Nordic
Studies in Education

Articles | Peer-reviewed | Vol. 42, No. 4, 2022, pp. 328–346

Connecting Student Perceptions and Classroom Observations as Measures of Cognitive Activation

Jóhann Örn Sigurjónsson
University of Iceland, Reykjavík, Iceland
Contact corresponding author: Jóhann Örn Sigurjónsson, email: johannorn@hi.is

Anna Kristín Sigurðardóttir
University of Iceland, Reykjavík, Iceland

Berglind Gísladóttir
University of Iceland, Reykjavík, Iceland

Jorryt van Bommel
Karlstad University, Karlstad, Sweden

ABSTRACT
Which dimensions of instruction can be reliably captured using student perception surveys, is subject for debate. The aim of this study is to empirically explore the validity and limitations of two different measures of cognitive activation: systematic classroom observations and student perceptions. 34 video-recorded lessons from ten lower secondary mathematics teachers in Iceland were analysed using an observation system and compared to 217 responses to the Tripod student perception survey. The results indicate that for the cognitive activation dimension, the connection between observer ratings and student perceptions is weak, raising questions about the validity of different measures of instructional quality.

Keywords: *cognitive activation, student perceptions, classroom observations, instructional quality*

Student perception surveys are frequently incorporated as a cost-effective data source in both research on instructional practice and educator accountability systems (Phillips et al., 2021). Studies have indicated some promising evidence for convergence between student ratings and teachers' value-added scores, although caution is advised in basing high-stakes decisions on such ratings (Buchholtz et al., 2020; Kuhfeld, 2017;

Manuscript received: 10.11.2021. Manuscript accepted: 24.08.2022.

ISSN 1891-5949, **https://noredstudies.org**

Citation: Sigurjónsson, J. Ö., Sigurðardóttir, A. K., Gísladóttir, B. & van Bommel, J. (2022). Connecting student perceptions and classroom observations as measures of cognitive activation. *Nordic Studies in Education, 42*(4), 328–346. https://doi.org/10.23865/nse.v42.3636

Sandilos et al., 2019). Questions remain about which dimensions of instruction can be reliably captured by student perception surveys, as student ratings may not be equally valid for every dimension (Praetorius, 2014; Wallace et al., 2016). Further knowledge is warranted on the validity of specific dimensions of instruction in student perception surveys. Their development may come alongside the exploration of synergies across classroom observation frameworks, thus moving the field of instructional practice studies forward and toward a more common lexicon (Grossman & McDonald, 2008; Praetorius & Charalambous, 2018).

Cognitive activation has widely been acknowledged as an important dimension of instructional quality (Baumert et al., 2010; Bell et al., 2019; Krauss et al., 2020). In mathematics classrooms where instruction is considered cognitively activating, students are invited to explain their thinking, justify their solutions, provide reasoning as well as generate ideas and conjectures. The teacher can create potential for cognitively activating instruction by selecting appropriately challenging tasks and engaging in mathematically rich practices, and thus facilitate students' cognitive activity (Praetorius & Charalambous, 2018). Such practices involve both presenting students with an adequate level of intellectual challenge as well as providing them with opportunities for mathematics-related discourse in the classroom. In research on teaching and learning, these factors have been measured by both systematic classroom video observations and task analysis (Sigurjónsson & Gísladóttir, 2020; Tekkumru-Kisa et al., 2020). While systematic analysis of classroom video is a time-consuming research method, it provides opportunities for thorough analysis of the complex interactions that take place during instruction (Blikstad-Balas, 2017; Snell, 2011). Observation systems provide the indispensable common vocabulary to describe and analyse instruction systematically despite being sensitive to rater error and possible biases when applied across contexts (Luoto et al., 2022; White, 2018).

The aim of this study is to contribute to the empirical knowledge of the validity and limitations of student perception surveys and classroom observations in measuring cognitive activation. In this paper, results are presented based on video observations and student perceptions in mathematics lessons at the lower secondary level in Iceland. The study builds on data from the Quality in Nordic Teaching (QUINT) research initiative whose vision is to investigate instructional quality in the Nordic countries using video observations and student perception surveys. By examining the connection between classroom observations and self-reported student perceptions of cognitive activation, it is illustrated how the results from two different approaches may align or differ. The research question addressed is: What is the nature of the connection between classroom observations and student ratings as measures of cognitive activation?

Cognitive activation

Although frameworks differ somewhat in their exact conceptualisations, there is a general agreement that instructional quality is a multidimensional construct (Grossman et al., 2013, 2014). Cognitive activation is claimed to emerge consistently as one of three crucial components of instructional quality, along with individual learning support and efficient classroom management (Kunter et al., 2013). Furthermore, cognitive activation is among the dimensions of teaching most often represented in both mathematics-specific and content-generic frameworks for analysing instructional quality (Bell et al., 2019; Praetorius & Charalambous, 2018). In an analysis of twelve frameworks that have been employed in studying mathematics instruction, Praetorius and Charalambous (2018) conceptualised cognitive activation as consisting of three aspects of teaching practice: (1) the teacher's selection of challenging tasks and use of mathematically rich practices, (2) facilitation of students' cognitive activity, and (3) supporting students' meta-cognitive learning from cognitively activating tasks. In mathematics, students have opportunities to be cognitively activated when they are given the chance to actively participate in idea generation and conjectures, explain their thinking, reason, or justify their solutions. In the field of mathematics education, several studies have been conducted on different aspects of cognitive activation. Cognitively activating instruction can be viewed as conducive to providing opportunities for students' productive struggle, meaning they expend reasonable effort to make sense of mathematics (Granberg, 2016; Hiebert & Grouws, 2007). In Russo and Hopkins' (2017) results from interviews of 73 elementary school students, evidence was found that students enjoyed the process of being challenged in mathematics and embraced the struggle involved with demanding tasks. These results contrasted with teacher concerns about older students' struggles to engage meaningfully in whole-class mathematical discussions (Leikin et al., 2006). However, more recent studies of teacher perceptions have shown less concern and generally a more positive attitude to posing challenging tasks involving teacher-facilitated discussion (Russo & Hopkins, 2019; Sigurjónsson & Kristinsdóttir, 2018). Thus, mathematics teachers seem to increasingly agree that facilitating opportunities for meaningful engagement with mathematical concepts through cognitively activating instruction is both desirable and good practice.

Cognitive activation has been measured in various ways and associations have been found with other educational factors. Importantly, evidence suggests that cognitively activating instruction positively impacts student achievement (Baumert et al., 2010) and that such teaching practices are significantly predicted by a teacher's professional content knowledge (Wilhelm, 2014). In a recent study including 163 mathematics classrooms, approximately 90% of the variance in learning gains was explained by a model in which observable teaching behaviour included three dimensions: cognitive activation, individual learning support and classroom management (Krauss et al., 2020). In the study, the potential for cognitively activating instruction was analysed using mathematical tasks as the unit of analysis, which is a common method (see e.g. Neubrand et al., 2013; Sigurjónsson & Gísladóttir, 2020; Tekkumru-Kisa et al., 2020). This method of analysis may include either tasks as exercises from textbooks and other learning materials, or tasks from tests and examinations. Although strictly analysing tasks may reliably indicate the teacher's selection of challenging tasks, this approach has limitations regarding the measurement of the two other equally important aspects of cognitive activation that involve interactions in the classroom: the use of mathematically rich practices and facilitation of cognitive activity (Praetorius & Charalambous, 2018). The extent to which students explain their own ideas and engage with others' ideas in classroom-situated discourse is another indicator of cognitively activating instruction in practice. In a study where video recordings were used to examine this aspect, the results indicate that in order to understand how teaching practices relate to student learning it is necessary to consider student participation (Ing et al., 2015). For student participation and classroom interaction to be considered in measuring cognitive activation, data collection methods such as classroom observation or student self-reports are required.

Video-based classroom observation

While observation has been at the heart of classroom research for decades, technological advances and improved access to suitable recording devices have made video-based observations an increasingly more viable research method (Xu et al., 2019). Among the distinctive features of video data is that it is a real-time sequential medium, meaning that researchers can review video segments as many times as they like (Jewitt, 2012). This has created opportunities for multiple methods of analysis as well as multiple researchers analysing and interpreting the same data. However, this distinctive feature also presents researchers with challenges. Blikstad-Balas (2017) outlines key challenges of using video in researching social practices. One challenge is preserving important contextual framing while providing detailed enough analyses. Another challenge is "magnification" of interesting yet atypical events in the data at the cost of potentially ignoring or missing other relevant information. A common approach in analysing large volumes of classroom video data is to gain an overview across the dataset using standardised scores and then analyse segments of particular interest in more detail (Klette, 2009; Snell, 2011). Such an approach requires observing and scoring the data in a systematic way.

Alongside the development toward rewatchable video data, research has been increasingly directed toward qualities of observation systems. Observation systems are assessment systems comprised of rating tools, rating processes, and sampling specifications, where the purpose is either to understand or improve teaching (Bell et al., 2019; Liu et al., 2019). One of the benefits of observation systems is the common vocabulary for classroom interactions that they offer for researchers and practitioners. While the nature of observation systems is to reduce the qualitative richness of classroom activity to quantified scores, this reduction both enables systematic comparison between classrooms and is argued to provide the common tools for moving the field of teaching and learning forward (Grossman & McDonald, 2008; Klette & Blikstad-Balas, 2018). Systematic scoring of lessons also requires assessing how accurately the given observation system captures instruction in the specific context in which it is applied. Continued developments of observation systems to provide valid and relevant results across contexts remains a challenge for educational researchers (Lietke, 2019; Luoto et al., 2022).

Student perceptions of instruction

Administering questionnaires of student perceptions as a method of collecting classroom data has decades of history (den Brok et al., 2006; Fraser, 1998; Fraser & Walberg, 1981). Since students spend time near daily in classrooms with various subject teachers, they can be considered to have valuable insights that external observers cannot feasibly gain. Several student questionnaires have been used as a measure of instructional quality, (see e.g. de Jong & Westerhof, 2001; Kane & Staiger, 2012) and care in choosing an appropriate data source for constructs to be measured has been recommended (Kunter & Baumert, 2006). Which dimensions of instructional quality have reliable and valid measures through student surveys is still a matter of debate.

The Tripod survey is one student perspective research instrument, originally developed to gather feedback from students for school improvement in seven dimensions of teaching (Ferguson, 2010; Phillips et al., 2021). Factor analyses have shown strong correlations between the seven dimensions, suggesting that the Tripod questionnaire can be reduced to two dimensions at the between classrooms level: classroom management and instructional support (Kuhfeld, 2017; Schweig, 2014; Wallace et al., 2016). In the results of the Tripod survey as employed in the large-scale Measures of Effective Teaching study (MET), the principal components accounting for most of the variance in student responses at the teacher level were also found to be two: classroom management and teachers' overall performance (Kane & Staiger, 2012). In a study of classrooms in Norway, the results of the Tripod survey were portrayed by individual survey items, showing the three highest and three lowest rated items along with the results of items that had been suggested as key for student achievement in the results of the MET study (Klette et al., 2017). In their results, all the survey items suggested to be key scored above the mean score for all items (M = 3.80). The lowest rated items had to do with student agency and learning enjoyment, while the highest rated items were about respect for the teacher and the quality of teacher explanations and assistance.

The relationship between student perceptions and classroom observations

Student perceptions and their relationship to other measures have been studied with somewhat mixed results. In the Global Teaching InSights study, most students reported being cognitively engaged in lessons despite observation scores being mostly on the low end of the scale in cognitive engagement (OECD, 2020). Further, Wallace et al. (2016) used the large-scale MET database to compare the general response factor of Tripod to observational scores in three instructional domains in middle school mathematics lessons. A weak correlation was found both with the emotional support domain and instructional support, whereas no correlation was found with classroom management. Drawing on data from 291 mathematics middle school grades from same dataset, Sandilos et al. (2019) examined correlations between the proposed seven dimensions of Tripod and two observation systems. Some correlations were found between the student perceptions and observation scores. A lower-medium correlation was found between the specific dimensions identified to represent rigor, i.e. Tripod's "challenge" dimension (measuring both reasoning and persistence) and the observation system's dimension "establishing a culture for learning". A moderate correlation was found between observation scores on the "questioning & discussion" factor and all Tripod dimensions, the strongest being dimensions regarding rigor ("challenge") and respecting student perspectives and promoting discussion ("confer"). Each instrument's relation to value-added measures was also studied. Three Tripod dimensions ("control", "challenge", and "clarify") were positively related to value-added measures, while score variability within Tripod "control" and "challenge" were negatively related to value-added measures. No significant relations to value-added measures were found for the instructional domains in the observation systems in the middle school mathematics grades. The mixed results from these two studies show a somewhat unclear connection between Tripod student ratings and systematic observer ratings of teaching, but some promising results in the relation between Tripod and value-added measures. As a result, caution has been advised in using student surveys for high-stakes teacher evaluations (Kuhfeld, 2017; Phillips et al., 2021). However, studies on the connection between student perceptions and classroom observations in specific dimensions of instruction, such as cognitive activation, are lacking.

Contrasts between what is observed in classrooms and student self-reported perceptions is not a new phenomenon. In what was dubbed as the "Expanded relevance paradox", Clarke (2006) described the paradoxical results of comparisons between classrooms in Sweden and Hong Kong. The application-oriented mathematics teaching found in Swedish classrooms was associated with students finding the subject irrelevant to their lives, while in Hong Kong classrooms the pure mathematics-oriented teaching was associated with students finding the subject important and relevant. A similar paradox may apply in other aspects of instruction.

Mathematics teaching in Iceland

Studies of mathematics teaching in Iceland have shown that teachers prioritise students' individual seatwork and practice in procedural fluency. In a report on lower secondary mathematics teaching, a majority of lessons were dominated by individual seatwork in textbooks with only a third of observed lessons including explicit whole-class instruction from the teacher with student discussions (Þórðardóttir & Hermannsson, 2012). Subsequent studies have reported the same common instructional pattern, with student collaboration and teacher-facilitated mathematical discussions existent but uncommon (Gunnarsdóttir & Pálsdóttir, 2015; Sigurgeirsson et al., 2014). In most mathematics lessons, the selected tasks are low-level, or the teachers' implementation of tasks results in low intellectual challenge (Sigurjónsson & Gísladóttir, 2020).

Student perceptions of instruction was one theme of a recent large-scale study on instructional practice in Iceland. A total of over 1600 Icelandic lower secondary students were asked to assess how important they found seven different educational factors. The results showed medium to strong correlation between every factor. The factor labelled "every student's well-being" was deemed most important by students, with 87% responding with either somewhat or strong agreement. The second-least important factor was "training students to think analytically and draw inferences", with 71% in either somewhat or strong agreement (Björnsdóttir & Jónsdóttir, 2014). Drawing on the same data, Sigþórsson et al. (2014) found 69% of students considered the instructional quality in their school either somewhat good or very good. Yet, significantly fewer students said they enjoyed school or were interested in their studies, with boys expressing less interest and enjoyment than girls. In follow-up interviews with students, they called for more diverse tasks and options in assignments, stating that demanding tasks would inspire more interest and enjoyment. However, the study did not aim to draw connections between the reported student perceptions and classroom observations. That is the intention of the current study.

Method

The sample consisted of ten mathematics teachers from separate schools in Iceland. The teachers were sampled with the aim of establishing heterogeneity of school variables, including school size, urban and rural locations, traditional and team teaching, and differing proportions of immigrant students. Three to four consecutive mathematics lessons were video recorded for each teacher, a total of 34 recorded lessons. Students in the observed mathematics classrooms (n = 217) filled out a survey measuring their perceptions of their teachers. All recorded lessons were in 8th grade classrooms with students aged 13–14. The survey administered to the students is a translated version of the Tripod student perception survey which was piloted for a Nordic context in Norway (Ferguson, 2010; Klette et al., 2017).

Data analysis: Classroom observations

The lessons were scored using the Protocol for Language Arts Teaching Observations (PLATO). The protocol has been employed successfully in other subjects such as mathematics (Mahan et al., 2021). In PLATO, lessons are scored in 15-minute segments. Each segment receives a score on a 4-point scale on various elements of teaching, depending on the amount of evidence found in support of each element (Grossman, 2019). PLATO consists of 12 elements, two of which relate to cognitive activation: Intellectual Challenge and Classroom Discourse (Bell et al., 2019; from here on abbreviated as IC and CD, respectively). Low-level IC (1 and 2) involves rote or procedural tasks where students apply given procedures. High-level IC (3 and 4) is where students engage in high-level thinking, e.g., by reasoning or justifying their solutions. A score in IC can be adjusted by one point if the teacher changes the task from how it was initially presented. The challenge can either be increased, for instance if the teacher asks students to further explain their thinking or solution method, or reduced, such as if the teacher solves the task for students. The CD element is divided into two sub-components: uptake of student responses and opportunities for student talk, which together create an overall score with uptake weighing more. Low-level CD (1 and 2) includes either no, automatic, or brief teacher uptake of student ideas. High-level CD (3 and 4) may show the teacher elaborating, revoicing, or asking for clarification of student ideas. Further, for CD to reach a high level at least a third of the segment must include opportunities for students to engage in content-related discourse.

The lessons were scored by three certified PLATO raters. To be certified, one must pass a course with a certification test (White, 2018). For inter-rater reliability assurance, the first segment of every other lesson was scored by two raters independently. In case of a disagreement, the two raters discussed the reasoning for their scores to reach an agreed score in accordance with the protocol.

Data analysis: Student perceptions

In this study, two subscales were constructed from the Tripod items that fit the specific aspects of instruction measuring cognitive activation in the observation protocol. One subscale formed a measure of student perceptions of the extent to which the teacher engaged students' reasoning and explanations of their answers. The other formed a measure of student perceptions of discourse in the classroom, i.e., the teacher respectfully inviting students to share their thoughts or ideas. The construction of the subscales may reveal possible nuances between similar constructs in PLATO and the Tripod survey. The reasoning scale was constructed from three survey items that best fit the IC element from PLATO. The discourse scale was similarly constructed from three survey items that best fit the CD element from PLATO. The reasoning and discourse subscales included specific items from Tripod's Challenge dimension and Confer dimensions, respectively. Items that specifically relate to cognitive activation were chosen, while items that do not directly relate to cognitive activation were not used. Table 1 shows the survey items used.

Table 1: Items used from the Tripod survey and Cronbach's alpha values for the two subscales.

ITEM	ITEM TEXT	SUBSCALE	α
REAS1	My teacher asks questions to be sure we are following along when s/he is teaching.	Reasoning	0.67
REAS2	My teacher asks students to explain more about the answers they give.		
REAS3	My teacher wants me to explain my answers—why I think what I think.		
DISC1	My teacher wants us to share our thoughts.	Discourse	0.65
DISC2	My teacher gives us time to explain our ideas.		
DISC3	My teacher respects my ideas and suggestions.		

A reliability analysis of the subscales yielded a Cronbach's alpha value of 0.67 for the reasoning scale and 0.65 for the discourse scale. The mean score for all items in the survey was 3.62.

It is worth noting that some of the items refer to the teachers' actions toward the entire class ("we" and "us") while others refer only to the specific student responding ("me" and "my"). In the Icelandic translation of the survey this feature was retained, preserving some potential to reflect differences in how students perceive that teachers act toward them individually.

To connect the observation scores to student ratings, the PLATO scores for IC and CD were used to determine different groups of teachers according to the observational evidence and an order of teachers within groups. For instance, teachers with at least one segment at the 4-level in either IC and CD comprise the group with the strongest evidence of cognitively activating instruction. Within groups, teachers are ordered according to the sum of their mean scores in IC and CD. Considering maximum segment scores in the relevant elements before mean scores ensures that observable evidence of cognitive activation at a high level is not devalued by having some segments at a low level.

Ethical considerations

Ethical issues are inherent with video recordings of classroom activity. All procedures about confidentiality, permission, storing and sharing of data have been successfully reviewed and acknowledged by the Science Ethics Committee of the University of Iceland. The study is conducted in accordance with the Data Protection Act no. 90/2018. For each recorded classroom an informed consent form was received from participating teachers, as well as from each student and their guardians. An informed consent form was received from participating teachers, each student, and their guardians in the recorded classrooms. The consent states agreement or disagreement with different parts of the larger study, such as being seen in the videos, use of video for professional development purposes, and video being shared with other researchers within the research centre. Participants were informed that they could withdraw from the study at any time.

Results

The results of the observation data analysis will first be delineated to show how the different groups of teachers have been identified and ordered. The results of the student perceptions are then presented in light of the observation results.

Observed features of cognitive activation

The distribution of scores for IC and CD by teachers are seen in figures 1 and 2, respectively.

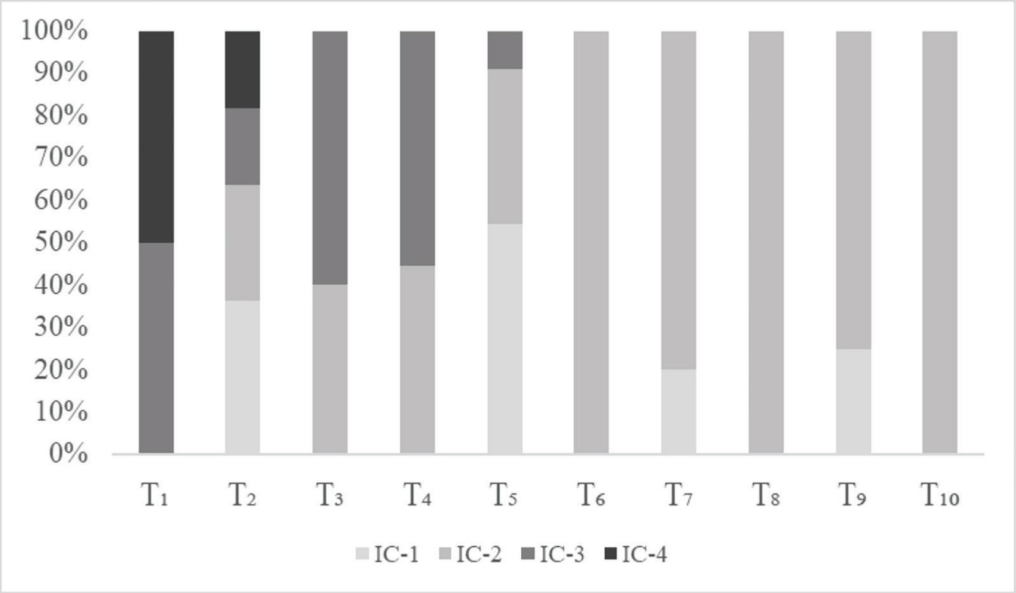

Figure 1: Distribution of segment scores by teachers in PLATO-IC (Intellectual Challenge).

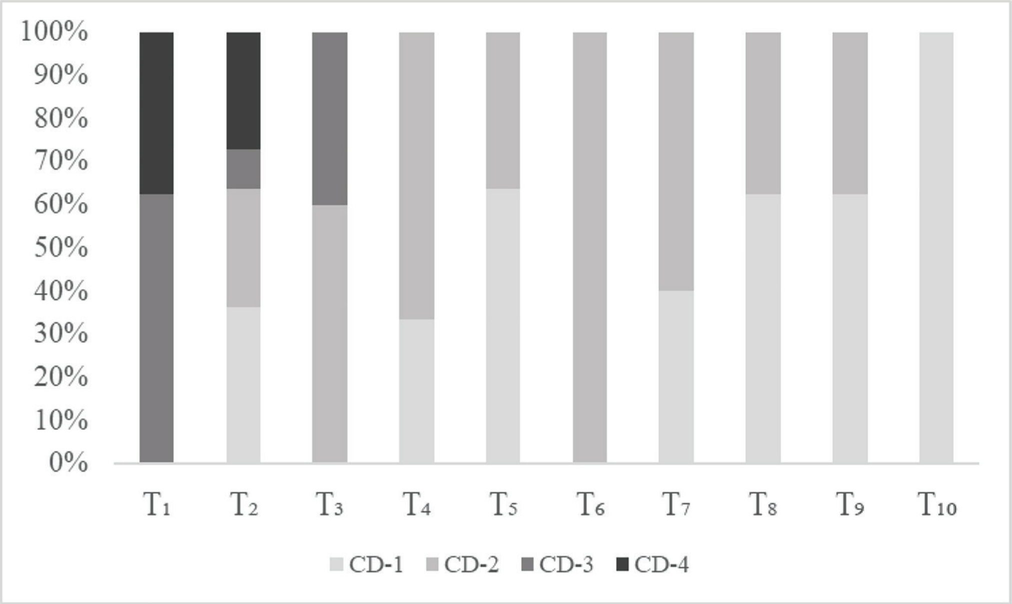

Figure 2: Distribution of segment scores by teachers in PLATO-CD (Classroom Discourse).

Table 2 shows three ordinal groups created by the maximum scores each teacher received for IC and CD in the observed lesson segments. Group A consists of teachers who had 4-level segment scores in both IC and CD, while Group B consists of teachers whose highest score in these elements was at the 3-level. Group C consists of teachers whose highest score in these two elements was at the 2-level. Within groups, teachers are ordered according to the sum of their mean scores in IC and CD.

Table 2: Ordering of teachers by ordinal groups from maximum scores in IC and CD, and within groups by sum of mean scores of IC and CD.

TEACHER	MAX(IC)	MAX(CD)	GROUP	MEAN(IC)	MEAN(CD)	MEAN(IC + CD)
T_1	4	4	A	3.50	3.38	6.88
T_2	4	4		2.18	2.27	4.45
T_3	3	3	B	2.60	2.40	5.00
T_4	3	2		2.56	1.67	4.22
T_5	3	2		1.55	1.27	2.82
T_6	2	2	C	2.00	2.00	4.00
T_7	2	2		1.80	1.60	3.40
T_8	2	2		2.00	1.38	3.38
T_9	2	2		1.75	1.38	3.13
T_{10}	2	1		2.00	1.00	3.00

Teachers T_1 and T_2 exhibited the strongest evidence of cognitive activation and comprise group A. Group B consists of teachers T_3, T_4 and T_5. They showed some evidence of cognitive activation but with some weaknesses, particularly in CD. Lastly, group C is teachers T_6 to T_{10} who showed the weakest evidence of cognitive activation in the observed lesson segments, with consistent scores at the 1-level and 2-level in both the IC and CD elements.

Student perceptions of cognitive activation

The student perceptions of cognitive activation can be connected to the observed cognitive activation by viewing their responses in context with the ordering of teachers based on the observed features of cognitive activation, where T_1 showed the strongest evidence and T_{10} the weakest. The means and standard deviations for the reasoning and discourse scales as rated by students in the Tripod survey are reported in table 3 along with observation scores for comparison. Student responses with a missing item response are omitted from table 3.

Table 3: Means and standard deviations for the Reasoning and Discourse subscales along with PLATO observation scores for IC and CD.

		REASONING		DISCOURSE		GROUP	PLATO-IC			PLATO-CD		
TEACHER	*N*	MEAN	SD	MEAN	SD		MAX	MEAN	SD	MAX	MEAN	SD
T_1	23	3.26	0.78	3.35	0.52	A	4	3.50	0.53	4	3.38	0.52
T_2	9	4.26	0.36	3.70	0.93		4	2.18	1.17	4	2.27	1.27
T_3	37	3.53	0.70	3.14	0.83	B	3	2.60	0.55	3	2.40	0.55
T_4	16	3.38	0.50	3.33	0.80		3	2.56	0.53	2	1.67	0.50
T_5	20	3.47	0.71	2.93	0.78		3	1.55	0.69	2	1.27	0.47
T_6	10	4.03	0.76	3.93	0.86	C	2	2.00	0	2	2.00	0
T_7	11	3.30	0.72	3.39	0.61		2	1.80	0.42	2	1.60	0.52
T_8	28	3.55	0.67	3.63	0.66		2	2.00	0	2	1.38	0.52
T_9	11	3.48	0.92	3.42	0.54		2	1.75	0.46	2	1.38	0.52
T_{10}	24	3.07	0.85	3.18	0.70		2	2.00	0	1	1.00	0
Total	189	3.46	0.76	3.32	0.75			2.15	0.77		1.80	0.87

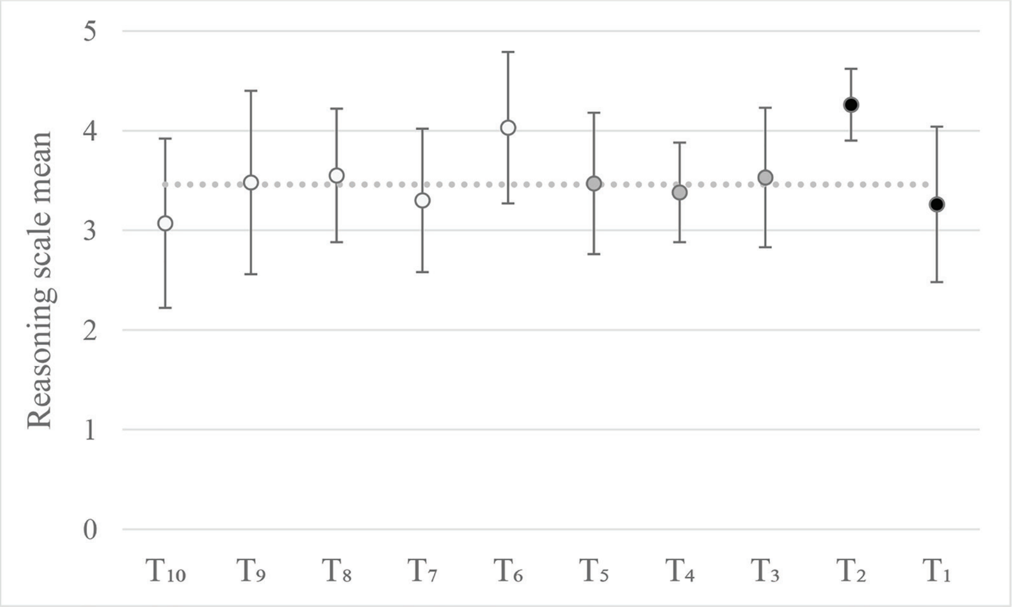

Figure 3: Reasoning subscale means by teachers in order by observed evidence of cognitive activation.

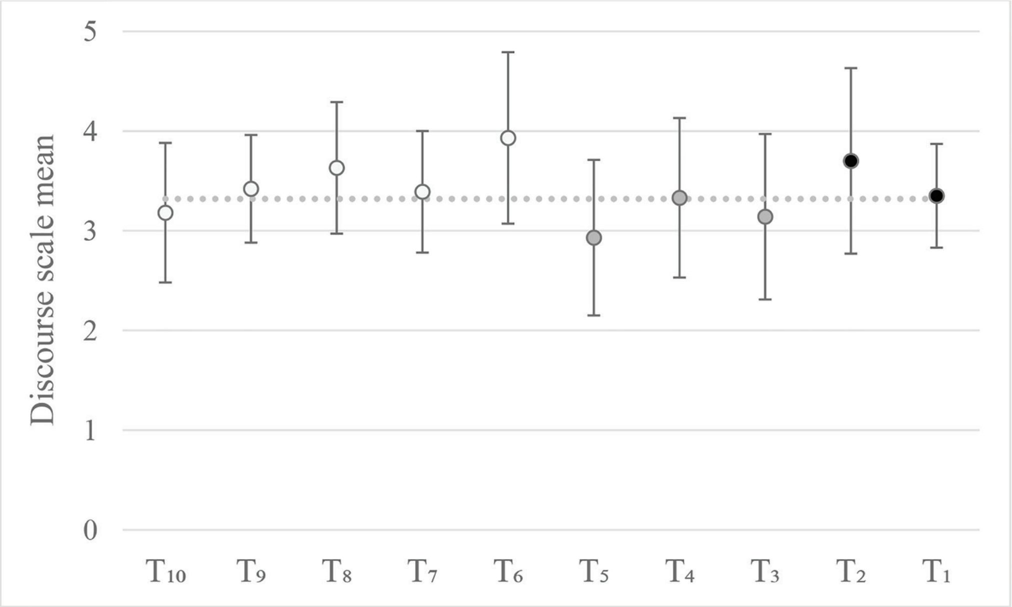

Figure 4: Discourse subscale means by teachers in order by observed evidence of cognitive activation.

Student ratings of T_2 are the highest in the reasoning subscale. Standard deviation is also comparatively low in this subscale, indicating that the students are generally in agreement that T_2 provides them with opportunities to explain their thinking and reason. However, T_1 is below the average in all reasoning subscale items with the second lowest overall mean rating. In group B, T_3, T_4 and T_5 share similar ratings in the reasoning subscale, close to the overall mean. Some teachers in group C, such as T_6 and

T_8, are rated at or above the average in reasoning, while T_{10} receives the lowest overall rating.

On the discourse scale, student ratings of teacher T_2 are second-highest, but with a higher standard deviation than in the reasoning subscale. Student ratings of T_1 are around the average with a relatively low standard deviation. From group B, T_3 and T_4 remain close to the mean while T_5 has the lowest overall mean rating. T_6 from group C receives the highest student rating on the discourse scale. Other teachers from group C hover around or above the overall mean for discourse, except T_{10} who is slightly below the mean. Figures 3 and 4 illustrate the mean student perceptions in the reasoning and discourse subscales by teachers in order by the observed evidence of cognitive activation. The error bars show one standard deviation in each direction.

Discussion

The study aimed to contribute to the empirical knowledge of the validity and limitations of student perception surveys and classroom observations in measuring cognitive activation. The results suggest that there is a discrepancy between student perceptions and classroom observations as measures of cognitive activation. The teachers that were observed to show strong evidence of cognitively activating instruction according to PLATO were evidently not rated highly by students on relevant items on the Tripod survey. The two teachers showing strong and consistent evidence in both intellectual challenge and classroom discourse showed different results. T_1 scored consistently at the 3-level and 4-level in those instructional dimensions but was consistently rated around the mean for each relevant item in the student survey. On the other hand, T_2 had more variability in the PLATO-scores but was the teacher who was rated highest on the reasoning subscale and second highest on the discourse subscale by the students. There was strong agreement among the students that T_2 asked them to provide reasoning for their solutions. The results from T_1 and T_2 may suggest that students in classrooms where cognitively activating instruction takes place in tandem with more rote activities may acknowledge or notice it to a greater extent than students who constantly receive cognitively activating instruction.

Variability in scores can indicate differences in student outcomes (Sandilos et al., 2019). T_2 had the lowest standard deviation in reasoning and the highest mean, but the highest standard deviation in discourse and a mean slightly above average. On the other hand, T_1 had an average standard deviation in reasoning and a mean slightly below average, but the lowest standard deviation in discourse with an average mean. Thus, these results may indirectly support the argument that high variability in student ratings relates negatively to student outcomes. There is some evidence for stronger agreement among students of teachers with higher observation scores, although the relation between the two warrants further study.

As the field of research on teaching and learning lacks a common language for different dimensions of instructional quality (e.g. Grossman & McDonald, 2008), it is worth discussing the terms used in the present paper on cognitive activation. Cognitive activation, according to Praetorius and Charalambous (2018), is the teacher's selection of challenging tasks and use of mathematically rich practices, facilitation of students' cognitive activity and supporting students' meta-cognitive learning from cognitively activating tasks. Using this conceptualisation, the IC and CD elements from PLATO were chosen as the observational measures of cognitive activation (Bell et al., 2019). For the student perceptions, specific items from the Tripod survey were chosen to construct two subscales that were deemed to represent student ratings of the cognitive activation dimension of instruction by privileging aspects of cognitive activation measured by the two PLATO elements: the opportunity teachers give students to provide reasoning for their solutions and answers, and opportunities to engage in content-related discourse.

In the reasoning subscale, there was some discrepancy between survey items. In REAS1, it is entirely possible that many teachers did indeed ask their students questions for understanding yet proceeded to assist them in ways that diminished their productive struggle, resulting in high student ratings but low observation scores. As the analysis suggests, this item does not seem to accurately measure the same aspects of cognitive activation that are captured in PLATO. It is in less agreement with other related questions, which matches previous empirical findings (Schweig, 2014). It is also worth reflecting on how student responses may differ if framed in the context of student actions, as is done in the classroom management dimension of Tripod, as opposed to the actions of their teacher. For instance, the REAS2 item ("My teacher asks students to explain more about the answers they give") might instead be phrased as "I explain more details about the answers that I give" – or REAS3 ("My teacher wants me to explain my answers – why I think what I think") might be phrased as "Students in my class explain why they think what they think in their answers".

The results of the study also raise the question of student raters' interpretation of survey items across different contexts (Wallace, 2016). Would student raters in other research contexts interpret the items pertaining to cognitive activation in a way that would produce different results? Most students participating in the Global Teaching InSights study reported that they felt cognitively engaged even though observation scores in that dimension were mostly on the low end (OECD, 2020). There appears to be emerging evidence that the "expanded relevance paradox" coined by Svan may apply elsewhere, and perhaps there also exists a certain "cognitive activation paradox" between the observed cognitive activation in instruction and students' perceptions of their cognitive engagement (Clarke, 2006). The extent to which this discrepancy is real and the extent to which it may be due to different interpretations is not entirely clear.

The discrepancy found in these results reveals possibilities for future research with a larger sample, and possibly warrants some caution in using student ratings for evaluating cognitive activation. Although research has shown that cognitive activation is an important part of instructional quality (Bell et al., 2019; Krauss et al., 2020), the results of this study give reason to doubt that cognitively activating instruction is directly connected to students' experiences of instruction being cognitively activating, despite moderate correlations found in a previous study using the (entire) relevant Tripod dimensions with another observation system's "questioning and discussion" construct (Sandilos et al., 2019). Hiebert & Grouws stated that instruction is not merely effective or not effective – instruction is effective for something (2007). Evidence suggests that cognitively activating instruction positively impacts student achievement (Baumert et al., 2010). Further research is justified on the connection between student perceptions and educational outcomes, such as student achievement, different instructional dimensions, or teacher value-added scores.

In this study, an effort was made to connect two different measures of instructional practice in mathematics lessons. In discussing the divergent results, it is important to note the difference between the measures. One was conducted by adult observers trained in a classroom observation framework utilised to analyse four lessons in segments from each teacher via video recording. The other was conducted by 13-year-olds filling out a student perception survey. Students are not trained in identifying and analysing various instructional practices, nor are they expected to be. However, they observe day-to-day classroom practices for the entire school year. Therefore, they can provide valuable insight for researchers who can only feasibly observe a fraction of the lessons that the students attend. Neither measure will show absolute truth – both measures will have an inherent measurement error. Inter-rater reliability remains a central issue in employing classroom observation frameworks, and uncertainty and error is part of video coding (White, 2018). As for student ratings, the Tripod instrument has been shown to measure classroom management as one factor, but evidence is lacking regarding whether the instrument can be used to measure other factors than "teaching in general" in more detail (Wallace et al., 2016). However, promising evidence has been found in positive correlations between the classroom management factor and student outcomes (Sandilos et al., 2019). Naturally, this leads to the question of which other dimensions of instructional quality can be obtained from student perception surveys.

Conclusion

Echoing Praetorius' (2014) concerns about equal validity between dimensions of student rating, the results of this study give reason to question the validity of students' capacity to assess more nuanced and complex pedagogical constructs such as cognitive activation. Even though Tripod does not explicitly claim to measure cognitive activation, the survey items that measure aspects of it exhibit a weak connection to cognitive activation as measured by observers through PLATO. In other words, the cognitive challenge that students perceive and report, i.e., opportunities to explain their thinking with the teacher and engaging in classroom discourse, seems to stem from other factors than intellectual challenge and content-related discourse as measured in classroom observations. Which factors influence those student experiences remains unanswered. For this study it has not been our expectation to find an overall explanation for the discrepancy, which can be the subject of future research. Among the dimensions of teaching measured by classroom observations, the results of this study suggest that for the cognitive activation dimension there is room for development in synthesizing instruments, such as the Tripod survey and PLATO, to measure its different aspects more accurately. Interpretations of results by these instruments, be it in research, professional development, or for educational accountability, should be made with care. Further inquiry into connections between specific dimensions, as observed by researchers and rated by students, will support more reliable and valid measures for understanding and improving instructional practice.

Acknowledgements

This work was supported by the QUINT Nordic Center of Excellence under NordForsk Grant 87663 through University of Oslo, Department of Teacher Education and School Research; and The Icelandic Research Fund under Grant 206754-051 through University of Iceland, School of Education.

REFERENCES

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180. https://doi.org/10.3102/0002831209345157

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Björnsdóttir, A., & Jónsdóttir, K. (2014). Viðhorf nemenda, foreldra og starfsmanna skóla. In G. G. Óskarsdóttir (Ed.) *Starfshættir í grunnskólum við upphaf 21. aldar* (pp. 29–56). Háskólaútgáfan.

Blikstad-Balas, M. (2017). Key challenges of using video when investigating social practices in education: contextualization, magnification, and representation. *International Journal of Research & Method in Education*, *40*(5), 511–523. https://doi.org/10.1080/1743727X.2016.1181162

Buchholtz, N., Klette, K., & Roe, A. (2020). Students' ratings of instructional quality and achievement in mathematics. In H. S. Siller, W. Weigel, & J. F. Wörler (Eds.), *Beiträge zum Mathematikunterricht 2020 auf der 54. Jahrestagung der Gesellschaft für Didaktik der Mathematik* (pp. 1165–1168). WTM-Verlag.

Clarke, D. (2006). Deconstructing dichotomies: Arguing for a more inclusive approach. In D. Clarke, J. Emanuelsson, E. Jablonka, & I. A. C. Mok (Eds.), *Making connections: Comparing mathematics classrooms around the world* (pp. 215–236). Sense Publishers.

de Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research*, *4*(1), 51–85. https://doi.org/10.1023/A:1011402608575

den Brok, P., Bergen, T., & Brekelmans, M. (2006). Convergence and divergence between students' and teachers' perceptions of instructional behaviour in Dutch secondary education. In D. Fisher & M. S. Khine (Eds.), *Contemporary approaches to research on learning environments: Worldviews* (pp. 125–160). World Scientific. https://doi.org/10.1142/9789812774651_0006

Ferguson, R. (2010). *Student perceptions of teaching effectiveness. Discussion brief from the National Center for Teacher Effectiveness and the Achievement Gap Initiative*. Harvard University.

Fraser, B. J. (1998). Science learning environments: assessment, effects and determinants. In B. J. Fraser & K. G. Tobin (Eds.), *The international handbook of science education* (pp. 527–564). Springer.

Fraser, B. J., Barry, J., & Walberg, H. J. (1981). Psychosocial learning environment in science classrooms: A review of research. *Studies in Science Education*, *8*(1), 67–92. https://doi.org/10.1080/03057268108559887

Granberg, C. (2016). Discovering and addressing errors during mathematics problem-solving: A productive struggle? *The Journal of Mathematical Behavior*, *42*(2), 33–48. https://doi.org/10.1016/j.jmathb.2016.02.002

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *43*(6), 293–303. https://doi.org/10.3102/0013189X14544542

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445–470. https://doi.org/10.1086/669901

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal*, *45*(1), 184–205. https://doi.org/10.3102/0002831207312906

Grossman, P. (2019). *Protocol for language arts teaching observations (PLATO 5.0)*. Stanford University. http://platorubric.stanford.edu/index.html

Gunnarsdóttir, G. H., & Pálsdóttir, G. (2015). Instructional practices in mathematics classrooms. In K. Krainer, & N. Vondrová (Eds.), *Proceedings of the Ninth Congress of the European Society for Research in Mathematics Education* (pp. 3036–3042). ERME.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Information Age.

Ing, M., Webb, N. M., Franke, M. L., Turrou, A. C., Wong, J., Shin, N., & Fernandez, C. H. (2015). Student participation in elementary mathematics classrooms: the missing link between teacher practices and student achievement? *Educational Studies in Mathematics*, *90*(3), 341–356. https://doi.org/10.1007/s10649-015-9625-z

Jewitt, C. (2012). *An introduction to using video for research*. NCRM Working Paper. (Unpublished).

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* [Research paper]. Bill & Melinda Gates Foundation.

Klette, K. (2009). Challenges in strategies for complexity reduction in video studies. Experiences from PISA+ Study: A video study of teaching and learning in Norway. In T. Janík & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 61–82). Waxmann Verlag.

Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal*, *17*(1), 129–146. https://doi.org/10.1177/1474904117703228

Klette, K., Blikstad-Balas, M., & Roe, A. (2017). Linking instruction and student achievement: A research design for a new generation of classroom studies. *Acta Didactica Norge*, *11*(3), 1–19. https://doi.org/10.5617/adno.4729

Krauss, S., Bruckmaier, G., Lindl, A., Hilbert, S., Binder, K., Steib, N., & Blum, W. (2020). Competence as a continuum in the COACTIV study: The "cascade model." *ZDM – Mathematics Education*, *52*(2), 311–327. https://doi.org/10.1007/s11858-020-01151-z

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod student survey. *Educational Assessment*, *22*(4), 253–274. https://doi.org/10.1080/10627197.2017.1381555

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. https://doi.org/10.1037/a0032583

Leikin, R., Levav-Waynberg, A., Gurevich, I., & Mednikov, L. (2006). Implementation of multiple solution connecting tasks: Do students' attitudes support teachers' reluctance? *Focus on Learning Problems in Mathematics*, *28*(1), 1–22.

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*(1), 61–95. https://doi.org/10.1007/s11092-018-09291-3

Luoto, J. M., Klette, K., & Blikstad-Balas, M. (2022). Possible biases in observation systems when applied across contexts: Conceptualizing, operationalizing and sequencing instructional quality. *Educational Assessment, Evaluation and Accountability*. https://doi.org/10.1007/s11092-022-09394-y

Mahan, K. R., Brevik, L. M., & Ødegaard, M. (2021). Characterizing CLIL teaching: new insights from a lower secondary classroom. *International Journal of Bilingual Education and Bilingualism*, *24*(3), 401–418. https://doi.org/10.1080/13670050.2018.1472206

Neubrand, M., Jordan, A., Krauss, S., Blum, W., & Löwen, K. (2013). Task analysis in COACTIV: Examining the potential for cognitive activation in German mathematics classrooms. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project* (pp. 125–146). Springer.

OECD. (2020). *Global Teaching InSights: A video study of teaching*. OECD. https://doi.org/10.1787/20d6f36b-en

Phillips, S. F., Ferguson, R. F., & Rowley, J. F. S. (2021). Do they see what I see? Toward a better understanding of the 7Cs framework of teaching effectiveness. *Educational Assessment*, *26*(2), 1–19. https://doi.org/10.1080/10627197.2020.1858784

Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings*. Waxmann Verlag.

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM – Mathematics Education*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Russo, J., & Hopkins, S. (2017). Student reflections on learning with challenging tasks: 'I think the worksheets were just for practice, and the challenges were for maths.' *Mathematics Education Research Journal*, *29*(3), 283–311. https://doi.org/10.1007/s13394-017-0197-3

Russo, J., & Hopkins, S. (2019). Teachers' perceptions of students when observing lessons involving challenging tasks. *International Journal of Science and Mathematics Education*, *17*(4), 759–779. https://doi.org/10.1007/s10763-018-9888-9

Sandilos, L. E., Sims, W. A., Norwalk, K. E., & Reddy, L. A. (2019). Converging on quality: Examining multiple measures of teaching effectiveness. *Journal of School Psychology*, 74(3), 10–28. https://doi.org/10.1016/j.jsp.2019.05.004

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *36*(3), 259–280. https://doi.org/10.3102/0162373713509880

Sigurgeirsson, I., Björnsdóttir, A., Óskarsdóttir, G., & Jónsdóttir, K. (2014). Kennsluhættir. In G. G. Óskarsdóttir (Ed.), *Starfshættir í grunnskólum við upphaf 21. aldar* (pp. 113–158). Háskólaútgáfan.

Sigurjónsson, J. Ö., & Gísladóttir, B. (2020). Vitsmunaleg áskorun í stærðfræðikennslu á unglingastigi. *Tímarit Um Uppeldi Og Menntun*, 29(2), 149–171. https://doi.org/10.24270/tuuom.2020.29.8

Sigurjónsson, J. Ö., & Kristinsdóttir, J. V. (2018). Upprifjunaráfangar framhaldsskóla í stærðfræði: Skapandi og krefjandi vinna eða stagl? *Tímarit Um Uppeldi Og Menntun*, 27(1), 65–86. https://doi.org/10.24270/uuom.2018.27.4

Sigþórsson, R., Pétursdóttir, A.-L., & Jónsdóttir, Þ. B. (2014). Nám, þátttaka og samskipti nemenda. In G. G. Óskarsdóttir, (Ed.), *Starfshættir í grunnskólum við upphaf 21. aldar* (pp. 161–196). Háskólaútgáfan.

Snell, J. (2011). Interrogating video data: Systematic quantitative analysis versus micro-ethnographic analysis. *International Journal of Social Research Methodology*, *14*(3), 253–258. https://doi.org/10.1080/13645579.2011.563624

Tekkumru-Kisa, M., Stein, M. K., & Doyle, W. (2020). Theory and research on tasks revisited: Task as a context for students' thinking in the era of ambitious reforms in mathematics and science. *Educational Researcher*, *49*(8), 606–617. https://doi.org/10.3102/0013189X20932480

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal*, *53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

White, M. C. (2018). Rater performance standards for classroom observation instruments. *Educational Researcher*, 47(8), 492–501. https://doi.org/10.3102/0013189X18785623

Wilhelm, A. G. (2014). Mathematics teachers' enactment of cognitively demanding tasks: Investigating links to teachers' knowledge and conceptions. *Journal for Research in Mathematics Education*, *45*(5), 636–674. https://doi.org/10.5951/jresematheduc.45.5.0636

Xu, L., Aranda, G., Widjaja, W., & Clarke, D. (2019). *Video-based research in education*. Routledge.

Þórðardóttir, Þ., & Hermannsson, U. (2012). *Úttekt á stærðfræðikennslu á unglingastigi grunnskóla*. Mennta- og menningarmálaráðuneyti.