Investigating differentiation: exploring the impact of task difficulty labelling on students' mathematics performance

MARIA HERSET, MOHAMED EL GHAMI AND ANNETTE HESSEN BJERKE

This study investigated the effect of task difficulty labelling on students' mathematics performance. Lower secondary school students ($n = 436$) participated in an experimental study in which the control group received three similar tasks without labels and the experimental groups received the same tasks labelled as *easy, medium* and *difficult*. The findings indicate that the students in the control group outperformed the students assigned tasks labelled as difficult. Additionally, the students' performance was worse when solving tasks labelled difficult compared to when they were given similar tasks labelled easy or medium, but the differences are not significant. We strongly advise exercising prudence when assigning tasks labelled difficult.

How best to differentiate teaching in heterogeneous classrooms seems to be a recurring concern, engaging teachers and researchers across the world (Pozas et al., 2020). In this regard, a promising and often used approach in mathematics classrooms is differentiated instruction (Pierce & Adams, 2005), which is a flexible way of adapting and adjusting teaching to meet students at their levels in a way that helps them achieve maximum growth (Tomlinson, 2014). In mathematics, where solving tasks (in one way or another) constitutes a primary activity, differentiation based on a student's existing proximity to specific knowledge, comprehension and skills, commonly referred to as readiness (Tomlinson & Imbeau, 2010), stands out as a recognized method for implementing differentiated instruction. This method, called tiering, involves adjusting tasks according to students' readiness levels to give them opportunities to draw on past experiences. Mathematics teachers have the option to assign varied tasks according to assessed readiness levels (Smale-Jacobse et al., 2019) or to present multiple versions of a learning task through a

Maria Herset, *Nord University*
Mohamed El Ghami, *Nord University*
Annette Hessen Bjerke, *OsloMet*

Herset, M., El Ghami, M. & Bjerke, A. H. (2025). Investigating differentiation: exploring the impact of task difficulty labelling on students' mathematics performance. *Nordic Studies in Mathematics Education*, 30 (1), 5–24.

tiered approach. Such an approach often involves categorizing task variations into three levels: below grade level, at grade level and above grade level (Pierce & Adams, 2005). These tiers are subsequently designated through appropriate labelling.

The research literature addressing this (e.g. Bal, 2016; Luster, 2008; Scott, 2012; Suarez, 2007) claims that adapting tasks based on individual students' readiness levels enhances mathematics performance. This approach is extensively utilized not only in Norwegian primary and secondary education but also in other countries, evident in both classroom practices (Eriksen et al., 2022; Grave & Pepin, 2015; Suarez, 2007) and in the content of mathematics textbooks (Auliya & Widjajanti, 2023; Brändström, 2005; Grave & Pepin, 2015; Kristensen, 2008; Olafsen & Maugesten, 2022).

Labelling mathematics tasks according to difficulty has deep roots (see e.g. Winther, 1965) and has endured despite facing criticism from various quarters (Herset et al., 2023). For instance, Botten et al. (2008) questioned whether labelling tasks according to difficulty in mathematics "has resulted in greater learning outcomes for all students" (p. 24, authors' translation), while Kristensen (2008) feared that labelling limits students' ambitions. However, recent studies show a persistent usage of textbooks in mathematics (e.g. Dolonen et al., 2016), in which easy, medium and difficult labels flourish (Herset et al., 2023; Kristensen, 2008; Mathiassen, 2009; Olafsen & Maugesten, 2022). More research is needed on how differentiated instruction links to outcomes (Smale-Jacobse et al., 2019).

Previous research states that the impact of revealing the *true* difficulty levels on students' performance is underexplored (Spielberg & Azaria, 2021). Spielberg and Azaria (2021) examined the effect of revealing the task difficulty on students' performance. The results indicated that there were no significant differences when comparing the number of correct answers between the control group (given tasks with no labels) and the experimental group (true labels). Moreover, since previous research found that a selection of tasks labelled difficult in mathematics textbooks are too easy (Brändström, 2005), and mathematics tasks which are divided into two levels, are too similar (Singh, 2017), we find it timely to investigate how labelling the same tasks with different labels (easy, medium and difficult) influences students' performance in mathematics. Hence, in this paper, we put forward the following research question: How does the labelling of mathematics tasks according to easy, medium and difficult labels affect students' performance?

We believe an examination of this research question will expose how easy, medium and difficult labels on mathematics tasks contribute to achieving one of the aims of differentiation, namely increased performance.

Literature review

The primary intention behind labelling mathematics tasks according to difficulty levels appears to be to facilitate students' engagement with tasks that align with their individual prerequisites (Kristensen, 2008). This approach, in turn, is believed to foster an optimal learning experience and cultivate a heightened sense of mastery in individuals (Mathiassen, 2009). In this regard, since students' experience of mastery is found to be a key measure (Skaalvik & Skaalvik, 2015), it is crucial that mathematics teachers use tasks adapted to students' readiness levels (Csikszentmihalyi, 2005; Pierce & Adams, 2005). This is supported by research reporting that self-efficacy – the "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura, 1997, p. 3) – is improved by differentiated instruction (e.g. Lai et al., 2020; Onyishi & Sefotho, 2021). This might be because experience of mastery fosters individuals' self-efficacy (e.g. Butz & Usher, 2015; Joët et al., 2011; Usher & Pajares, 2009), which again is predictive of students' performance (e.g. Chen, 2003; Doménech-Betoret et al., 2017; Özcan & Eren Gümüş, 2019; Pajares & Kranzler, 1995; Pajares & Miller, 1994; Skaalvik et al., 2015).

As Ghalem et al. (2016) highlight, performance is a multifaceted construct that depends on the specific context. Notably, its definition is frequently absent from the literature, with explanations often arising indirectly through the delineation of measurement methodologies. This tends to work since performance in mathematics is often assessed through tests and examinations. We follow Zakariya (2022) and see performance as "students' examination scores or grades in mathematics courses that they followed" (p. 3).

Mathematics performance is influenced by a number of factors, such as students' intelligence, self-esteem, personality traits (Zuffianò et al., 2013), mathematics self-concept (Pajares & Miller, 1994), mathematics anxiety (Özcan & Eren Gümüş, 2019; Pajares & Kranzler 1995; Pajares & Miller, 1994), mental ability (Pajares & Kranzler 1995) and self-efficacy (Özcan & Eren Gümüş, 2019; Pajares & Kranzler, 1995; Pajares & Miller, 1994, 1995; Zakariya, 2021), as well as how each student perceives the usefulness of mathematics (Pajares & Miller, 1994). Through this list, research shows that self-efficacy is an important predictor of performance (Özcan & Eren Gümüş 2019; Pajares & Kranzler, 1995; Pajares & Miller, 1994; Zakariya, 2021; Zuffianò et al., 2013). Here, we conceptualize mathematics self-efficacy as "a situational or problem specific assessment of an individual's confidence in her or his ability to successfully perform or accomplish a particular task or problem" (Hackett & Betz, 1989, p. 262). Collins (1984) confirmed the positive effect of high

mathematics self-efficacy on performance, and more recent studies have also found a positive relationship between students' mathematics self-efficacy and performance in mathematics (Schöber et al., 2018) and scores on national tests (Street et al., 2017).

For several decades, tiered mathematics lessons according to readiness have been highlighted as an important tool for differentiating the subject matter. Little et al. (2009) found tiering useful, as it enables the teacher to provide "challenging tasks while ensuring sufficient scaffolding for struggling students and reducing repetition for more advanced students" (p. 36). However, caution is advised: on the one hand, if the task is too easy, students might get bored and lose concentration and energy levels, and on the other hand, if the task is too hard, students might become frustrated and experience anxiety (Csikszentmihalyi, 2005). For students to achieve an optimal learning effect and experience mastery, tasks must be adapted so that they feel joy and commitment (Csikszentmihalyi, 2005). Skaalvik and Skaalvik (2015) pointed out that it is important that students complete tasks, since experience of mastery has been shown to be necessary for students to develop and maintain expectations of mastery. Students' experience of working with tasks adapted to their readiness will allow them to grow in their learning, as they get the opportunity to work with tasks that challenge them (Little et al., 2009).

Previous research on tiered teaching techniques in which lessons were planned according to students' learning styles and levels of readiness found that students taught with differentiated teaching methods had higher mathematical success than those who did not (Bal, 2016; Luster, 2008). In addition, Scott (2012) found that differentiated instruction improved student performance but only for students with above-average abilities. However, more critical voices have proposed that instead of differentiation initiatives, the focus should be on students' well-being and productive disposition rather than cognitive performance outcomes, as this may result in "labelling" of the students (Anthony et al., 2019). Labelling students as having "less developed readiness" can contribute to social and structural inequalities (Bannister, 2016, p. 345).

In the literature on tiering by readiness, the focus tends to be on the effect on students' performance of giving them different mathematics tasks according to readiness (Bal, 2016; Luster, 2008; Scott, 2012; Suarez, 2007). In an experiment Bal (2016) conducted, students' learning styles and readiness levels were determined before applying the tiered teaching technique. The students were divided into two groups (with low and medium readiness) and assigned tasks according to their readiness level. Bal (2016) found that the tiered teaching technique increased students' performance in mathematics. However, it is not clear whether the

experiment included tasks labelled according to readiness or whether the students received information about the difficulty of the tasks. Suarez (2007), who used labelling of tasks in their study, found that students were more motivated to participate when the teacher labelled mathematics tasks according to readiness. A higher motivation occurred as the teacher(s) applied tiered instruction by labelling each mathematics task with three levels of mastery, and the students were told to select tasks according to which level of challenge they felt was appropriate (Suarez, 2007). Another study reported that homework differentiated according to readiness influenced students' attitudes (Keane & Heinz, 2019). In Keane and Heinz's (2019) study, the students did not receive any information about the difficulty level (easy, medium or difficult). Each week, the students were told to choose one of three tasks, which eventually had a positive impact on homework engagement. In addition, Keane and Heinz (2019) found that the students' choice of tasks was appropriate for their level.

We add to this body of research not by investigating students' performance when solving tasks that match their readiness levels (as the reviewed literature does), but rather by examining how different labels (easy, medium and difficult) on similar mathematics tasks affect students' mathematics performance. An important argument, in addition to the fact that labels are not always correct in textbooks (Brändström, 2005; Singh, 2017), is that Krauthausen (2018) notes that a task's level of difficulty is a subjective evaluation. Based on this, we think it would be problematic to investigate a "true" difficulty level, as students' assessments of the level of difficulty may vary between students. In addition, students' opinions of a task's difficulty level can change during the school year and even during the day (Krauthausen, 2018).

Elsewhere, we report that the time students spent solving a task was significantly shorter when a task was labelled as difficult compared to when the same task was not labelled (Herset & El Ghami, 2022). We also found that labelling easy mathematics tasks as difficult had a negative effect on students' self-efficacy (Herset et al., 2023). Since perceived task difficulty in mathematics affects students' self-efficacy (Liu et al., 2020; Street et al., 2017, 2022), we assert that it is important to study how labelling tasks as easy, medium and difficult affects students' performance in mathematics.

Materials and methods

In this paper, we draw on data collected for a larger project through an online questionnaire distributed to Norwegian lower secondary students.

Since the population is large and widely dispersed, we used cluster sampling (Cohen et al., 2018) by randomly selecting schools across Norway and contacting school principals from the list of schools. Because of the Covid-19 pandemic and school lockdowns, we included 23 schools whose principals agreed to take part. To ensure that the data were collected in the same way in all schools, the students had to respond to the survey during class, while the mathematics teachers ensured that the students worked individually without calculators. Hence, the students were required to be in class. The final sample included 349 students: 172 girls (49 %) and 177 boys (51 %).

The online survey given to these 13–15-year-old students enabled the use of three research designs: a pretest–posttest control-group design, a posttest-only control-group design (Creswell & Creswell, 2018) and a repeated-measures design (Cohen et al., 2018). The same survey was administered to all students, and the online survey randomly assigned them to different control and experimental groups. Therefore, the students' location, school, grade level and gender did not affect their group allocation. The survey included no more than 11 tasks (and some additional questions on, for example, age, gender and self-reported level of self-efficacy and effort). The control group received all 11 tasks without labels for the tasks, and the experimental groups were given the same tasks labelled as easy, medium or difficult. The survey was carefully set up with the aim of exploring the effect of labelling mathematics tasks as easy, medium and difficult on students' performance, self-efficacy, preferences (when it comes to mathematics tasks) and persistence (in solving tasks). Here, we focus on 3 of the 11 tasks while exploiting the repeated-measures design and the posttest-only control-group design of the survey. The overall project was approved by the Norwegian Social Science Data Service, and we followed its ethical guidelines.

Table 1. *The three selected tasks (authors' translations)*

Task A	In Barcelona, you find the not-yet-completed church known as the Sagrada Família. They started building it in 1882, and it was supposed to be finished in 2026. How many years do they expect it will take to build the Sagrada Família?
Task B	Rita is on holiday in Greece. She wants to rent a scooter. It costs NOK 25 per 5 minutes. How much does it cost to rent the scooter for 1 hour?
Task C	Silja wants to take a swimming test. To do that, she has to swim 200 m without taking a break. The length of the pool is 12.5 m. How many lengths does Silja have to swim?

Note. Tasks A–C are similar and are all at mastery level 2 of 5 (i.e. 70 % of the students are expected to solve the tasks correctly).

Measures

We measured the students' performance by assessing their answers to three similar tasks (tasks A–C; see table 1). The definition of performance (drawing on Zakariya, 2022) allowed us to see the performance simply as *correct or incorrect solutions to tasks A–C*. Performance was assessed as correct numerical answers to the given mathematics tasks, while both empty answers and calculation errors were assessed as incorrect solutions. Persistence, effort and strategy were not included as measures of performance.

These tasks were retrieved from the 2020 national test in mathematics. The tasks were selected because of their similar difficulty level (mastery level 2 out of 5; see Björnsson, 2016), their similar underpinning mathematical theme (arithmetic and algebra) and similar word length and layout. The repeated-measures design provided several observations from the same student (Cohen et al., 2018); by choosing similar tasks, similar student behaviour was expected when they engaged with the tasks.

Data collection

The main idea behind the design was that all the participants would be given tasks A, B and C, which were all assessed as equally difficult and labelled with varying difficulty levels (easy, medium or difficult). In addition, Creswell and Creswell (2018) require participants to be randomly assigned to different groups and subgroups to conduct a true experiment.

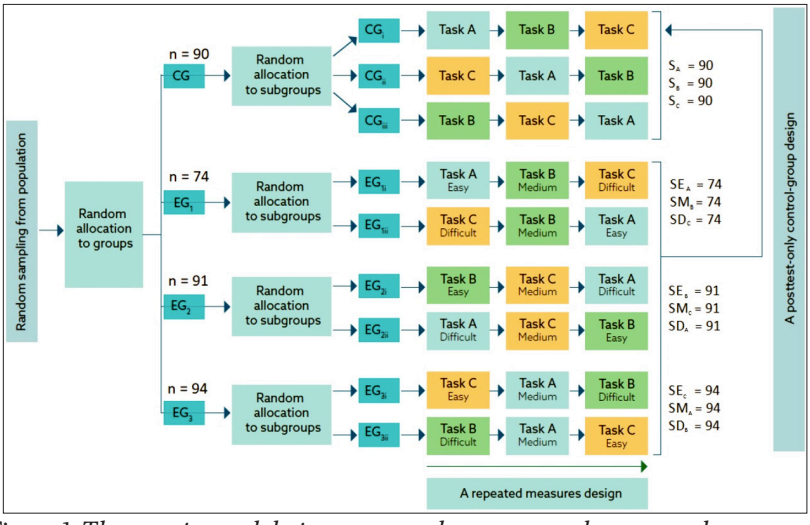

Figure 1. *The experimental design: a repeated-measures and posttest-only control-group design based on the three tasks*

Figure 1 shows how the participants were first randomly assigned into four groups when opening the online survey, either to the control group (CG) or to one of three experimental groups (EG_1, EG_2 or EG_3). Next, they were assigned to subgroups. This was necessary to avoid a treatment effect (Cohen et al., 2018). In the subgroups, the order of the tasks and labels was controlled by changing the order of the easy and difficult labelling for the different subgroups (see figure 1). For example, the students assigned to subgroup EG_{1i} first received task A, labelled easy, and received last task C, labelled difficult. For EG_{1ii}, the order of tasks A and C and the labels were reversed. Since the medium label was assumed to have the smallest treatment effect, it was always given as the second task. The reason for including a CG in the design (see the top of figure 1) was to ensure that the participants perceived tasks A–C (when not labelled) to be equally difficult. We used the related-sample Cochran's Q test to compare the differences between performance on tasks A–C. As expected, there were no significant differences when comparing the repeated measures of the CG performance on tasks A–C ($Q = 1.135$, $df = 2$, $p = .567$) (remember that tasks A–C were assumed to be similar, i.e. the same mathematical topic, difficulty level, word length and layout). Furthermore, to avoid a floor or ceiling effect (Everitt, 2002), tasks at mastery level 2 from the national test in mathematics were assumed to be an appropriate difficulty level, since 70 % of Norwegian students are expected to complete tasks at mastery level 2 (Björnsson, 2016). Thus, we anticipated that approximately 70 % of the students would solve tasks A–C correctly when encountering them without labels. To validate this, we used a binomial test by comparing the performance on tasks A–C to the binomial assumption of 70 %. Our analysis showed that 68 % of the CG solved tasks A–C correctly. According to the binomial test, the frequency distribution of the sample corresponded to the population ($p = .231$). Hence, we could exploit the post-test-only control-group design (Creswell & Creswell, 2018) that allowed us to examine the effect on students' performance of labelling tasks as easy, medium and difficult by comparing the performance by the EGs (tasks A–C with labels) with the performance by the CG (tasks A–C without labels).

Analysis

To answer our research question, we exploited two designs of the survey, as illustrated in figure 1, and proposed the following two hypotheses, both grounded in the reviewed literature:

H1 Labelling mathematics tasks as easy, medium and difficult affects students' performance.

H2 When solving three similar mathematics tasks, students' performance decreases with increasing task difficulty labelling.

To examine H1, we first merged the solutions from the three EGs to all the tasks labelled as easy. From EG_1, there were 74 solutions to task A labelled easy (solution easy task A = SE_A = 74), from EG_2 there were 91 solutions to task B labelled easy (SE_B = 91) and from EG3, there were 94 solutions to task C labelled easy (SE_C = 94). Thus, we had 259 answers to the tasks labelled easy. These numbers are given on the right in figure 1. We repeated this for the solutions stemming from all the tasks labelled medium and those labelled difficult, leaving us with 259 medium solutions and 259 difficult solutions. In the same manner, we had 270 answers to the unlabelled tasks in the CG (90 from task A [S_A = 90], 90 from task B [S_B = 90] and 90 from task C [S_C = 90]). We used a chi-squared test to compare the CG's solutions with those from the EGs. A repeated-measures design involves "multiple observations of a single individual" (Creswell & Creswell, 2018, p. 169). We exploited this when examining H2, since the aim was to test three experimental conditions: easy, medium and difficult labels. The three measurement points were the students' solutions to tasks A–C labelled easy, medium and difficult. This design had no control condition; that is, the students in the experimental groups did not solve any tasks without labels. However, according to Cohen et al. (2018), a repeated-measures design has considerable potential for control. This is justified by the fact that the same students received the three labels. As mentioned, since the first labels may influence the second or third labels, we randomized the order of the labels, as Cohen et al. (2018) recommend. To test H2, we used a related-sample Cochran's Q test followed by pairwise comparison to examine the extent to which easy, medium and difficult labels on tasks A–C affected a single student's correct solutions.

Results

The students' results on tasks A–C are provided in table 2, revealing that 69 % of the students in the CG (no labels) solved task A correctly, while 59 % of those receiving tasks A labelled as easy answered correctly. The corresponding numbers for the medium label were 65 % and 59 % for the difficult label. The column to the right in table 2 shows, for example, that these measures correspond to a reduction of 14 % in correct answers when

Table 2. *Performance on tasks A–C*

Task	Label	*n*	Correct solutions (%) (CS)	Effect of labelling tasks on students' solutions $\frac{CS_{EG}-CS_{CG}}{CS_{CG}}$
A	Easy	94	58.5 %	-15 %
	Medium	74	64.9 %	-6 %
	Difficult	91	59.3 %	-14 %
	No label	90	68.9 %	0 %
B	Easy	91	74.7 %	7 %
	Medium	94	70.2 %	0 %
	Difficult	74	55.4 %	-21%
	No label	90	70.0 %	0 %
C	Easy	74	62.2 %	-4%
	Medium	91	60.4 %	-6 %
	Difficult	94	56.4 %	-13 %
	No labels	90	64.4 %	0 %

Note. Even if the tasks are labelled with levels of difficulty (easy, medium or difficult) or not labelled with level, the tasks are similar and are all at mastery level 2 of 5.

comparing the results from the group that received task A labelled difficult (59 %) compared to those in the CG who received task A without labels (69 %). The results for tasks B and C are presented in table 2 in the same way.

Further, we measured the effect of easy, medium and difficult labels regardless of whether the task was A, B or C. Table 3 shows the proportion of correct answers in the EGs when the tasks were labelled easy, medium or difficult and the proportion of correct answers in the CG when the same tasks were not labelled. For example, 57 % of the students in the EGs correctly solved a task with a difficult label, while when the CG received the same tasks, 68 % of the solutions were correct.

Table 3. *Performance on tasks A–C*

Group	*n*	Task	Number of solutions	Label	Performance
EG	259	A–C	259	Easy	65.3 %
EG	259	A–C	259	Medium	65.3 %
EG	259	A–C	259	Difficult	57.1 %
CG	90	A–C	270	No labels	67.8 %

Note. Tasks A-C are similar and at mastery level 2 of 5 (i.e. 70 % of the students are expected to solve the tasks correctly).

To test H1, we conducted a chi-squared test (Cohen et al., 2018). Since we had two categorical variables and the data were bivariate, the chi-squared test was a test of independence (Cohen et al., 2018).

The test results (see table 4) revealed that the proportion of correct answers was significantly lower when the tasks were labelled difficult compared to when they were not labelled ($\chi^2 = 6.38$, $df = 1$, $p = .007$). The phi coefficient was 0.11, which is a small effect size (Cohen et al., 2018). No such significant differences were found when comparing the easy label and no labels or between the medium label and no labels.

To test H2, we used the related-sample Cochran's Q test (Cohen et al., 2018). As table 5 shows, the result was statistically significant ($Q = 6.125$, $df = 2$, $p = .047$). Follow-up pairwise comparisons using Cochran's Q test showed that a significantly lower proportion of participants solved a task correctly when it was labelled difficult compared to when a similar task was labelled medium ($p = .032 < .05$) and when it was labelled difficult compared to easy ($p = .032 < .05$). However, a Bonferroni correction (adjusted significant value, $p < .017$) for multiple tests showed that the p-value was not significant ($p = .032 > .017$).

Table 4. *Observed frequencies of the level labelling and results*

Label	Results		χ2	*df*	*p*
	Correct	Wrong			
Easy	169	90	0.38	1	.300
No labels	183	87			
Medium	169	90	0.38	1	.300
No labels	183	87			
Difficult	148	111	6.38	1	.007*
No labels	183	87			

Note. Even if the tasks are labelled as easy, medium or difficult, the tasks are similar and are identical to the tasks with no label. * The *p*-value was significant at the .01 level.

Table 5. *Observed frequencies of level labelling and results*

Label	Results		χ2	*df*	*p*
	Correct	Wrong			
Easy	169	90	6.125	2	.047*
Medium	169	90			
Difficult	148	111			

Note. Even if the tasks are labelled as easy, medium or difficult, the tasks are similar and are all at mastery level 2 of 5. * The *p*-value was significant at the .05 level.

Discussion and concluding remarks

A strand within research on teaching in heterogeneous classrooms focuses on differentiated instruction (Pierce & Adams, 2005), often in relation to readiness (or ability) (Tomlinson, 2014), using tiering strategies (Pierce & Adams, 2005). By viewing labelling of mathematics tasks as a tiering strategy, we assert that the results we report here contribute to shedding new light on the well-established practice of labelling tasks according to their level in classroom teaching (Eriksen et al., 2022; Grave & Pepin, 2015) and in mathematics textbooks (Grave & Pepin, 2015; Kristensen, 2008; Mathiassen, 2009; Olafsen & Maugesten, 2022).

Labelling mathematics tasks in textbooks according to the level of difficulty has deep roots (Winther, 1965) and still flourishes (Olafsen & Maugesten, 2022), despite critical voices (Botten et al., 2008; Kristensen, 2008). In the current study, we found that labelling mathematics tasks as difficult had a significant negative effect on students' performance: students who encountered tasks labelled difficult (even if they were not) got a lower proportion of correct solutions (57 %) than the students who received the same tasks without labels (68 %). In addition, we found that the proportion of correct solutions was lower when students tackled tasks labelled difficult (57 %) compared to when they were given similar tasks labelled easy (65 %) or medium (65 %). However, these results were not significant.

Csikszentmihalyi's (2005) flow model can explain why the students' performance became significantly lower when the tasks were labelled difficult compared to when the tasks were not labelled. In our case, students in the EGs may have thought that the tasks were difficult and hence became more frustrated and anxious compared to the students in the CG, who received tasks without labels. This raises an evaluative question (which perhaps adds to the critical voices): "Is it necessary to label mathematics tasks according to difficulty levels?". Based on our results, the answer is no, first and foremost because our analysis showed no effect when easy tasks were labelled easy. This result is in line with Spielberg and Azaria (2021), who found that revealing the difficulty level of easy tasks did not affect students' performance. In addition, our analysis uncovered something new: a misleading label may go against the initial intention of using such a differentiation initiative (which is to improve students' performance in mathematics). Hence, we assert that we should be careful when labelling tasks according to their difficulty levels. Since our results show that none of the labels positively affected the students' performance, we believe that teachers and textbook authors should be careful about labelling tasks according to difficulty level – especially since our results show that labelling easy tasks as difficult decreased students'

performance. Another argument is that declaring a mathematics task easy, medium or difficult is not an objective concern (Krauthausen, 2018), meaning that some students might think a mathematics task is difficult and others not.

Nevertheless, we would like to point out that it is crucial that mathematics teachers use tasks adapted to students' readiness levels (Csikszentmihalyi, 2005; Pierce & Adams, 2005), especially since researchers seem to agree that self-efficacy is improved by differentiated instruction (e.g. Lai et al., 2020; Onyishi & Sefotho, 2021). Herset et al. (2023) found that labelling tasks difficult has a negative effect on students' self-efficacy. Coupled with the results we report here, which assert that a difficult label has a negative effect on students' performance, it is tempting to propose a model in which self-efficacy has a mediator role. A mediation model could be that a difficult label (independent variable) negatively affects students' self-efficacy (mediator variable), which in turn decreases students' performance (dependent variable). This model is supported by Collins (1984), who found that mathematics ability has an indirect effect, via mathematics self-efficacy, on mathematics performance. Remember, experiences of mastery were found necessary for students to develop and preserve expectations of mastery (Skaalvik & Skaalvik, 2015). More research is needed to confirm (or reject) the proposed mediation model.

Taken together, when attempting to answer the research question on how easy, medium and difficult labels on mathematics tasks affect students' performance, we responded to a call from Smale-Jacobse et al. (2019), who systematically reviewed research publications on differentiated instruction. They saw the need for more research into how differentiated instruction is linked to outcomes (i.e. performance). By comparing a CG with EGs, we found that when mathematics tasks were labelled difficult, this had a statistically significant negative effect on performance. Moreover, when mathematics tasks were labelled easy or medium, this had no positive (or negative) effect on performance. Hence, we assert that this finding adds the labelling of tasks (as a tiering strategy) to the long list of factors that influence mathematics performance.

While this study's results contribute to a new understanding of the effects of using tasks labelled as easy, medium and difficult in mathematics as a tiering strategy, the study has some limitations. We used a narrow definition of performance (drawing on Zakariya, 2022) that allowed us to see performance simply as *correct* or *wrong solutions*. This led us to exclude, for instance, students' written work (e.g., their preferred strategies), persistence and effort. A broader focus on measuring performance would have strengthened our findings. We also limited ourselves to one topic. It may be that labels do not affect students' solutions in other

mathematical topics. Moreover, our choice to use cluster sampling makes it hard to generalize our results. On the other hand, we used 23 different schools, and, according to Cohen et al. (2018), it is safer to use several clusters than fewer heavy-sampled clusters.

Based on the limitations addressed above, we suggest that more research is needed to fully understand the relationship between the effect of labelling tasks as easy, medium and difficult and students' performance. We suggest this can be done by examining how labels affect the students' performance by focussing on the numbers of blank answers. This is justified by the fact that the students who were given tasks labelled difficult may have thought that their immediate solution idea sounded too simple, which may have led to blank answers. In addition, further research should examine how labels affect mediating factors (e.g. self-efficacy), and how the effect may vary based on students' ability levels, motivation and gender differences. We also suggest further research using other tasks within different topics. The reason we found no positive significant effect on students' performance when tasks were labelled easy may be because we chose to use rather easy tasks in this study (mastery level 2 out of 5; see Björnsson, 2016). We therefore suggest more research using tasks with different levels of difficulty (e.g. investigating the effect of labelling difficult tasks as easy). In addition, Suarez (2007) made an important point when asserting that students are more motivated to participate when they can choose among mathematics tasks labelled according to three mastery levels. As it is not clear whether it was the labelling of the task that motivated the students or whether it was because they were allowed to choose between tasks, we suggest more research on the relationship between labels and motivation in mathematics.

Acknowledgements

The datasets are available from the first author on reasonable request and after approval of the Norwegian social science data service and extended ethical approval. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. All authors listed made substantial, direct and intellectual contributions to the work and approved it for publication. This research received no external funding.

References

Anthony, G., Hunter, J. & Hunter, R. (2019). *Working towards equity in mathematics education: Is differentiation the answer?* Mathematics Education Research Group of Australasia.

Auliya, K. & Widjajanti, D. B. (2023). Singaporean and Japanese maths textbooks: character, structure, and content. *Mosharafa. Jurnal Pendidikan Matematika*, 12 (1), 155–168. https://doi.org/10.31980/mosharafa.v12i1.764

Bal, A. P. (2016). The effect of the differentiated teaching approach in the algebraic learning field on students' academic achievements. *Eurasian Journal of Educational Research*, 16 (63), 185–204. http://dx.doi.org/ 10.14689/ejer.2016.63.11

Bandura, A. (1997). *Self-efficacy: the exercise of control*. W.H. Freeman.

Bannister, N. A. (2016). Breaking the spell of differentiated instruction through equity pedagogy and teacher community. *Cultural Studies of Science Education*, 11 (2), 335–347. https://doi.org/10.1007/s11422-016-9766-0

Björnsson, J. K. (2016). *Metodegrunnlag for nasjonale prøver.* Utdanningsdirektoratet. https://www.udir.no/globalassets/filer/vurdering/nasjonaleprover/metodegrunnlag-fornasjonale-prover-august-2018.pdf

Botten, G., Daland, E. & Dalvang, T. (2008). Tilpasset opplæring innenfor fellesskapet. *Tangenten*, 19 (2), 23–27.

Brändström, A. (2005). *Differentiated tasks in mathematics textbooks. An analysis of the levels of difficulty* [Unpublished licentiate thesis]. Luleå University of Technology.

Butz, A. R. & Usher, E. L. (2015). Salient sources of early adolescents' self-efficacy in two domains. *Contemporary Educational Psychology*, 42, 49–61. https://doi.org/10.1016/j.cedpsych.2015.04.001

Chen, P. P. (2003). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences*, 14 (1), 77–90. https://doi.org/ 10.3200/JEXE.75.3.221-244

Cohen, L., Manion, L. & Morrison, A. K. (2018). *Research methods in education*. Routledge.

Collins, J. L. (1984). *Self-efficacy and ability in achievement behavior* [Unpublished doctoral thesis]. Stanford University.

Creswell, J. W. & Creswell, J. D. (2018). *Research design: quantitative, qualitative and mixed methods*. Sage.

Csikszentmihalyi, M. (2005). *Flow og engagement i hverdagen*. Dansk Psykologisk Forlag.

Dolonen, J. A., Furberg, A., Gilje, O., Ingulfsen, L., Kluge, A. et al. (2016). *Med ARK&APP. Bruk av læremidler og ressurser for læring på tvers av arbeidsformer.* University of Oslo.

Doménech-Betoret, F., Abellán-Roselló, L. & Gómez-Artiga, A. (2017). Self-efficacy, satisfaction, and academic achievement: the mediator role of students' expectancy-value beliefs. *Frontiers in Psychology*, 8, 1193. https://doi.org/10.3389/fpsyg.2017.01193

Eriksen, E., Solomon, Y., Bjerke, A. H., Gray, J. & Kleve, B. (2022). Making decisions about attainment grouping in mathematics: teacher agency and autonomy in Norway. *Research Papers in Education*, 1–21. https://doi.org/10.1080/02671522.2022.2135014

Everitt, B. S. (2002). *The Cambridge dictionary of statistics*. Cambridge University Press.

Ghalem, Â., Okar, C., Chroqui, R. & Semma, E. (2016). Performance: a concept to define. *Logistica*, 1–13. https://doi.org/10.13140/RG.2.2.24800.28165

Grave, I. & Pepin, B. (2015). Teachers' use of resources in and for mathematics teaching. *Nordic Studies in Mathematics Education*, 20 (3-4), 199–222.

Hackett, G. & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education*, 20 (3), 261–273. https://doi.org/10.2307/749515

Herset, M. & El Ghami, M. (2022). *The effect of level-marking mathematical tasks on students' time spent on such tasks and correct solutions: an experimental study*. CERME12. https://hal.science/hal-03745597v1/file/TWG08_06_Herset.pdf

Herset, M., El Ghami, M. & Bjerke, A. H. (2023). The effect of level-marked mathematics tasks on students' self-efficacy: an experimental study. *Frontiers in Psychology*, 14. https://doi.org/10.3389/fpsyg.2023.1116386

Joët, G., Usher, E. L. & Bressoux, P. (2011). Sources of self-efficacy: an investigation of elementary school students in France. *Journal of Educational Psychology*, 103 (3), 649–663. https://doi.org/10.1037/a0024048

Keane, G. & Heinz, M. (2019). Differentiated homework: impact on student engagement. *Journal of Practitioner Research*, 4 (2), 1–23. https://doi.org/10.5038/2379-%20%099951.4.2.1111

Krauthausen, G. (2018). Natural differentiation – an approach to cope with heterogeneity. In G. Kaiser, H. Forgasz, M. Graven, A. Kuzniak, E. Simmt, & B. Xu (Eds.), *Invited lectures from the 13th international congress on mathematical education* (pp. 325–341). Springer.

Kristensen, T. E. (2008). Tilpasset opplæring innenfor fellesskapet. *Tangenten*, 19 (2), 9–14.

Lai, C. P., Zhang, W. & Chang, Y. L. (2020). Differentiated instruction enhances sixth-grade students' mathematics self-efficacy, learning motives, and problem-solving skills. *Social Behavior and Personality: an international journal*, 48 (6), 1–13. https://doi.org/10.2224/sbp.9094

Little, C. A., Hauser, S. & Corbishley, J. (2009). Constructing complexity for differentiated learning. *Mathematics Teaching in the Middle School*, 15 (1), 34–42.

Liu, Q., Liu, J., Cai, J. & Zhang, Z. (2020). The relationship between domain- and task-specific self-efficacy and mathematical problem posing: a large-scale study of eighth-grade students in China. *Educational Studies in Mathematics*, 105 (3), 407–431. https://doi.org/10.1007/s10649-020-09977-w

Luster, R. (2008). *A quantitative study investigating the effects of whole-class and differentiated instruction on student achievement* [Unpublished doctoral thesis]. Walden University.

Mathiassen, K. (2009). Lektor – adjunkt – lærer: artikler for studiet i praktisk-pedagogisk utdanning. In I R. Mikkelsen & H. Flademoe (Eds.), *Differensiert undervisning* (pp. 123–136). Universitetsforlaget.

Olafsen, A. R. & Maugesten, M. (2022). *Matematikkdidaktikk i klasserommet* (3rd ed.). Universitetsforlaget.

Onyishi, C. N. & Sefotho, M. M. (2021). Differentiating instruction for learners' mathematics self-efficacy in inclusive classrooms: Can learners with dyscalculia also benefit? *South African Journal of Education*, 41 (4), 1–15. https://doi.org/ 10.15700/saje.v41n4a1938

Özcan, Z. Ç., and Eren Gümüş, A. (2019). A modeling study to explain mathematical problem-solving performance through metacognition, self-efficacy, motivation, and anxiety. *Australian Journal of Education*, 63 (1), 116–134. https://doi.org/10.1177/000494411984007

Pajares, F. & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology*, 20 (4), 426–443. https://doi.org/10.1006/ceps.1995.1029

Pajares, F. & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: a path analysis. *Journal of Educational Psychology*, 86 (2), 193–203. https://doi.org/10.1037/0022-0663.86.2.193

Pajares, F. & Miller, M. D. (1995). Mathematics self-efficacy and mathematics performances: the need for specificity of assessment. *Journal of counseling psychology*, 42 (2), 190.

Pierce, R. L. & Adams, C. M. (2005). Using tiered lessons in mathematics. *Mathematics Teaching in the Middle School*, 11 (3), 144–149. https://doi.org/10.5951/MTMS.11.3.0144

Pozas, M., Letzel, V. & Schneider, C. (2020). Teachers and differentiated instruction: exploring differentiation practices to address student diversity. *Journal of Research in Special Educational Needs*, 20 (3), 217–230. https://doi.org/10.1111/1471-3802.12481

Schöber, C., Schütte, K., Köller, O., McElvany, N. & Gebauer, M. M. (2018). Reciprocal effects between self-efficacy and achievement in mathematics and reading. *Learning and Individual Differences*, 63, 1–11. https://doi.org/10.1016/j.lindif.2018.01.008

Scott, B. E. (2012). *The effectiveness of differentiated instruction in the elementary mathematics classroom* [Unpublished doctoral thesis]. Ball State University.

Singh, O. (2017). Danningsperspektiver på utforming av lærersubjektet i læreverket i matematikk. *Norsk Pedagogisk Tidsskrift*, 101 (3), 266–277. https://doi.org/10.18261/issn.1504-2987-2017-03-07

Skaalvik, E. M., Federici, R. A. & Klassen, R. M. (2015). Mathematics achievement and self-efficacy: relations with motivation for mathematics. *International Journal of Educational Research*, 72, 129–136. https://doi.org/10.1016/j.ijer.2015.06.008

Skaalvik, E. M. & Skaalvik, S. (2015). *Motivasjon for læring: teori og praksis*. Universitetsforlaget.

Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M. &Maulana, R. (2019). Differentiated instruction in secondary education: a systematic review of research evidence. *Frontiers in Psychology*, 10, 2366. https://doi.org/10.3389/fpsyg.2019.02366

Spielberg, Y. & Azaria, A. (2021). Revelation of task difficulty in AI-aided education. In *Proceedings of ICTAI* (pp. 1403–1408). IEEE.

Street, K. E. S., Malmberg, L.-E. & Stylianides, G. J. (2017). Level, strength, and facet-specific self-efficacy in mathematics test performance. *ZDM*, 49 (3), 379–395. https://doi.org/10.1007/s11858-%20%09017-0833-0

Street, K. E., Stylianides, G. J. & Malmberg, L. E. (2022). Differential relationships between mathematics self-efficacy and national test performance according to perceived task difficulty. *Assessment in Education: Principles, Policy & Practice*, 29 (3), 288–309. https://doi.org/10.1080/0969594X.2022.2095980

Suarez, D. (2007). Differentiation by challenge: using a tiered program of instruction in mathematics. In W. Powel & O. K. Powel (Eds), *Making the difference: differentiation in international schools* (pp. 199–227). CreateSpace Independent Publishing Platform.

Tomlinson, C. A. (2014). *The differentiated classroom: responding to the needs of all learners*. ASCD.

Tomlinson, C. A. & Imbeau, M. B. (2010). *Leading and managing a differentiated classroom*. ASCD.

Usher, E. L. & Pajares, F. (2009). Sources of self-efficacy in mathematics: a validation study. *Contemporary Educational Psychology*, 34 (1), 89–101. https://doi.org/ 10.1016/j.cedpsych.2008.09.002

Winther, K. (1965). *Oppgavesamling i regning og matematikk: for 8. skoleår i den niårige folkeskolen*. H. Aschehoug & Co.

Zakariya, Y. F. (2022). Improving students' mathematics self-efficacy: a systematic review of intervention studies. *Frontiers in Psychology*, 13. https://doi.org/10.3389/fpsyg.2022.986622

Zakariya, Y. F. (2021). Self-efficacy between previous and current mathematics performance of undergraduate students: an instrumental variable approach to exposing a causal relationship. *Frontiers in Psychology*, 11, 1–11. https://doi.org/10.3389/fpsyg.2020.556607

Zuffianò, A., Alessandri, G., Gerbino, M., Kanacri, B. P. L., Di Giunta, L. et al. (2013). Academic achievement: the unique contribution of self-efficacy beliefs in self-regulated learning beyond intelligence, personality traits, and self-esteem. *Learning and Individual Differences*, 23, 158–162. https://doi.org/10.1016/j.lindif.2012.07.010

Maria Herset

Maria Heret is a PhD candidate at the Faculty of Education and Arts, Nord University, Norway. Her research interests include self-efficacy, mathematics teaching, and differentiated instruction in mathematics education.

maria.herset@nord.no

Mohamed El Ghami

Mohamed El Ghami is a professor and leader of research group *Knowledge building and knowledge in education* at the Faculty of Education and Arts, Nord University, Norway. His research interest is mainly in computational science, knowledge, integration of ICT in education, curricula, and education, with a focus on teaching mathematics. His research in computational science is close to the top at the international level, with solid publication output.

mohamed.el-ghami@nord.no

Annette Hessen Bjerke

Annette Hessen Bjerke is an associate professor in mathematics education at Oslo Metropolitan University, Norway, where she currently leads the doctoral program in *Educational sciences for teacher education*. She has participated in several research projects focusing on mathematics teacher education and the teaching of mathematics, focusing in particular on the theory-practice divide, and pre-service teachers' developing self-efficacy and subject matter knowledge.

anetsen@oslomet.no