, I have accepted that it is impossible to assess those transferable or working skills and that it is practically impossible to measure those on any reasonable scale. If you want to develop the system somehow, you should at least be aware of this challenge. I cannot provide any real advice for this situation because the learning objectives themselves are good. But if the system forces you to assess those objectives reliably, they [criteria] no longer work.” (T10, 12 years’ experience, physics)

The teacher recognises the value of having a wide range of learning objectives and reliable assessment but sees achieving both simultaneously as impossible. They hope that this tension would be acknowledged by authorities responsible for the reform.

Despite their beliefs regarding how this tension should be resolved, teachers felt that the demands of the reform were conflicting. They saw that if the aim were to improve the comparability of grades, more measurable learning objectives would be necessary. Alternatively, if the aim was to improve the validity of grades, it should be accepted that grades may never be comparable.

3.4. Tension 4: transparency of assessment promotes student learning but challenges teacher authority

The fourth tension concerned the accountability of assessment and its pedagogic and societal aspects. The recent reform had increased the demand for transparency in assessment: Teachers were supposed to introduce the learning objectives to students at the beginning of the learning process, inform them of the evidence on which their summative assessment was based and provide formative feedback on students’ progress towards each objective. The numbered list of learning objectives and associated assessment criteria allowed and compelled teachers to provide students with more detailed information than before. Teachers considered the requirements of transparency to be rational and fair. They also used the criteria as a pedagogical tool to guide students’ learning in several ways. For example, they motivated students with the criteria and used them in self-assessment tasks.

However, transparency combined with a greater emphasis on the measuring purpose of assessment created a tension. Teachers saw that there had been a societal shift in which parents had become overly interested in grades, as described in the following extract.

“It [the thought of having to justify a grade] is distressing at this point in the year, but even more so when you are tired in the last few weeks of the autumn or spring term and when you know that you have students with difficult parents. At its worst, it can rob you of your sleep. Maybe this is more of a societal change. It used to be that the teacher gave a grade and there was not much discussion about it, at least not about the details.” (T2, 12 years’ experience, physics)

The teacher describes their experience of parents increasingly questioning grades, noting that this has become a major source of stress for teachers. They attribute this phenomenon to a societal change. Experiences of parents trying to influence teachers’ judgement were common. Most teachers described either having personal experience of serious conflicts with parents or having a colleague that had. Teachers saw transparency as giving parents tools to participate in grading or, perhaps more accurately, giving parents the impression that they had enough knowledge to participate. The criteria were seen not only as a stimulus for the problem but also as a solution, providing teachers with a

backbone to justify their grades. Teachers also saw parental involvement as a positive development. However, they wished parents would be involved in formative rather than summative assessment, as a teacher described in the following quote:

“Let me draw a caricature of a parent involved in a student’s assessment. They have no interest in the seventh grade and no interest in the eighth grade. In the ninth grade, they start asking ‘on what grounds,’ ‘how’ and so on. It would have been productive in seventh and eighth grade to ask, ‘Can I do something for my child? Could I look after the homework?’ If the parent is involved, we can make a difference. But if the evidence is already given when the parent starts digging into the criteria... Okay, they may even force [the teacher to give] a better grade for their child, but the child has not learned anything.” (T1, 18 years’ experience, physics)

In the quote, the teacher describes a typical situation where parents are more interested in their child’s grades than in their learning process. The teacher sees learning as more important than grades and hopes that parents would share this view.

4. Discussion

In the present study, we examined the tensions in teachers’ conceptions of assessment shortly after implementing an assessment reform that provided teachers with more definite assessment criteria. We identified four central tensions between teachers’ conceptions of assessment.

The first tension occurred between assessment *of* and assessment *for* learning. The requirement for greater comparability of grades evoked this tension. Teachers did not see assessment as irrelevant per se, as described in Brown’s (2004) study. Rather, the competition between the purposes of assessment led them to express negative views about different aspects of assessment. This tension has, of course, been widely discussed in the assessment literature (Harris & Brown, 2009; Remesal, 2011; Winstone & Carless, 2021). It could be called the mother of all tensions in assessment! However, the finding that solving this tension by ignoring the societal function of assessment was relatively popular among secondary school teachers seems unique in terms of international comparison (cf. Brown et al., 2011; Remesal, 2011). We suggest that it is due to the non-competitive context of Finnish basic education, which sees education as valuable in its own right rather than as an instrument for accessing further education (Autio, 2021).

The second tension concerned more directive assessment guidelines, especially criteria that are seen as not only driving and inspiring teaching but also strangling teachers’ pedagogical choices. Teachers saw advantages in a high degree of teacher autonomy in assessment but called for tighter external control. So far, Finnish teachers’ broad autonomy in assessment has not led to diverse practices but instead to a widespread use of written examinations (Atjonen et al., 2019; Pulkkinen et al., 2024). In this study, teachers opined that new, stricter guidelines in assessment had led them to use more diverse and innovative forms of assessment. However, in line with Daugherty et al.’s findings (Daugherty et al., 2008), teachers felt that the requirement for equal assessment limited their possibilities for supporting individual students’ learning.

The third concern regarded learning objectives that were considered functional for teaching but not necessarily for comparable assessment, and vice versa. This tension relates to the conceptualisation of assessment as a measurement practice, which raises questions of validity and reliability. Teachers acknowledged the value of complex, discipline-endorsing learning objectives, such as research skills in physics, but they recognised that their accurate assessment was difficult, if not impossible. They felt torn between the need to deliver comparable grades and the need to include all objectives in grade formation. Ultimately, this tension digs deep into educational and societal values (Biesta, 2009). As the assessment culture moves towards more accurate measurement, greater emphasis must be placed on the reliability of

assessment, which in turn requires more measurable learning objectives. What is unique about this tension is its nuance. Although teachers considered this tension unsettling, they claimed to be relatively comfortable with either pole of this tension: improving comparability or giving it up. This may reflect two current convincing but contradictory narratives: the international idea of comparability and competition in assessment and the historical assessment context of Finnish basic education.

The fourth tension concerned transparency in assessment, which was generally seen as fair and supportive of student learning. However, along with the growing emphasis on grades, parents were described as using transparency for detrimental purposes. This tension is reminiscent of the first tension, as some parents were seen to be more interested in children’s grades than in their actual learning. Traditionally, Finnish parents have largely seen assessment as a tool for learning rather than as a mechanism for selection and competition (Nieminen et al., 2021). However, there is reason to suspect that the assessment reform itself has given rise to a growing if still comparatively low interest in grading among parents.

In the low-stakes assessment context of Sweden, there has been similar speculation that a change in the grading system has led to grade inflation through increased competition between students (Wikström, 2005), increasing pressure on teachers to give high grades. At the same time, the grown interest in grades and perhaps the Finnish reform itself, are more symptoms than causes of an evaluation society in which assessments are receiving increasing attention in education and all other sectors (Dahler-Larsen, 2012; Teltemann & Jude, 2019). Torrance (2011) argues that assessment reforms are not only consequences of progress but also reflect “developments in the social and economic aspirations which we hold for the education system, and thus what it is that we are trying to design assessment to accomplish” (p. 460).

What is noteworthy about the findings is that a seemingly small and practical reform has contributed to teachers’ assessment practices and societal discourses in unintended ways. The reform has provoked discussions about validity, measurability and accountability in education, which are not mainstream issues among teachers in Finland. Teacher accountability is seldom touched upon, as teachers have a high degree of autonomy, which is usually taken for granted (Hwa, 2022). Neither validity nor measurability has been a common theme, partly because of teachers’ routine assessment practices and partly because of the context, in which grades do not monopolise students’ life choices. Apparently, only a minor shift towards high-stakes assessment activated beliefs about these issues (Fives & Buehl, 2012). This underscores how even minor alterations to assessment policies can cause considerable tensions in teachers’ conceptions.

It is important to note that the *existence* of tensions does not indicate that an assessment policy is unsuccessful, as dealing with tensions is an integral part of teachers’ assessment work (Remesal, 2011; Xu & Brown, 2016). Moreover, if the goal of an assessment reform were solely to eliminate tensions, this could be done. For example, the tension between summative and formative assessment could be eliminated by abandoning summative assessment and using random choice for student selection. However, tensions need to be processed and balanced, and the reform’s framing as small and practical may have been uninviting to such a process and thus made the tensions more confusing.

Prior research has generally seen the resolution of assessment tensions as the responsibility of individual teachers or schools (Harris & Brown, 2009; Meijer et al., 2023; Xu & Brown, 2016). We present a contrasting view. Reforms are necessary, and tensions are inevitable, but simply implementing reforms and leaving teachers and schools to deal with the consequences is not sustainable. It has been demonstrated, for example, that the tension between assessment of and assessment for learning is shaped and intensified by countries’ assessment policies (Volante et al., 2024). Consequently, the point of our article is not to call for individualistic solutions, such as more or better teacher training for developing teachers’ assessment literacies. Instead, we call for *culturally*

sensitive assessment reforms. This is because it is perhaps impossible to have it all: accurate, comparable, valid summative assessment and intensive formative assessment that promote learning and support student agency (Torrance, 2011). When reforms attempt to pursue too much, the tensions can become unworkable for teachers (Deneen & Brown, 2016; Yan & Brown, 2021). Suppose a reform is intended to move the assessment culture towards more accurate grading. In that case, stakeholders must consider its influences on the low-stakes features that the culture already has and reflect on the extent to which these can be maintained after the reform. The same rings true the other way around. In moving towards a low-stakes culture, giving up some of the high-stakes elements may be necessary to avoid unworkable tensions.

Culturally sensitive assessment reforms would ideally consider the educational context's socio-cultural and -historical dimensions. This calls for an in-depth understanding of the history, goals and ideologies of education in a particular national context (e.g. Cooper & Jordan, 2003). Our study has described the tensions that arise in teachers' understandings of assessment at a time of reform based on the idea of assessment as measurement (albeit, globally speaking, on a small scale). By its very nature, this idea of measurement is seemingly acontextual and universal, but it reflects global and international accountability cultures in educational assessment (Smith, 2016; Teltemann & Jude, 2019).

The study findings show that even in the low-stakes Finnish context, conceptions of the need for accurate measurement are familiar and relatable to teachers. Tensions arise when teachers are asked to implement these acontextual ideals of assessment in their own lived, embodied classroom contexts. In Finland, teachers have historically been seen as legitimate assessors within the relatively low-stakes context. In many ways, the educational system has allowed incomparable grades to exist, which has been possible in a system with little competition between students. Consequently, equity has always been a core value in Finnish basic education (Autio, 2021; Niemi, 2012). As the principal aim of the reform was to categorise students at the end of their learning, this may have created tensions with the idea of equity. The tensions might have been smaller if the rationale for introducing assessment criteria had been, for example, to illustrate the paths of learning in different skills or to better diagnose and support students at risk.

5. Limitations

Our study has its limitations. One is the limitation to one cultural context and one assessment reform. More research is needed to examine the generalisability of the finding that a shift in assessment culture in one direction requires teachers to make compromises in the other direction. A close examination of tensions following reforms might reveal the extent to which it is true that not everything can be achieved through assessment. Comparing different types of assessment reform in one context or similar types of reform across various contexts would provide a fuller understanding of the dynamics.

Another limitation is the exclusive focus on mathematics and physics teachers. Although the teachers interviewed were a rather heterogeneous group with different backgrounds and experiences, their disciplinary approaches had an influence on their conceptions of knowledge and their experiences of being a teacher. For example, students and parents are typically more interested in grades in mathematics than, say, in the arts.

When interpreting our study results, it is crucial to bear in mind that the interviewees did not represent a selective sample of Finnish teachers. They had completed a survey before the interviews, which had perhaps guided their attention or at least made them spend extra time reflecting on the reform. The teachers interviewed were all volunteers, and during the interviews, it became clear that they were, on average, more familiar with the reform than teachers in general. For example, many of the teachers had responsibilities related to assessment in their school or

district. Therefore, the teachers' conceptions did not represent those of all Finnish teachers in 2022. In this study, the selective sample worked to our advantage because reforms take a long time to implement, and the teachers interviewed were already aware of the reform and had experienced its impact on their work.

6. Final words

The message of our study can be crystallised in three points. First, even a minor change in assessment cultures can create significant tensions in teachers' conceptions of assessment. Second, when an assessment culture shifts in a new direction, teachers not only adapt to the new features but also consider how these features compromise the prevailing assessment culture. Third, this compromise making adds to the tension of teachers' assessment work. It may be tempting for policymakers to pick the best parts of other countries' assessment cultures, but this does not necessarily work (e.g. Yan & Brown, 2021). Based on this study, we cannot say whether this dynamic is inevitable. Perhaps there might be a change that is only for the better, with no need for compromise making. What we can say is that we have not yet heard of a country that has managed to reap the full benefits of both low- and high-stakes assessment cultures.

Funding

This work was supported by the Ministry of Education and Culture in Finland [grant number: VN/25994/2021] and Kone Foundation [grant number: 202303077]

CRedit authorship contribution statement

Laura Ketonen: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Juuso Henrik Nieminen:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization.

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Acknowledgements

Thanks to our colleague Toni Mäkipää for his help in testing the coding framework. It was an important step in the development of the analysis.

References

- Alonzo, D., Labad, V., Bejano, J., & Guerra, F. (2021). The policy-driven dimensions of teacher beliefs about assessment. *Australian Journal of Teacher Education*, 46(3), 36–52. <https://search.informit.org/doi/10.3316/aeipt.228959>.
- Arrafii, M. A. (2022). “We must assess all, even a student farting [is also assessed] for the behavioural aspect of learning”: Teachers' conceptions of assessment in the context of assessment reform in Indonesia. *Curriculum Journal*, 33(1), 103–125. <https://doi.org/10.1002/curj.130>
- Atjonen, P., Laivamaa, H., Levonen, A., Orell, S., Saari, M., Sulonen, K., Tamm, M., Kamppi, P., Rumpu, N., Hietala, R., & Immonen, J. (2019). *Että tietää missä on menossa*: Oppimisen ja osaamisen arviointi perusopetuksessa ja lukiokoulutuksessa [“So that we know where we stand”. In *Assessment of learning and competence in basic education and general upper secondary education*]. The Publications of the Finnish Education Evaluation Centre no. 7. <http://hdl.handle.net/10138/300928>.

- Autio, T. (2021). From knowledge and bildung toward competences and skills in Finnish curriculum policy? Some theoretical, historical, and current observations related to Finland. In W. Zhao, & D. Tröhler (Eds.), *Euro-Asian encounters on 21st-Century competency-based curriculum reforms: Cultural views on globalization and localization* (pp. 41–56). Springer. https://doi.org/10.1007/978-981-16-3009-5_3.
- Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: The engine of systemic curricular reform? *Journal of Curriculum Studies*, 32(5), 623–650. <https://doi.org/10.1080/00220270050116923>
- Barnes, N., Fives, H., & Dacey, C. M. (2014). Teachers' beliefs about assessment. In H. Fives, & M. G. Gill (Eds.), *International handbook of research on teachers' beliefs* (pp. 284–300). Routledge.
- Barnes, N., Fives, H., & Dacey, C. M. (2017). US teachers' conceptions of the purposes of assessment. *Teaching and Teacher Education*, 65, 107–116. <https://doi.org/10.1016/j.tate.2017.02.017>
- Biesta, G. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability*, 21(1), 33–46. <https://doi.org/10.1007/s11092-008-9064-9>
- Birenbaum, M., DeLuca, C., Earl, L., Heritage, M., Klenowski, V., Looney, A., Smith, K., Timperley, H., Volante, L., & Wyatt-Smith, C. (2015). International trends in the implementation of assessment for learning: Implications for policy and practice. *Policy Futures in Education*, 13(1), 117–140. <https://doi.org/10.1177/1478210314566733>
- Black, P., & William, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551–575. <https://doi.org/10.1080/0969594X.2018.1441807>
- Bonner, S. M. (2016). Teachers' perceptions about assessment: Competing narratives. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 21–39). Routledge.
- Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, 22(2), 151–167. <https://doi.org/10.1080/713695728>
- Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, 11(3), 301–318. <https://doi.org/10.1080/0969594042000304609>
- Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students*. Nova Science Publishers.
- Brown, G. T., & Gao, L. (2015). Chinese teachers' conceptions of assessment for and of learning: Six competing and complementary purposes. *Cogent Education*, 2(1), Article 993836. <https://doi.org/10.1080/2331186X.2014.993836>
- Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education*, 27(1), 210–220. <https://doi.org/10.1016/j.tate.2010.08.003>
- Brown, G. T., & Remesal, A. (2012). Prospective teachers' conceptions of assessment: A cross-cultural comparison. *Spanish Journal of Psychology*, 15(1), 75–89. https://doi.org/10.5209/rev_SJOP.2012.v15.n1.37286
- Chan, K. T., & Tan, K. (2022). How teachers experience assessment tension and its effect on formative assessment practices. *Educational Research for Policy and Practice*, 21(3), 447–464. <https://doi.org/10.1007/s10671-022-09316-1>
- Chen, J., & Teo, T. (2020). Chinese school teachers' conceptions of high-stakes and low-stakes assessments: An invariance analysis. *Educational Studies*, 46(4), 458–475. <https://doi.org/10.1080/03055698.2019.1599823>
- Cooper, R., & Jordan, W. J. (2003). Cultural issues in comprehensive school reform. *Urban Education*, 38(4), 380–397. <https://doi.org/10.1177/00420859030380040>
- Dahler-Larsen, P. (2012). *The evaluation society*. Stanford University Press. <https://doi.org/10.1515/9780804778121>
- Darmody, M., Lysaght, Z., & O'Leary, M. (2020). Irish post-primary teachers' conceptions of assessment at a time of curriculum and assessment reform. *Assessment in Education: Principles, Policy & Practice*, 27(5), 501–521. <https://doi.org/10.1080/0969594X.2020.1761290>
- Daugherty, R., Black, P., Ecclestone, K., James, M., & Newton, P. (2008). Alternative perspectives on learning outcomes: Challenges for assessment. *Curriculum Journal*, 19(4), 243–254. <https://doi.org/10.1080/09585170802509831>
- DeLuca, C., Lapointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, 28, 251–272. <https://doi.org/10.1007/s11092-015-9233-6>
- Deneen, C. C., & Brown, G. T. L. (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent Education*, 3(1), Article 1225380. <https://doi.org/10.1080/2331186X.2016.1225380>
- Finnish National Agency for Education (FNAE). (2004). *The national core curriculum for basic education 2004*. Finnish National Agency for Education.
- Fives, H., & Buehl, M. M. (2012). Spring cleaning for the “messy” construct of teachers' beliefs: What are they? Which have been examined? What can they tell us?. In K. R. Harris, S. Graham, T. Urdan, S. Graham, J. Royer, & M. Zeidner (Eds.), *APA educational psychology handbook* (Vol. 2, pp. 471–499) American Psychological Association. <https://doi.org/10.1037/13274-019>. Individual differences and cultural and contextual factors.
- Fives, H., & Buehl, M. M. (2016). Teachers' beliefs, in the context of policy reform. *Policy Insights from the Behavioral and Brain Sciences*, 3(1), 114–121. <https://doi.org/10.1177/2372732215623554>
- FNAE. (n.d.). Tietoa perusopetuksen arvioinnin uudistusprosessista [Information on the process of assessment reform in basic education]. Finnish National Agency for Education. <https://www.oph.fi/fi/koulutus-ja-tutkinnot/tietoa-perusopetuksen-arvioinnin-uudistusprosessista>. Accessed May 20, 2024.
- FNAE. (2014). *The national core curriculum for basic education 2014* (Vol. 2016, p. 5). Finnish National Agency for Education. Publications.
- FNAE. (2020a). *Oppilaan oppimisen ja osaamisen arviointi perusopetuksessa [Assessment of and assessment for learning in basic education]*. Finnish National Agency for Education. https://www.oph.fi/sites/default/files/documents/perusopetuksen-arviointiluku-10-2-2020_1.pdf. (Accessed 20 May 2024).
- FNAE. (2020b). *Perusopetuksen päättöarvioinnin kriteerit [Criteria for the final assessment in basic education]*. Finnish National Agency for Education. (Accessed 20 May 2024).
- Fulmer, G. W., Lee, I. C. H., & Tan, K. H. K. (2015). Multi-level model of contextual factors and teachers' assessment practices: An integrative review of research. *Assessment in Education: Principles, Policy & Practice*, 22(4), 475–494. <https://doi.org/10.1080/0969594X.2015.1017445>
- Gebril, A., & Brown, G. T. (2014). The effect of high-stakes examination systems on teacher beliefs: Egyptian teachers' conceptions of assessment. *Assessment in Education: Principles, Policy & Practice*, 21(1), 16–33. <https://doi.org/10.1080/0969594X.2013.831030>
- Govaerts, M. J., van der Vleuten, C. P., & Holmboe, E. S. (2019). Managing tensions in assessment: Moving beyond either-or thinking. *Medical Education*, 53(1), 64–75. <https://doi.org/10.1111/medu.13656>
- Harlen, W. (2005). Teachers' summative practices and assessment for learning: Tensions and synergies. *Curriculum Journal*, 16(2), 207–223. <https://doi.org/10.1080/09585170500136093>
- Harris, L. R., & Brown, G. T. (2009). The complexity of teachers' conceptions of assessment: Tensions between the needs of schools and students. *Assessment in Education: Principles, Policy & Practice*, 16(3), 365–381. <https://doi.org/10.1080/09695940903319745>
- Hopfenbeck, T. N., Flórez Petour, M. T., & Tolo, A. (2015). Balancing tensions in educational policy reforms: Large-scale implementation of assessment for learning in Norway. *Assessment in Education: Principles, Policy & Practice*, 22(1), 44–60. <https://doi.org/10.1080/0969594X.2014.996524>
- Hwa, Y.-Y. (2022). “Our system fits us”: Comparing teacher accountability, motivation, and sociocultural context in Finland and Singapore. *Comparative Education*, 58(4), 542–561. <https://doi.org/10.1080/03050068.2022.2102754>
- Ketonen, L. (2024). Opettajien näkemyksiä matematiikan ja fysiikan päättöarvioinnin kriteereistä [Teachers' views of the final assessment criteria in mathematics and physics]. In *Ainedidaktikka*. Advance online publication. <https://doi.org/10.23988/ad.142709>.
- Kumpulainen, K., & Lankinen, T. (2012). Striving for educational equity and excellence: Evaluation and assessment in Finnish basic education. In H. Niemi, A. Toom, & A. Kallioniemi (Eds.), *Miracle of education: The principles and practices of teaching and learning in Finnish schools* (pp. 71–82). Sense.
- Lappalainen, H.-P. (2004). *Kerroin kaiken tietämäni. Perusopetuksen äidinkielen ja kirjallisuuden oppimistulosten arviointi 9. vuosiluokalla 2003 [I told everything I knew: The assessment of learning outcomes in mother tongue and literature in ninth grade in 2003]. Oppimistulosten arviointi 2/2004*. Finnish National Agency for Education.
- McMillan, H. J. (2013). Why we need research on classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 3–16). Sage.
- Meijer, K., Baartman, L., Vermeulen, M., & de Bruijn, E. (2023). Teachers' conceptions of assessment literacy. *Teachers and Teaching*, 29(7–8), 695–709. <https://doi.org/10.1080/13540602.2023.2190091>
- Ministry of Education and Culture. (2021). The Government has decided to raise the leaving age in compulsory education from the start of Autumn Term 2021. *Ministry of Education and Culture*. <https://julkaisut.valtioneuvosto.fi/handle/10024/162875>. (Accessed 20 May 2024).
- Moses, M. S., & Nanna, M. J. (2007). The testing culture and the persistence of high stakes testing reforms. *Education and Culture*, 23(1), 55–72. <https://doi.org/10.1353/eac.2007.0010>
- Niemi, H. (2012). The societal factors contributing to education and schooling in Finland – the Finnish education: Equity and quality as its main objectives. In H. Niemi, A. Toom, & A. Kallioniemi (Eds.), *Miracle of education* (pp. 23–40). Brill.
- Nieminen, J. H., Atjonen, P., & Remesal, A. (2021). Parents' beliefs about assessment: A conceptual framework and findings from Finnish basic education. *Studies In Educational Evaluation*, 71, Article 101097. <https://doi.org/10.1016/j.stueduc.2021.101097>
- Ouakrim-Soivio, N. (2013). *Toimivatko päättöarvioinnin kriteerit? Oppilaiden saamat arvosanat ja Opetushallituksen oppimistulosten seuranta-arviointi koulujen välisten osamieserojen mittareina [Are the criteria for final grading functioning? Student grades and NBE reviews on learning outcomes as indicators of between-school differences]. Raportit ja selvitykset 2013:9*. Finnish National Agency for Education. <http://urn.fi/URN:ISBN:978-952-13-5579-0>.
- Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, 62(3), 307–332. <https://doi.org/10.3102/003465430620033>
- Postareff, L., Virtanen, V., Katajavuori, N., & Lindblom-Ylänne, S. (2012). Academics' conceptions of assessment and their assessment practices. *Studies In Educational Evaluation*, 38(3–4), 84–92. <https://doi.org/10.1016/j.stueduc.2012.06.003>
- Pulkkinen, J., Hildén, R., Ketonen, L., Khawaja, A., Kotila, J., Mäkipää, T., Nissinen, K., Puhakka, E., Rantala, J., Rautopuro, J., Sulkunen, S., Söyrinki, S., & Tammelin-Laine, T. (2024). Arviointiuidistuksen alkutaipeilla. In *Perusopetuksen päättöarviointitutkimuksen loppuraportti [The early stages of assessment reform: Report of the research on final assessment in Finnish basic education]* (p. 6). The Publications of the Ministry of Education and Culture 2024. <http://urn.fi/URN:ISBN:978-952-263-723-9>.
- Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education*, 27(2), 472–482. <https://doi.org/10.1016/j.tate.2010.09.017>

- Rosin, T., Vaino, K., Soobard, R., & Rannikmäe, M. (2022). Examining Estonian science teachers' beliefs about teaching and assessment. *Cogent Education*, 9(1), Article 2104472. <https://doi.org/10.1080/2331186X.2022.2104472>
- Smith, W. (2016). *The global testing culture: Shaping education policy, perceptions, and practice*. Symposium Books.
- Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of teacher assessment: Evidence of what works best and issues for development*. <http://dspace.stir.ac.uk/handle/1893/32425>. (Accessed 20 May 2024).
- Stobart, G., & Eggen, T. (2014). *High-stakes testing in Education: Value, fairness and consequences*. Routledge.
- Teltemann, J., & Jude, N. (2019). Assessments and accountability in secondary education: International trends. *Research in Comparative and International Education*, 14(2), 249–271. <https://doi.org/10.1177/1745499919846174>
- Torrance, H. (2011). Using assessment to drive the reform of schooling: Time to stop pursuing the chimera? *British Journal of Educational Studies*, 59(4), 459–485. <https://doi.org/10.1080/00071005.2011.620944>
- Volante, L., DeLuca, C., Barnes, N., Birenbaum, M., Kimber, M., Koch, M., Looney, A., Poskitt, J., Smith, K., & Wyatt-Smith, C. (2024). International trends in the implementation of assessment for learning revisited: Implications for policy and practice in a post-COVID world. *Policy Futures in Education*, Article 14782103241255855. <https://doi.org/10.1177/14782103241255855>
- Whittaker, A., & Young, V. M. (2002). Tensions in assessment design: Professional development under high-stakes accountability. *Teacher Education Quarterly*, 29(3), 43–60. <http://www.jstor.org/stable/23478389>.
- Wikström, C. (2005). Grade stability in a criterion-referenced grading system: The Swedish example. *Assessment in Education: Principles, Policy & Practice*, 12(2), 125–144. <https://doi.org/10.1080/09695940500143811>
- Winstone, N. E., & Carless, D. (2021). Who is feedback for? The influence of accountability and quality assurance agendas on the enactment of feedback processes. *Assessment in Education: Principles, Policy & Practice*, 28(3), 261–278. <https://doi.org/10.1080/0969594X.2021.1926221>
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149–162. <https://doi.org/10.1016/j.tate.2016.05.010>
- Yan, Z., & Brown, G. T. L. (2021). Assessment for learning in the Hong Kong assessment reform: A case of policy borrowing. *Studies In Educational Evaluation*, 68, Article 100985. <https://doi.org/10.1016/j.stueduc.2021.100985>
- Yates, A., & Johnston, M. (2018). The impact of school-based assessment for qualifications on teachers' conceptions of assessment. *Assessment in Education: Principles, Policy & Practice*, 25(6), 638–654. <https://doi.org/10.1080/0969594X.2017.1295020>