

Journal of Educational Psychology

The Potential of Relevance Interventions for Scaling Up: A Cluster-Randomized Trial Testing the Effectiveness of a Relevance Intervention in Math Classrooms

Hanna Gaspard, Cora Parrisius, Heide Piesch, Markus Kleinhansl, Eike Wille, Benjamin Nagengast, Ulrich Trautwein, and Chris S. Hulleman

Online First Publication, October 4, 2021. <http://dx.doi.org/10.1037/edu0000663>

CITATION

Gaspard, H., Parrisius, C., Piesch, H., Kleinhansl, M., Wille, E., Nagengast, B., Trautwein, U., & Hulleman, C. S. (2021, October 4). The Potential of Relevance Interventions for Scaling Up: A Cluster-Randomized Trial Testing the Effectiveness of a Relevance Intervention in Math Classrooms. *Journal of Educational Psychology*. Advance online publication. <http://dx.doi.org/10.1037/edu0000663>

The Potential of Relevance Interventions for Scaling Up: A Cluster-Randomized Trial Testing the Effectiveness of a Relevance Intervention in Math Classrooms

Hanna Gaspard¹, Cora Parrisius¹, Heide Piesch¹, Markus Kleinhansl¹, Eike Wille¹, Benjamin Nagengast¹, Ulrich Trautwein¹, and Chris S. Hulleman²


¹ Hector Research Institute of Education Sciences and Psychology, University of Tübingen


² Center for the Advanced Study of Teaching and Learning, University of Virginia


Relevance interventions have shown a great potential to foster motivation and achievement (Lazowski & Hulleman, 2016). Yet, further research is warranted to test how such interventions can be successfully implemented in practice. We conducted a cluster-randomized trial in ninth-grade math classrooms to test the effectiveness of a relevance intervention, which was shown to be efficacious when implemented by researchers, for fostering motivation and achievement under real-world conditions. The 78 participating classrooms ($N = 1,744$ students) were randomly assigned to one of two intervention conditions or a waitlist control condition. The intervention was implemented by master's students or the regular math teachers. Intervention effects were evaluated using self-reports, teacher ratings, and achievement tests 4 weeks and 3 months after the intervention, controlling for the initial levels of the outcomes. Compared with the control condition, both intervention conditions showed similar positive effects on utility value. Unexpectedly, students in both intervention conditions also reported higher perceived cost compared with students in the control condition after the intervention. When implemented by master's students, additional intervention effects on students' growth mindsets and a standardized achievement test could be observed. Only small differences in effectiveness were observed between the intervention conditions, although master's students showed a higher level of adherence. In both intervention conditions, higher levels of adherence and lower levels of discipline problems were associated with more positive changes in utility value. Overall, the intervention thus showed mixed effects. Future research should therefore continue to examine the conditions under which relevance interventions work in practice.


Educational Impact and Implications Statement


This study tested the effectiveness of a 90-min relevance intervention for fostering students' motivation and achievement in ninth-grade math classrooms when it was delivered by master's students or the students' regular math teachers. The intervention consisted of (a) following a presentation containing research results on the importance of effort and one's attitude toward achievement in mathematics and examples of the usefulness of mathematics in different life domains and (b) evaluating quotes from young adults who talked about the relevance of mathematics. Although students reported a decline in their self-reported utility value on average, the intervention groups reported higher utility value 3 months after the intervention compared with the control group. When the intervention was delivered by master's students, the students participating in the intervention also believed that their performance in mathematics was driven more by effort and less by talent, and they performed better on a standardized math test compared with the control group. At the same time, however, students in both intervention groups reported higher perceived cost compared with the control group, and other measures were unaffected.

Hanna Gaspard  <https://orcid.org/0000-0001-8830-8031>

Cora Parrisius  <https://orcid.org/0000-0003-1277-6028>

Markus Kleinhansl  <https://orcid.org/0000-0002-1460-1224>

Benjamin Nagengast  <https://orcid.org/0000-0001-9868-8322>

Ulrich Trautwein  <https://orcid.org/0000-0003-0647-0057>

Eike Wille is now at the Competence Center for Teaching, Nuertingen-Geislingen University.

This research was supported by the Eliteprogramme for Postdocs of the Baden-Württemberg Stiftung and the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63). The work of

Hanna Gaspard was additionally supported by the Postdoc Academy of the Hector Research Institute of Education Sciences and Psychology, Tübingen, funded by the Baden-Württemberg Ministry of Science, Research, and the Arts. Cora Parrisius, Heide Piesch, and Markus Kleinhansl were/are doctoral students at the LEAD Graduate School and Research Network (GSC1028), which was funded within the framework of the Excellence Initiative of the German federal and state governments.

Correspondence concerning this article should be addressed to Hanna Gaspard, who is now at the Center for Research on Education and School Development, TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany. Email: hanna.gaspard@tu-dortmund.de

Keywords: cluster-randomized trial, expectancy-value theory, implementation fidelity, mathematics, relevance intervention

Supplemental materials: <https://doi.org/10.1037/edu0000663.supp>

“Why do I have to learn all this stuff?” or “When will I ever need this information outside of school?” Math teachers often find themselves confronted with similar questions raised by their students. Many adolescents do not see the relevance of mathematics for their lives (Harackiewicz et al., 2010), and on average, their perceived value of mathematics decreases throughout secondary school (e.g., Gaspard et al., 2017; Jacobs et al., 2002; Watt, 2004). When students perceive a high value in a particular domain, however, they are more engaged and show better learning outcomes (e.g., Hulleman et al., 2008; Song et al., 2020). Given the great need for skilled personnel in math-related careers (Noonan, 2017), it is therefore a key challenge for math teachers to help their students see the value in what they are learning in class so that they keep motivated and perform up to their potential (Brophy, 1999; Hidi & Harackiewicz, 2000). Over the last years, field experiments have shown that relevance interventions (also known as utility-value interventions; Hulleman & Harackiewicz, 2021) can positively affect students’ motivation and achievement (Lazowski & Hulleman, 2016; Rosenzweig & Wigfield, 2016) and thus seem to be a promising tool for educational practice.

In previous research, evidence for the efficacy of short relevance interventions has been provided, stemming from field studies with students in secondary school (e.g., Gaspard, Dicke, Flunger, Brisson, et al., 2015; Hulleman & Harackiewicz, 2009; Rosenzweig et al., 2019) and in college (e.g., Canning et al., 2018; Hulleman et al., 2010, 2017; Rosenzweig et al., 2020). Although the intervention approaches used to foster perceived relevance varied between studies, the interventions were typically implemented with a high degree of control over the implementation by the researcher. However, before interventions can be brought to scale successfully, a solid test of these interventions under more realistic conditions is warranted, for example when the regular teachers or other persons are trained to implement the intervention more broadly. When the researchers who initially developed the intervention are not involved any longer, this can lead to more variation in how the intervention is implemented and thus to lower degrees of implementation fidelity (i.e., the degree to which the intervention is implemented as designed; Greene, 2015; Hulleman & Corray, 2009; O’Donnell, 2008). This is particularly the case for interventions that are delivered in the regular classroom setting.

This study was aimed at testing the effectiveness of the “Motivation in Mathematics” (MoMa) intervention, which is a 90-min relevance intervention developed for ninth-grade math classrooms in academic track schools in Germany. Although there is certainly a need to foster students’ motivation and achievement across different school tracks, research has found that students in this track tend to show a particularly pronounced decline in their interest in mathematics after the transition to secondary school (Frenzel et al., 2010), and lower self-concept and interest compared with students in the lower tracks, at least when controlling for their individual achievement (Trautwein et al., 2006). A previous efficacy study found positive effects of the MoMa intervention on students’ values, expectancies,

teacher-rated effort, and standardized achievement scores in mathematics (Brisson et al., 2017; Gaspard, Dicke, Flunger, Brisson, et al., 2015). In this first efficacy study, the intervention was implemented in the classroom by the researchers. With the present effectiveness study, we aimed to test whether the positive effects of the intervention could be replicated when the intervention was delivered by master’s students in education science (as an easily-accessible group of people in this particular context) or by the students’ regular math teachers, both trained for this purpose, and thus tested the intervention under conditions that would potentially allow it to be scaled up. Thus, this study is the first to test the effectiveness of a relevance intervention that was previously shown to be efficacious and thereby contributes to general calls for effectiveness studies to test the replicability of positive intervention effects under conditions that reflect the reality of the school context (e.g., Kim, 2019). Such studies allow researchers to identify the conditions under which interventions will work in practice. To be able to contribute to answering this question, we also investigated several aspects of implementation fidelity (i.e., adherence, quality of delivery, student responsiveness) rated from different perspectives (i.e., observers and students) and whether variation in implementation fidelity was associated with the effectiveness of the intervention.

Expectancy-Value Theory

Relevance interventions are grounded in Eccles et al.’s (1983) expectancy-value theory (for a review, see Wigfield et al., 2016), one of the most prominent theories in motivation research. Expectancy-value theorists posit that students’ academic behavior on a particular task and their choices related to this task are most directly predicted by their expectancies for how well they will do on this task and the value they attach to it.

Expectancies of success are conceptually close to other constructs referring to competence-related beliefs, such as self-concept and self-efficacy (Wigfield & Eccles, 2000). Students’ self-concept in a specific domain (e.g., math) refers to their subjective evaluation of their ability in this domain (Marsh, 2007), whereas self-efficacy is defined as students’ beliefs about their ability to perform given academic tasks at designated levels (Bandura, 1997). These constructs have many conceptual similarities; yet, there are also important differences between them, including time orientation (past- vs. future-oriented), temporal stability, and the degree to which these beliefs are shaped by frames of reference or standards against which people judge their competencies (Bong & Skaalvik, 2003; Marsh et al., 2019). Whereas self-concepts are highly stable and heavily affected by social comparisons with students’ classmates as one important frame of reference, self-efficacy is supposed to be more malleable and has been shown to be affected by frame-of-reference effects to a smaller degree (Bong & Skaalvik, 2003; Marsh et al., 2019; Möller et al., 2009). For these reasons, students’ self-concept might be more difficult to affect through interventions than self-efficacy.

However, these different competence-related beliefs are typically highly correlated and not always separated in empirical research (Eccles & Wigfield, 2002). In line with our theoretical background, we therefore use the term *expectancies* to refer to all competence-related beliefs, although we investigated students' self-concept and self-efficacy as separate outcomes.

Eccles and colleagues (1983) distinguish four major components that contribute to subjective task value: Whereas intrinsic value (enjoyment of a given task), attainment value (perceived personal importance to do well on a task), and utility value (perceived usefulness of a given task for achieving one's goals) contribute positively to subjective task value, cost (perceived negative consequences of engaging in a task) contributes negatively to it (for further discussion of these components, see Barron & Hulleman, 2015; Eccles et al., 2005; Wigfield et al., 2017). Recent research has further shown that some of these components can be broken down into subfactors: Attainment value can be separated into importance of achievement and personal importance, utility value can be differentiated into the usefulness for different life domains in the short and in the long term, and cost can be assessed in terms of effort required, emotional cost, and opportunity cost (Gaspard, Dicke, Flunger, Schreier, et al., 2015; Perez et al., 2014).

A large body of research including longitudinal studies conducted in different countries has supported the basic assumptions of expectancy-value theory, showing that students' expectancies and values are important predictors of their engagement and achievement in different domains as well as their academic choices in related domains (e.g., Marsh et al., 2005; Simpkins et al., 2006; Watt et al., 2012). It has further been shown that students' expectancies and values interact in predicting their academic outcomes such that the highest outcomes are predicted when both expectancies and values are high (Lauermann et al., 2015; Nagengast et al., 2011; Trautwein et al., 2012).

Given the richness of value components described in expectancy-value theory, this opens up different potential avenues for enhancing students' values. In prior intervention studies, researchers have focused on utility value because it is assumed to be more malleable through external interventions compared with attainment and intrinsic value (Gaspard, Dicke, Flunger, Brissou, et al., 2015; Harackiewicz et al., 2014; Hulleman et al., 2010). Even though the main target of these interventions is utility value, we use the term *relevance intervention* here to denote that these interventions rely on mechanisms that include not only utility but also target relevance as "a personally meaningful connection to the individual" (Priniski et al., 2018; p. 12) and can have impacts beyond utility value, including identification with and interest in the domain (Gaspard, Dicke, Flunger, Brissou, et al., 2015; Hulleman & Harackiewicz, 2021).

Relevance Interventions: Prior Research and Remaining Questions Before Bringing Them to Scale

Different intervention approaches to enhance students' perceived relevance have been applied so far. An important distinction is between directly communicating utility value (e.g., by providing arguments for the usefulness of the learning material) and relevance-inducing reflection tasks (e.g., by writing an essay on the relevance of the learning material). With respect to communicating

utility value, several lab studies have shown that this approach can promote interest and performance for students with high initial motivation, but it can backfire and undermine interest and performance for students with low expectancies (Canning & Harackiewicz, 2015; Durik et al., 2015; Durik & Harackiewicz, 2007; for theoretical explanations of these moderation effects, see Binning & Brownman, 2020; Durik et al., 2015). However, such negative effects for students with low expectancies can apparently be ameliorated when the communication of utility value is combined with the self-generation of utility value or an expectancy boost (Canning & Harackiewicz, 2015; Durik et al., 2015).

With respect to relevance-inducing reflection tasks, this approach asks students to reflect upon the relevance of the learning material to their lives and to make specific, personal connections (Brisson et al., 2020; Harackiewicz et al., 2016; Hulleman et al., 2017). Several studies with high school (Hulleman & Harackiewicz, 2009) and college students (Canning et al., 2018; Harackiewicz et al., 2016; Hulleman et al., 2010, 2017; Kosovich et al., 2019; Rosenzweig et al., 2020) have shown that this intervention approach yields positive effects on utility value, interest, and performance, although the effects are often limited to students at risk for low motivation (i.e., students with low expectancies; e.g., Hulleman et al., 2010, 2017; Hulleman & Harackiewicz, 2009; or students from minority groups; e.g., Harackiewicz et al., 2016). However, a few studies have also provided evidence that younger students or students at two-year colleges, who may be less prepared and less compliant to write essays about the relevance of the learning material, do not similarly benefit from such interventions (Brisson et al., 2017; Canning et al., 2019; Gaspard et al., 2015).

Both intervention approaches (i.e., directly communicating utility value and relevance-inducing reflection tasks) have thus shown positive effects on motivation and performance for some groups of students: Whereas directly communicating utility value was shown to promote motivation and performance only for students with high interest or expectancies, relevance-inducing tasks were often found to be particularly beneficial for students with low expectancies. The moderating role of students' expectancies on the effects of these interventions, in particular, is in line with the interaction of expectancies and values postulated in expectancy-value theory and found in prior research that used correlational data (e.g., Nagengast et al., 2011; Trautwein et al., 2012). For relevance interventions to be effective, it thus seems important that they also foster students' expectancies (Canning & Harackiewicz, 2019; Durik et al., 2015; Hulleman et al., 2017; Rosenzweig et al., 2020).

What is the most effective way to enhance perceived relevance in practice so that as many students as possible can benefit from them? One potential approach might be to combine the communication of utility value with relevance-inducing reflection tasks (Canning & Harackiewicz, 2015). To prevent the communication of utility value from potentially backfiring for students with low expectancies, it might also be important to provide them with an expectancy boost (Durik et al., 2015). One possible intervention approach that could be applied to target students' expectancies could be a growth mindset intervention (see Hulleman & Barron, 2016), which is aimed at enhancing students' persistence and their willingness to take challenges by teaching them that their abilities are malleable (Blackwell et al., 2007; Yeager et al., 2016, 2019).

However, as far as we know, prior research has not investigated students' mindsets in the context of relevance interventions.

Before they can be brought to scale successfully, questions also remain regarding the best way to implement such interventions in educational practice. In fact, prior intervention studies can additionally be differentiated in terms of the level at which the intervention was implemented. Whereas some interventions were implemented at the student level, typically as a task to be completed outside of class (e.g., Canning et al., 2018; Hulleman et al., 2010; Rosenzweig et al., 2020), others were implemented at the classroom level within the regular class (Gaspard, Dicke, Flunger, Brisson, et al., 2015; Shin et al., 2019; Woolley et al., 2013). Classroom-based interventions, in particular, have a high relevance for educational practice: They correspond to the natural learning setting and involve all students (however, see Binning & Browman, 2020). Compared with interventions conducted outside of the classroom, students' engagement in the intervention might therefore depend less on individual characteristics. They also open up the opportunity to engage students in discussions as a group, thereby potentially leading to changes in motivation at the classroom level that could reinforce positive effects for individual students.

However, in classroom-based interventions, in particular, there can be variation in how interventions are implemented by the person delivering the intervention in the classroom. Prior classroom-based relevance interventions were implemented either by the researchers themselves (Gaspard, Dicke, Flunger, Brisson, et al., 2015) or by a small, selected group of teachers (Woolley et al., 2013) or preservice teachers (Shin et al., 2019) who received intensive training and support for the delivery in the classroom, thereby ensuring high levels of implementation fidelity. It therefore remains unclear whether relevance interventions can be effectively implemented under conditions that enable scaling up, for instance, when they are delivered by the regular teachers, who have limited resources for training. Under such conditions, there might be more variation in how the intervention is implemented, which could potentially lead to reduced effects of the intervention (Durlak & DuPre, 2008; Hulleman & Cordray, 2009; O'Donnell, 2008; Weiss et al., 2014). When investigating the effects of classroom-based interventions, it is thus imperative to consider implementation fidelity, because it allows researchers to better understand and interpret the impact of the intervention and to study the sources of variation in intervention effects. Important aspects of implementation fidelity are adherence (i.e., the extent to which the intervention is implemented as intended), exposure (e.g., the number of sessions or the length of each session), quality of delivery (i.e., qualitative aspects of intervention delivery not directly related to the implementation of prescribed content), participant responsiveness (participants' responses to the intervention), and program differentiation (i.e., ensuring that there is no diffusion of treatments; Dane & Schneider, 1998).

MoMa as a Classroom-Based Relevance Intervention in Secondary School

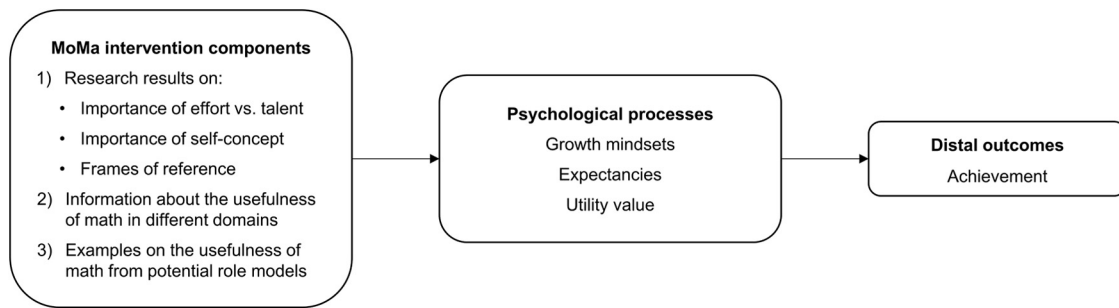
The MoMa intervention that was evaluated in this study is a classroom-based relevance intervention that targets students' utility value in mathematics and was developed for ninth-grade academic track students in Germany. Given the low levels of utility value that have been found for students in this age group, also in the academic track

(Gaspard et al., 2017; Harackiewicz et al., 2010), a relevance intervention in this group of students can be seen as a "universal prevention" against further decreases in perceived usefulness. The 90-min intervention consists of a psychoeducational presentation for the whole class and relevance-inducing reflection tasks, on which students work individually. The psychoeducational presentation, in turn, has two main components. First, research results on the importance of effort and self-concept for math achievement are presented and students are told about frame-of-reference effects that can occur within the classroom. This first part of the intervention is aimed at inoculating students against potential negative effects of highlighting the relevance of mathematics for students with low expectancies (Durik, Hulleman, et al., 2015; Durik, Shechter, et al., 2015). Given that the importance of effort compared with talent for acquiring math skills is emphasized, this part also targets students' growth mindsets (Blackwell et al., 2007). Second, students are provided with various examples on the utility of mathematics for future education, career opportunities in different fields, and leisure time activities. This part thus communicates the usefulness of mathematics in different domains directly and is meant to prepare students for the relevance-inducing tasks. Finally, students work on relevance-inducing tasks, during which they are asked to reflect about the personal relevance of mathematics for their lives and thus personalize and internalize the previously heard content. Figure 1 represents the intervention logic Model for the MoMa intervention, detailing the intervention components and the psychological processes triggered by the intervention (i.e., expectancies, mindsets, utility value), which should ultimately enhance students' achievement as a more distal outcome.

In a first cluster-randomized trial with 82 ninth-grade mathematics classrooms, the efficacy of this intervention when delivered by researchers in the classroom was evaluated (Brisson et al., 2017; Gaspard, Dicke, Flunger, Brisson, et al., 2015). More specifically, a group of five female doctoral candidates who were involved in the development of the intervention delivered it in the classroom. There were two intervention conditions, which differed with respect to the relevance-inducing tasks that students worked on: Students either got to read and evaluate interview quotations from young adults related to the usefulness of mathematics (quotations condition) or they were asked to write an essay about the usefulness of mathematics in their lives (text condition). Compared with a waitlist control condition, both intervention conditions showed positive effects on students' utility value six weeks and five months after the intervention, with stronger effects in the quotations condition ($d = .30$ at posttest, $d = .26$ at follow-up) compared with the text condition ($d = .14$ at posttest, $d = .16$ at follow-up). Whereas the text condition only showed additional effects on homework self-efficacy, the quotations condition was also shown to foster attainment value, intrinsic value (at the follow-up), self-concept (at the posttest), homework self-efficacy, teacher-rated effort, and test scores. The quotations condition was thus more successful than the text condition.

To conclude, this first efficacy study showed that a 90-min relevance intervention can have positive effects on students' values, expectancies, effort, and achievement in mathematics when it is implemented with a high level of control by the researchers involved in its development. However, it remains an important question for educational practice whether such positive effects can also be found when the intervention is delivered by someone else. Different options to scale up this intervention might be to train college students who could then be deployed to the classrooms as part of their course work or to train the regular math teachers.

Figure 1
The Motivation in Mathematics (MoMa) Intervention Logic Model



The Present Study

With the present study, we aimed at investigating the effectiveness of the MoMa intervention under conditions that would enable implementing the intervention at scale. We tested two conditions representing further steps to scale: The more successful intervention in the efficacy study (i.e., quotations) was implemented either by master's students or by the regular math teachers, both trained for this purpose. This study extends prior research on relevance interventions because it provides a solid test of the effects of this type of intervention in practice: An adequately powered cluster-randomized trial was conducted to test the effectiveness of the intervention on a large set of outcomes, the design and hypotheses were preregistered (see <https://osf.io/d4vp9>), and the implementation of the intervention by teachers was tested and compared with a condition that was closer to the conditions used in the efficacy trial.

In our view, there are different potential strengths and weaknesses from having master's students versus teachers implement the intervention. In the German university context in which the study was conducted, master's students represent a group that can be easily accessed through courses. In this study, the master's students were enrolled in a program at the Institute where the study was conducted (Education Sciences and Psychology) and delivered the intervention as part of a project-based class focusing on intervention research in theory and practice. Similar to how adolescents perceived the doctoral candidates who delivered the intervention in the efficacy trial, adolescents might see such master's students as potential role models (Bandura, 1977) and as experts coming from the university to provide them with insights from research. As individuals from outside of school who were enrolled in a program not directly related to mathematics, master's students might also represent an authentic source of information with respect to the usefulness of math skills. Because of their training, they should be more familiar with the requirements of randomized trials and the importance of high implementation fidelity compared with teachers. However, they usually do not have teaching experience. Math teachers, on the other hand, bring their professional competencies (Kunter et al., 2013) and should be able to draw on these when delivering a classroom-based intervention. They are also familiar with their students and might be able to use their relationship with their students and their knowledge about their students' needs and interests to help them make connections between mathematics and their lives. Finally, teachers have limited resources available for training. Because of their different time resources and their participation for course credit

versus voluntary participation, the master's students received more extensive training in this study than the math teachers.

With this study, we aimed to investigate two sets of research questions: The first set focused on the effectiveness of the intervention, whereas the second set focused on measures of implementation fidelity in the classroom. Because there was not a lot of variation in exposure to the intervention, and because the MoMa intervention entailed a clear-cut program with a low risk of diffusion (i.e., the waitlist control group did not have access to the intervention materials), we focused on adherence, quality of delivery, and student responsiveness. More specifically, we had four major research questions.

First, can we replicate the positive effects of the MoMa intervention on students' values, expectancies, effort, and achievement in mathematics when the intervention is implemented by trained master's students or math teachers instead of researchers? As preregistered, we expected that the relevance intervention in both conditions would have positive effects on students' utility value, attainment value, intrinsic value, expectancies, effort, and achievement in mathematics based on the results of the efficacy study (Brisson et al., 2017; Gaspard, Dicke, Flunger, Brisson, et al., 2015). Because of the focus of the intervention, utility value was considered to be the primary outcome and the other measures were considered to be secondary outcomes. We also wanted to explore whether the intervention affected students' growth mindsets, because the psychoeducational presentation included a part that focused on the importance of effort as compared with the importance of talent. Because this was not tested in the efficacy study, we did not preregister any hypotheses regarding students' growth mindsets.

Second, is the MoMa intervention differentially effective when implemented by master's students or the regular math teachers? Given that both master's students and math teachers bring strengths and weaknesses when it comes to delivering such an intervention as detailed above, we had no a priori expectations as to whether the intervention would be more effective when implemented by master's students or by the regular math teachers.

Third, are there differences between master's students and math teachers in how the intervention is implemented in the classroom? Based on the differences in training and background between master's students and teachers detailed above, we preregistered the following hypotheses: We expected that master's students would show a higher degree of adherence compared with teachers and that teachers would show better classroom management as an indicator of

quality of delivery compared with master's students. Additionally, we explored differences between the two conditions in other indicators of quality of delivery and student responsiveness.

Fourth, are differences in the implementation of the intervention associated with the effectiveness of the intervention? For this research question, we investigated changes in utility value as the primary outcome of the intervention. As preregistered, we expected that both a higher degree of adherence and a higher degree of classroom management would be positively associated with higher effectiveness of the intervention. On the basis of previous research on teaching quality and teacher and student motivation, we also explored whether quality of delivery in terms of clarity of instruction, a supportive climate (Lipowsky et al., 2009; Pianta & Hamre, 2009), enthusiastic teaching (Keller et al., 2016), authenticity (Kreber et al., 2010), and perceived autonomy support (Ryan & Deci, 2020) was associated with changes in the intervention conditions. Finally, we explored associations between student responsiveness (i.e., participation and interest in the intervention) and the effectiveness of the intervention.

Method

Sample and Procedure

Data for this second large-scale test of the MoMa intervention (MoMa 2) were collected in ninth-grade classrooms in academic track schools in the German state of Baden-Württemberg from October 2017 to March 2018. We report all preregistered outcomes and analyses here to provide a comprehensive picture of the main results of this cluster-randomized trial.¹ The Ministry of Education and Cultural Affairs in Baden-Württemberg approved the study and the collection of the data (date of approval: July 26, 2017; file number: 31-6499.20/1105). The Ethics Committee for Psychological Research at the University of Tübingen confirmed that the procedures were in line with ethical standards of research with human subjects (date of approval: August 1, 2017; file number: 2017/0724/75). The design of the study was preregistered on November 9, 2017; after the first wave of data collection, but before the implementation of the intervention and before any analyses had been run (<https://osf.io/d4vp9>).

In Baden-Württemberg, as in most German federal states, students are tracked from grade 5 on. The academic track is the highest track and leads to a general university entrance qualification. Parents choose their child's track on the basis of recommendations made by elementary school teachers. More than 40% of all students attending elementary school go on to attend academic track schools in Baden-Württemberg (Autorengruppe Bildungsberichterstattung, 2018). In the school year 2017/18, ninth-grade students in Baden-Württemberg attending the academic track were found to have an average Highest International Socio-Economic Index of Occupational Status (HISEI; see Ganzeboom & Treiman, 1996) of 61.0 ($SD = 6.1$) compared with 40.1 ($SD = 7.3$) for students attending other school types.

Within each school, the participating classes were randomly assigned to three conditions: (a) relevance intervention implemented by master's students, (b) relevance intervention implemented by the regular math teachers, and (c) waitlist control condition. To determine the necessary sample size for the study, we

conducted a power analysis for a multisite cluster-randomized trial with the treatment implemented at Level 2 (i.e., classes within schools are randomly assigned to experimental conditions) with Optimal Design (Raudenbush et al., 2011). To derive realistic estimates for the necessary parameters, we drew on data from the efficacy study (see Gaspard, Dicke, Flunger, Brisson, et al., 2015). The results of the power analysis indicated that we would achieve a power of .79 to detect intervention effects of $\delta = .20$ (comparing a single intervention condition with the control condition) with a total sample of 25 schools (with one class per experimental condition and $n = 25$ students per class; for more details about the power analysis and the assumptions, see the online supplemental materials).

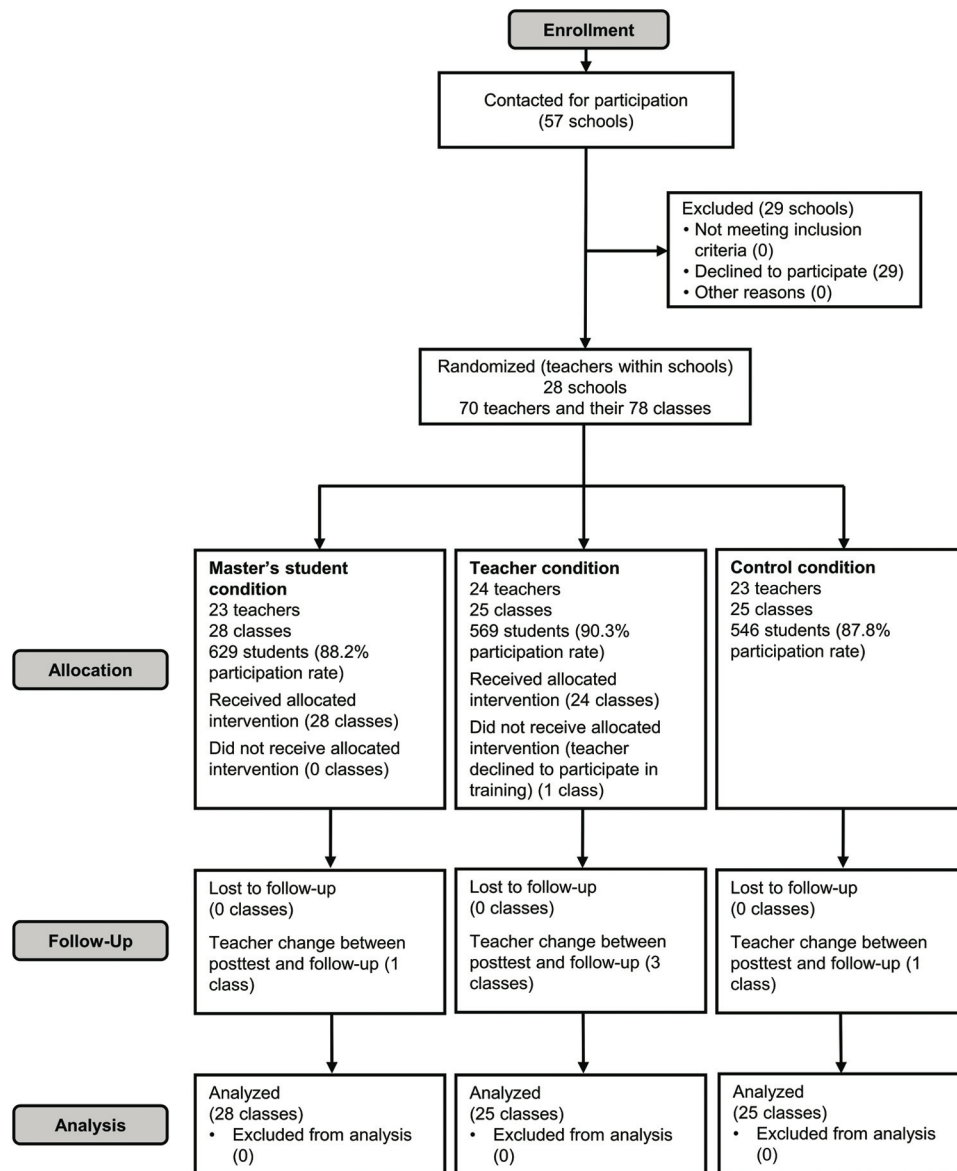
Figure 2 provides an overview of all phases of the study procedure starting with the enrollment. To recruit the participating classes, a total of 57 academic track schools in Baden-Württemberg was contacted. These regular academic track schools were contacted either because they were situated close to the university conducting the study or because of previous cooperation with these schools. The headmasters of the schools were first contacted and were then asked to forward the information materials about the study to the math teachers. There was no incentive for participation, neither for the teachers nor for the classes. A total of 70 teachers (44.2% female; age $M = 38.7$, $SD = 9.8$; years of teaching experience $M = 10.4$, $SD = 8.5$) from 28 schools agreed to participate in the study with their 78 classes (1–5 classes per school). Because we had met our goal with this sample size, we stopped the recruitment at this point and subsequently randomized these teachers and their classes to the three different conditions (within each school). Eight teachers participated with two classes each; all other teachers participated with one class each. To reduce the risk of diffusion effects, classes of the same teacher were put in the same condition and the randomization was thus based on the math teachers. This randomization process resulted in 28 classes in intervention Condition 1 (master's student), 25 classes in intervention Condition 2 (teacher), and 25 classes in the waitlist control condition.

Of the 28 participating academic track schools, 25 schools (with 68 participating classes) followed an eight-year-long curriculum from grades 5 to 12, preparing students for their higher education qualification. Three schools (with 10 participating classes), in contrast, provided a nine-year-long curriculum from grades 5 to 13. Because the classes were allocated to the three conditions within schools, these classes were equally distributed to the conditions.

Students' participation was voluntary and nonincentivized, and parents and students had to provide written consent. A total of 1,744 students participated in the study, which corresponds to an 88.7% overall participation rate ($n = 629$ in the master's student condition, $n = 569$ in the teacher condition, $n = 546$ in the waitlist control condition). Students' mean age was 14.63 years ($SD = 0.48$) at the beginning of the study, and 53.8% of the participating students were female. Furthermore, 31.7% of the participating

¹ Data from this study were additionally used to investigate the effects of the MoMa relevance intervention on precursors of career choices (i.e., vocational interests, career orientation, STEM career aspirations, perceived importance of math and physics for students' career aspirations) as more distal outcomes (Piesch et al., 2020). Furthermore, data from this study were used in three publications that did not focus on the effects of the intervention but used this data set for different research questions and used the intervention conditions as a covariate (Gaspard & Lauermann, 2020; Parrisius, Gaspard, Trautwein, et al., 2020; Parrisius, Gaspard, Zitzmann, et al., 2020).

Figure 2
Flow Diagram of the Study Design



students had a migration background (i.e., the student or one of their parents was not born in Germany) and 72.9% had at least one parent who obtained a general university entrance qualification.² In terms of migration background and parents' level of education, our sample was approximately representative for academic track students in Baden-Württemberg (Stanat et al., 2019; Statistisches Bundesamt [Destatis], 2018).

The study consisted of three waves of data collection. Students were administered questionnaires by trained research assistants before the intervention in October 2017 (pretest = T1), on average four weeks (14–40 days) after the intervention in December 2017 (posttest = T2), and on average three months (11–17 weeks) after the intervention in February 2018 (follow-up = T3). All 78

classrooms participated in all waves of data collection. On the day of the intervention, 22 students (4%) in the teacher condition and 34 students (5.4%) in the master's student condition were absent. In line with an intention-to-treat approach and the preregistered analyses, these students were included in the analyses as part of the condition their classroom was assigned to. Teachers were asked to rate their students' effort at the same time points. In five classes, the teacher changed between posttest and follow-up. In

² To assess parents' level of education, students were asked to report the highest school leaving certificate that their parents had obtained. Parents' school leaving certificates were then coded based on whether they allow entering a university.

one of these classes, the previous teacher still rated the students; in the other classes, the new teachers did so.

Relevance Intervention

As described above, the intervention consisted of a 90-min lesson on the relevance of mathematics which included a psychoeducational presentation for the whole class (about 45 minutes) and relevance-inducing tasks, on which the students worked individually (about 40 minutes). In these tasks, students were given a total of six interview quotations of young adults describing situations in which mathematics was useful to them and were asked to evaluate these quotations based on their personal relevance. The intervention materials (i.e., PowerPoint slides and individual tasks) were identical for both intervention conditions. For an overview of the intervention components, including their approximate duration, and examples for the intervention materials, see the online supplemental materials.

The intervention materials were based on the efficacy study (Brisson et al., 2017; Gaspard et al., 2015). However, they were subsequently optimized based on the experiences in this study, on further pilot tests, and on feedback from math teachers to make it easier for the teachers to deliver the intervention. The core intervention components, however, were kept intact. Compared with the efficacy trial, the following changes were made. First, as an icebreaker, students were asked to signal their agreement with five statements (e.g., "I like math") using green, yellow, and red traffic light cards instead of an open discussion about students' attitudes toward mathematics. Students were asked about what most students thought about mathematics instead of their own opinion, to avoid pressuring students who may be afraid of openly sharing their opinion. Second, take-home messages on separate slides were introduced in the first part of the presentation to reinforce the content (e.g., "Learning math is like a workout: The more you practice, the better you get"). Third, new examples for the utility of mathematics for daily life that students might relate to more easily compared with the previously used ones (e.g., social media instead of house construction) were included in the presentation. Fourth, the content of the quotations was partly changed based on the ratings of the quotes in the efficacy study and further pilot work. Two quotes were kept the same, two quotes were partly reworded, and two quotes were newly introduced.

Classes in the waitlist control condition received the intervention after the last wave of the data collection. The math teachers in this condition were asked to choose whether they preferred to deliver the intervention themselves or whether they preferred a master's student to deliver the intervention in their classroom.

Training for Master's Students and Teachers

The master's students who delivered the intervention (in the master's student condition) were trained for this purpose as part of a two-semester project-based class on intervention research in theory and practice. The class was part of the curriculum in the master's program Education Sciences and Psychology at the university where the study was conducted. The program is supposed to prepare students for a PhD in education or a career in the education sector (e.g., educational administration) but does not qualify for teaching. Students were able to choose between this course and an alternative course that was aimed at developing a training for teacher candidates. The two-semester course provided information about the theoretical background of motivation interventions as

well as the design of intervention studies, and students received intensive practical training on how to deliver the intervention in the classroom. A group of nine students initially started the class. After the first semester, one student changed master's programs and thus left the class. Two students were not German native-speakers; thus, instead of delivering the intervention, they observed it (see implementation fidelity). A total of six master's students (five female and one male, age $M = 24.7$, $SD = 1.5$) thus delivered the intervention in the classes of this condition (four to five classes per student). These master's students had obtained bachelor's degrees in different majors (education, economics, sociology, and business psychology). Although three of them indicated prior experience of some sort in teaching adolescents (e.g., tutoring), none of them had a teaching certificate or was qualifying for it. They received a script with detailed notes for the presentation and got feedback on their presentation in individual training sessions.

Teachers ($n = 24$; 45.8% female; age $M = 40.2$, $SD = 9.8$; years of teaching experience $M = 11.8$, $SD = 8.9$) assigned to the teacher condition were asked to participate in a 3-hr workshop in small groups. One teacher in this condition declined to participate in the workshop and thus did not deliver the intervention in the classroom (see Figure 2). In line with our preregistration, we followed the intention-to-treat approach in our analyses and included this class in the teacher condition so that the random assignment was kept intact (cf. Sagarin et al., 2014).³ A total of four identical workshops were offered at different sites with two to eight teachers participating in each of these. Two teachers were not able to come to the workshop sessions for organizational reasons and therefore received an individualized workshop. In the workshops, teachers were provided with brief information about the theoretical background of the intervention, the importance of conducting randomized experiments, and the results of the previous efficacy trial. Teachers were then walked through the intervention, followed by a discussion about potential challenges when delivering the intervention in the classroom. Teachers were provided with all necessary intervention materials including a lesson plan with an overview of the different intervention components and their aims (see the online supplemental materials), the PowerPoint slides for the psychoeducational presentation, and a script with detailed notes for the presentation. These materials were identical to those for the master's students, except for small details that naturally needed to be different between the conditions (e.g., the introduction). Teachers were asked to deliver the intervention as intended for scientific reasons, but they were also encouraged to use their own examples for illustration. Teachers indicated that they engaged with the content of the intervention and the preparation of the class outside of the workshop for one to three hours ($n = 10$, $M = 1.80$, $SD = 0.79$).

Instruments

Self-Reported Motivation

Students answered the same set of items on their motivation in math at the pretest, posttest, and follow-up. All items were rated

³To check the consistency of findings, we also ran analyses excluding this class in line with a per-protocol analysis (Sagarin et al., 2014). The pattern of results remained the same with only very slight changes in the regression coefficients. The results can be found in the online supplemental materials.

on a 4-point Likert scale from 1 (*completely disagree*) to 4 (*completely agree*). Sample items and Cronbach's alpha for these scales are provided in Table 1, and the full set of items can be found in the online supplemental materials. The mean scores of the respective scales were used for the analyses, with mean scores being computed under the condition that more than half of the items had valid responses.

Students' *values* were measured with a scale developed by Gaspard, Dicke, Flunger, Schreier, et al. (2015) and used in the efficacy trial by Gaspard, Dicke, Flunger, Brisson, et al. (2015), which was slightly adapted and shortened for the present study. In addition to the four value components, this instrument allows to differentiate between subscales describing multiple facets of utility value, attainment value, and cost. Support for the separability of these subfacets as well as a second-order model was found in previous studies (Gaspard, Dicke, Flunger, Schreier, et al., 2015; 2017). Students' utility value was measured with a total of 12 items tapping general utility, utility for job, utility for daily life, and utility for school. Attainment value was assessed with six items tapping importance of achievement and personal importance. Intrinsic value was measured with three items. Additionally, this scale assessed perceived cost with nine items tapping emotional cost, effort required, and opportunity cost. Students' *mind-sets* were assessed at the pretest and posttest only. Students rated the importance of effort and the importance of talent for math achievement on four and three items, respectively. The scale assessing importance of effort was taken from Rakoczy et al. (2005). The scale assessing importance of talent was taken from the German field test of the Program for International Student Assessment (PISA) 2000 (Kunter et al., 2002) and adapted to mathematics. Students' *expectancies* were measured with four items indicating their academic self-concept in math. This scale included items from previous large-scale studies (e.g., Marsh et al., 2005) and was previously used in the efficacy trial by Brisson et al. (2017). Additionally, students' self-efficacy in math was assessed with four items. This scale was based on the German version of the student questionnaire of PISA 2003 (Ramm et al., 2006). Students reported on their *effort* in math on three items. This scale was adapted from Trautwein et al. (2009).

Teacher-Reported Effort

Teachers rated individual students' math effort on two items on a 4-point Likert scale ranging from 1 (*completely disagree*) to 4 (*completely agree*). A sample item and reliabilities are provided in Table 1.

Achievement

Information on students' previous math grades was collected from school records (ranging from 1 = *insufficient* to 6 = *very good*).⁴ At the pretest and at the follow-up, students worked on a 3 min 30 s normed speed test, which measures students' fluency of solving typical math operations with 50 questions. The sum score was used for the analyses. This speed test is a part of the German mathematics test for grade 9 (Schmidt et al., 2013). Validity studies showed that this short speed test is a very good proxy for students' achievement in longer assessments using standardized, curriculum-based math tests (Ennemoser et al., 2011; Schmidt et al., 2013). This test was also used in the efficacy study by Brisson

et al. (2017). The internal consistency of the test was good at both time points (Kuder Richardson-20 = .88), and the retest reliability was high ($r_{TIT3} = .77$).

At the follow-up, students additionally worked on a curricular math test including a total of 21 tasks tapping four of the topics that are typically covered in the first half of grade 9: Pythagorean theorem (four tasks), central dilation (three tasks), intercept theorems (five tasks), and potencies (nine tasks). The tasks tapping the Pythagorean theorem were taken from the German mathematics test for grade 9 (Schmidt et al., 2013). The tasks tapping the other three topics were adapted from a diagnostic tool for math teachers testing the curricularly defined competencies in grades 9–10 (Kronberger & Weizenegger, 2009). There are clear education standards for mathematics in Baden-Württemberg, which detail which competencies students should acquire over the course of two school years and which should be reached by the end of grade 10, but teachers are free to decide the order in which they cover different topics. There was thus some variance in the topics that the classes had worked on until the follow-up. Five classes following a nine-year academic track curriculum had not covered any of the topics and were thus not able to complete the curricular math test at all. In addition, 13 classes were not able to complete the tasks for one of the topics, and two classes were not able to complete the tasks for two of the topics. To deal with this missing data, we estimated a two-parameter logistic item response theory model using full information maximum likelihood estimation and saved the factor scores for further analysis. The expected-a-posteriori (EAP) reliability of the test for all students with any data on the math test was .66. However, the EAP reliability for students with complete data was .92, suggesting that the low reliability was mainly due to the amount of missing data.

Implementation Fidelity: Observation

Two trained observers (in pairs drawn from a total of eight trained observers) attended each intervention and rated implementation fidelity (i.e., adherence, quality of delivery, student responsiveness). The pairs of observers were created depending on the availability of the observers while ensuring that each combination of observers was realized; all eight observed both intervention conditions. They were provided with all the intervention materials and detailed guidelines on how to rate the different items in the observation scheme beforehand. Additionally, they were trained by using a video example from a pilot intervention in the classroom and by observing training sessions of the master's students who would conduct the intervention.

During the observations, the observers sat nonintrusively in the back or at the side of the classroom. Overall, the intervention was implemented as planned in both intervention conditions, with the observers noting only very few severe deviations from the standardized procedure. More specifically, the intervention consisted of 15 phases that were predefined in advance (see lesson plan in the online supplemental materials). All of them were implemented in all classes.

⁴ As preregistered, we also tried to obtain information about students' scores on a standardized math test conducted at the end of grade 8 from the participating schools. However, many schools were not able to provide this information because they no longer had access to the scores in grade 9. Because of the large amount of missing data (57.6%), we did not include this variable as a covariate in our analyses.

Table 1*Sample Items and Reliabilities for Scales at All Measurement Waves*

Scale	Sample item	Number of items	α_{T1}	α_{T2}	α_{T3}	r_{T1T2}	r_{T1T3}
Utility value		12	.88	.89	.89	.64	.60
General utility	Math is very useful to me.	2	.76	.73	.74	.56	.55
Utility for job	A good knowledge of math will help me in my future job.	3	.84	.86	.87	.63	.57
Utility for daily life	Knowing about the subject of math brings me many advantages in my daily life.	3	.92	.91	.93	.55	.55
Utility for school	Being good at math will help me in the remaining years at school.	4	.72	.81	.82	.49	.46
Attainment value		6	.87	.88	.88	.71	.68
Importance of achievement	It is important to me to be good at math.	3	.88	.88	.88	.67	.63
Personal importance	Math is very important to me personally.	3	.84	.84	.84	.68	.67
Intrinsic value	Math is fun to me.	3	.93	.93	.93	.78	.74
Cost		9	.91	.93	.93	.77	.75
Emotional cost	Doing math makes me really nervous.	3	.80	.84	.83	.69	.67
Effort required	Doing math is exhausting to me.	3	.88	.89	.89	.68	.66
Opportunity cost	I have to give up a lot to be good at math.	3	.89	.90	.91	.66	.66
Importance of effort	I believe that working diligently is the most important thing in math.	4	.80	.84	—	.57	—
Importance of talent	To be good at math, you need to have a talent for it.	3	.80	.83	—	.61	—
Self-concept	I am good at math.	4	.91	.91	.89	.82	.79
Self-efficacy	I am convinced that I can achieve good results on math homework and tests.	4	.85	.90	.88	.73	.71
Effort	I do my best on math tasks.	3	.77	.84	.83	.57	.53
Teacher-rated effort	This student works thoroughly on all of his/her math tasks and homework assignments.	2	.75	.80	.80	.72	.66

At the end of the intervention, the two observers rated the instructor's adherence to the script as well as several measures related to quality of delivery and student responsiveness. Adherence was rated with one item (i.e., "How closely did the instructor follow the intervention script?") on a scale from 1 = *the instructor did not follow the script at all* to 10 = *the instructor followed the script almost word-for-word*. Discipline problems as an indicator of classroom management (e.g., Fauth et al., 2020) were rated with three items (e.g., "During the lesson, the instructor had to warn students a lot to keep them quiet.") on a scale from 1 = *completely disagree* to 4 = *completely agree*. Additionally, the raters rated clarity of instruction, supportive climate, enthusiasm, and authenticity as other indicators of quality of delivery and class participation as an indicator of student responsiveness on several items, most of which were adapted from established scales for assessing teaching quality (e.g., Fauth et al., 2020; Kunter et al., 2011, 2013); all on a scale ranging from 1 = *completely disagree* to 4 = *completely agree* (see the online supplemental materials for sample items and descriptive statistics of these scales, which we considered in a more exploratory way). According to the guidelines provided by Cicchetti (1994), the two observers showed excellent interrater reliabilities for adherence and participation ($ICC \geq .75$) and good interrater reliabilities ($ICC \geq .60$) for all scales except for clarity of instruction and supportive climate. As the average deviation index indicated sufficient agreement for all scales (Burke et al., 1999; see the online supplemental materials for the statistics), the ratings were averaged across the two raters. Still, the results for clarity of instruction and supportive climate should be interpreted with caution. With the exception of supportive climate ($\alpha = .63$), all scales with multiple items showed high internal consistencies ($\alpha = .81$ to $.92$).

Implementation Fidelity: Student Reports

At the end of the intervention, students were also asked to fill out a short questionnaire in which they rated the quality of the instructor's delivery as well as their own responsiveness to the intervention (for sample items and descriptive statistics, see the online supplemental materials)⁵. With respect to quality of delivery, students rated clarity of instruction, supportive climate, enthusiasm, and authenticity on items that were parallel to those rated by the observers (but only partly parallel for authenticity). Furthermore, they rated perceived autonomy support (adapted from Flunger et al., 2020) as another indicator of quality of delivery and their interest in the intervention as an indicator of student responsiveness (self-developed). All items were rated on a scale ranging from 1 = *completely disagree* to 4 = *completely agree*. The scales had sufficient internal consistencies ($\alpha = .74$ to $.82$).

Statistical Analyses

To evaluate the effects of the intervention on different outcomes (Research Question 1), we conducted two-level regression analyses with the students at Level 1 and classrooms at Level 2 in Mplus 7.31 (Muthén & Muthén, 1998–2012). The clustering of classrooms within

⁵ Some aspects of implementation fidelity (i.e., quality of delivery, student responsiveness) were also rated by the instructors (i.e., the master's students or the teachers) at the end of the intervention. However, we did not include them here because understanding the students' perspectives in addition to having an outside perspective (i.e., the observers) should be more important for understanding variation in the effectiveness of the intervention.

schools was accounted for using the design-based correction of standard errors implemented in Mplus (with *type = complex* and school as a stratification variable; McNeish et al., 2017).⁶ The analyses were carried out separately for the different outcomes at the posttest and follow-up. We had preregistered self-reported values (i.e., intrinsic value, attainment value, utility value, and cost), expectancies (i.e., self-concept and self-efficacy), and effort; teacher-rated effort; and achievement as outcomes. In addition, we investigated effects of the intervention on students' mindsets (i.e., importance of effort and importance of talent) and on different facets of utility value, attainment value, and cost to gain a better understanding of the processes targeted through the intervention, although we had not preregistered any hypotheses about these outcomes.

To estimate the effects of the intervention, two dummy variables indicating the two intervention conditions as compared with the control condition were used as predictors at the classroom level. To compare the effects of the two intervention conditions (Research Question 2), we additionally used Wald's chi-square tests to compare the two regression coefficients. In line with our power analysis, we included the pretest score for the respective outcome variable as a covariate in these analyses. As preregistered, we additionally included as covariates those variables for which we found substantial differences between the experimental conditions before the intervention ($\delta > .05$) to yield more precise estimates of the intervention effects (Raudenbush, 1997; What Works Clearinghouse, 2020). We therefore investigated mean differences for the pretest scores of the major study variables. Because we found small pretest differences between the experimental conditions on intrinsic value, cost, self-concept, self-efficacy, effort, math speed test, previous math grades, and teacher-rated effort ($d = .00$ – $.10$ between master's student and control conditions, $d = .01$ – $.19$ between teacher and control conditions, and $d = .09$ – $.19$ between master's student and teacher conditions), we included all of these variables as covariates in our analyses (see Table 2 for the descriptive statistics and the online supplemental materials for detailed effect sizes). Because there was a higher percentage of girls in the waitlist control condition (59.0%) compared with the intervention conditions (51.4% and 51.7%), we additionally controlled for gender. The effects of the covariates at both levels were freely estimated to account for contextual effects (Marsh et al., 2009). The covariates at the student level were group-mean centered (Enders & Tofighi, 2007), and manifest aggregation was used for the class-level predictors (Marsh et al., 2009). To facilitate the interpretation of the results, all continuous variables were standardized before running the analyses. Thereby, the regression coefficients of the dummy variables indicating the effects of the intervention conditions compared with the control condition can directly be interpreted as effect sizes (see Marsh et al., 2009; Tymms, 2004).

To compare implementation fidelity between the two intervention conditions (Research Question 3), we computed *t*-tests for the observer ratings (measured at the class level) and two-level regression analyses for the student ratings (student ratings nested in classrooms). To investigate associations between the measures of implementation fidelity and changes in utility value as the primary outcome of the intervention (Research Question 4), we conducted two-level regression analyses with students at Level 1 and classrooms at Level 2, in which we accounted for the nesting in schools using the design-based correction of standard errors in Mplus. Because implementation fidelity could not be assessed in the control condition, these analyses used only the subsample of students in the two intervention conditions ($n = 1,198$). The analyses were

thus set up to examine whether variations of implementation fidelity within the intervention conditions were associated with changes in the outcome. In these analyses, we regressed utility value at T2/T3 on utility value at T1, the measures of implementation fidelity, a dummy variable indicating the intervention condition (1 = master's student), and two interaction terms between the implementation fidelity measures and the intervention condition. For the observers' ratings, the implementation-fidelity-related variables were included only at the classroom level. For the students' ratings, the variables were included at both the individual and classroom levels. Additionally, we included the same covariates (i.e., gender, intrinsic value, cost, self-concept, self-efficacy, effort, math speed test, previous math grades, and teacher-rated effort) as for the analyses examining intervention effects.

For all our analyses, we used $p \leq .05$ as the inference criterion. As preregistered, one-tailed tests were used where we had formulated hypotheses (i.e., effects of the intervention conditions as compared with the control condition for preregistered outcomes and preregistered analyses regarding implementation fidelity). Two-tailed tests were used when we had no *a priori* hypotheses (i.e., exploratory outcomes, differences in the effects of the two intervention conditions, and exploratory implementation fidelity analyses).

We investigated effects of the intervention based on intention-to-treat analyses. That is, the only inclusion criterion was parental consent for the students, and no further exclusions were made. Full information maximum likelihood estimation was used to deal with missing data (Graham, 2009), which resulted from the absence of students at single measurement waves and nonresponse to single items (see Table 2 for the amount of missing data). Teacher-rated effort was additionally missing at the class level because some teachers did not fill out the ratings for their class at single time points (four classes at the pretest, three classes at the posttest, and one class at the follow-up). For the analyses with the curricular math test, we excluded the five classes ($n = 117$ students) that did not take this test because they had not covered the respective topics in class.

Results

Effects of the Intervention When Implemented by Master's Students and Teachers

Descriptive statistics for all major study variables at the different measurement waves by experimental condition are displayed in Table 2. Correlations between all major study variables can be found in the online supplemental materials. The effects of the two intervention conditions as compared with the control condition are presented in Table 3.

⁶We had originally preregistered three-level analyses in which intervention effects would have been allowed to vary between schools. However, when we included the set of covariates for which we found pretest differences, a three-level model with random slopes allowing for the variation of intervention effects between schools did not converge. We therefore chose to account for the nesting of classrooms within schools using the design-based correction of standard errors as recommended in the statistical literature (McNeish et al., 2017).

Table 2*Descriptive Statistics in the Three Conditions at All Measurement Waves*

Measure	Master's student (<i>n</i> = 629, 51.4% female)			Teacher (<i>n</i> = 569, 51.7% female)			Control (<i>n</i> = 546, 59.0% female)		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Utility value									
T1	595	2.83	0.50	558	2.83	0.50	526	2.85	0.50
T2	587	2.83	0.51	515	2.83	0.51	519	2.78	0.48
T3	557	2.78	0.51	520	2.77	0.52	489	2.75	0.50
Attainment value									
T1	595	2.86	0.60	555	2.88	0.59	528	2.88	0.61
T2	584	2.89	0.61	515	2.90	0.62	521	2.91	0.61
T3	553	2.86	0.63	520	2.88	0.62	488	2.89	0.63
Intrinsic value									
T1	590	2.50	0.85	546	2.42	0.86	521	2.48	0.88
T2	569	2.43	0.85	511	2.42	0.84	508	2.50	0.84
T3	536	2.41	0.84	514	2.36	0.83	476	2.47	0.86
Cost									
T1	596	2.08	0.67	558	2.19	0.72	526	2.10	0.68
T2	587	2.09	0.68	515	2.20	0.73	519	2.02	0.68
T3	557	2.10	0.68	519	2.18	0.70	490	2.04	0.68
Self-concept									
T1	597	2.82	0.77	552	2.72	0.77	528	2.79	0.76
T2	584	2.82	0.76	515	2.73	0.80	519	2.84	0.78
T3	554	2.83	0.76	514	2.72	0.77	486	2.79	0.76
Self-efficacy									
T1	596	2.80	0.63	555	2.71	0.62	528	2.79	0.61
T2	586	2.88	0.66	515	2.83	0.69	520	2.91	0.67
T3	553	2.90	0.67	519	2.82	0.68	487	2.92	0.64
Effort									
T1	596	3.27	0.57	554	3.32	0.60	523	3.27	0.57
T2	580	3.22	0.60	512	3.27	0.65	516	3.27	0.63
T3	549	3.21	0.62	513	3.19	0.67	481	3.22	0.63
Importance of effort									
T1	593	2.84	0.64	553	2.80	0.66	523	2.86	0.67
T2	582	2.91	0.63	512	2.85	0.68	516	2.83	0.68
Importance of talent									
T1	587	2.23	0.68	547	2.28	0.74	513	2.23	0.72
T2	579	2.19	0.70	508	2.30	0.73	512	2.26	0.73
Teacher-rated effort									
T1	560	3.06	0.76	534	2.99	0.77	509	3.00	0.82
T2	601	3.03	0.81	530	3.02	0.84	540	2.96	0.83
T3	624	3.05	0.76	532	3.01	0.78	539	3.04	0.82
Math speed test									
T1	597	29.29	7.63	557	28.62	7.85	529	29.18	7.67
T3	565	32.73	8.42	521	31.66	8.51	501	32.28	8.05
Curricular math test T3	504	0.09	0.82	503	−0.04	0.90	474	−0.06	0.86
Math grade previous year	620	2.73	1.00	557	2.92	0.94	528	2.74	1.01

Note. T = time point.

For utility value, the target outcome of the intervention, we found positive effects of both intervention conditions at the posttest and the follow-up, although the effects at the follow-up were smaller and only significant when using one-tailed testing as preregistered. These effects did not differ by intervention condition. No effects on attainment value and intrinsic value were observed. However, students in both intervention conditions reported higher cost at both the posttest and the follow-up as compared with the control condition controlling for the initial levels and other covariates. We had not preregistered any hypotheses about cost. As was found for the effects on utility value, these effects on cost did not depend on the intervention condition.

To gain a deeper understanding of the intervention effects on students' values, we additionally explored intervention effects on subfacets of utility value, attainment value, and cost (see the online supplemental materials). We note that these analyses were not preregistered and are thus exploratory. For subfacets of utility value, we found positive effects of both intervention conditions on general utility at the posttest and of the teacher condition at the follow-up. We also found positive effects of the two intervention conditions on utility for job at both time points. For utility for daily life, positive effects were found only in the teacher condition at both the posttest and the follow-up. No positive effects on utility for school were observed, but rather there was a negative effect of the teacher condition at the follow-up. For subfacets of attainment

Table 3*Effects of the Two Intervention Conditions on Student Outcomes at Posttest and Follow-Up*

Time	Utility value			Attainment value			Intrinsic value			Cost			Importance of effort			Importance of talent		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Posttest																		
Master's student	.15	.06	.018*	.03	.05	.595	-.08	.06	.161	.11	.05	.026*	.16	.05	.001*	-.11	.05	.020*
Teacher	.18	.06	.003*	.04	.06	.465	-.03	.07	.741	.15	.07	.030*	.15	.08	.067	-.01	.06	.854
Follow-up																		
Master's student	.10	.06	.098 [†]	-.03	.06	.635	-.11	.06	.053	.11	.04	.007*						
Teacher	.09	.05	.075 [†]	-.01	.06	.860	-.07	.06	.220	.10	.04	.011*						
Time	Self-concept			Self-efficacy			Effort			Teacher-rated effort			Math speed test			Curricular math test		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Posttest																		
Master's student	-.03	.05	.499	-.07	.05	.151	-.07	.07	.325	.04	.05	.400						
Teacher	-.07	.06	.288	-.04	.07	.530	-.03	.07	.668	.09	.05	.073 [†]						
Follow-up																		
Master's student	.00	.05	.980	-.04	.05	.475	-.06	.08	.425	-.03	.05	.638	.11	.06	.073 [†]	.08	.10	.392
Teacher	-.04	.06	.512	-.07	.06	.278	-.14	.07	.029*	-.02	.05	.640	.10	.06	.132	.02	.13	.890

Note. Standardized regression coefficients represent effects of the intervention conditions as compared with the control condition and can be interpreted like effect sizes. Coefficients are taken out of two-level regression analyses in which the respective pretest score, gender, previous math grade, intrinsic value, cost, self-concept, self-efficacy, effort, teacher-rated effort, and math speed test at T1 were controlled for at both levels. Importance of effort and importance of talent were not assessed at the follow-up, and the math speed test and the curricular math test were not assessed at the posttest. We applied one-tailed testing for outcomes with preregistered hypotheses (i.e., utility value, attainment value, intrinsic value, self-concept, self-efficacy, effort, teacher-rated effort, math speed test, and curricular math test) and two-tailed testing for outcomes without preregistered hypotheses (i.e., cost, importance of effort, and importance of talent).

* $p < .05$ (two-tailed). [†] $p < .05$ (one-tailed).

value, no significant intervention effects were observed, although there were opposite tendencies for the subfacets (negative for importance of achievement and positive for personal importance). For subfacets of cost, we found that students reported higher emotional cost at the posttest and the follow-up in the two intervention conditions, and higher effort required in the master's student condition at the follow-up when controlling for the initial levels and covariates. No significant effects were observed for opportunity cost.

Although we had not preregistered analyses for students' mindsets, we investigated these as a further outcome targeted in the first part of the intervention (i.e., the psychoeducational presentation). In line with the message of this part of the intervention, we found that students in classes in which the intervention was delivered by master's students reported a higher importance of effort and a lower importance of talent as compared with the control condition controlling for the initial levels and covariates. When using two-tailed testing, no effects of the teacher condition on students' mindsets were found.

Furthermore, no effects of the two intervention conditions on self-concept and self-efficacy were observed at the posttest or the follow-up. For self-reported effort, against our hypotheses, we found a negative effect in the teacher condition at the follow-up. For teacher-rated effort, in line with our hypotheses, we found that teachers in classes in which the intervention was delivered by the teachers rated their students' effort as higher at the posttest (using one-tailed testing) as compared with the control condition. This effect, however, could no longer be observed at the follow-up. Finally, we found that students in classes in which the intervention was delivered by master's students showed higher

performance on the math speed test at the follow-up as compared with the control condition controlling for initial levels and covariates (using one-tailed testing as preregistered). The effect of the teacher condition, however, was not significant. No effects were found for the curricular math test.

Implementation Fidelity

In line with our research questions regarding implementation fidelity, we additionally investigated differences in implementation fidelity between the two intervention conditions (see Table 4). As expected, the master's students showed a higher adherence to the intervention script than the teachers. There were no significant differences in discipline problems as an indicator of classroom management during the intervention, although, against our expectations, there was a tendency for more discipline problems when the intervention was implemented by the regular math teacher. Furthermore, additional exploratory analyses (see the online supplemental materials) suggested that the raters observed greater clarity of instruction and greater authenticity in the master's student condition but a higher degree of class participation in the teacher condition. The student ratings universally suggested higher quality of delivery and greater interest in the intervention when the intervention was delivered by master's students in comparison with teachers.

Finally, we examined whether implementation fidelity was associated with changes in utility value in the two intervention conditions (see Table 5). When regressing utility value in the two intervention conditions at the posttest onto measures of implementation fidelity, both a higher adherence and lower levels of discipline problems during the intervention predicted more positive changes in utility value. In addition, there was a significant negative effect of the master's student

Table 4*Mean Differences in Measures of Implementation Fidelity Between the Two Intervention Conditions*

Variable	Master's student			Teacher			Master's student – Teacher			
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
Adherence	28	9.05	0.55	24	7.29	1.34	6.01	29.57	<.001	1.72
Discipline problems	28	1.71	0.59	24	1.92	0.74	–1.11	44.00	.272	–0.31

Note. Differences between the two intervention conditions were evaluated using *t*-tests assuming unequal variances (as indicated by a Levene test of variance homogeneity).

condition as compared with the teacher condition. That is, when controlling for implementation fidelity, a larger effect of the teacher condition on utility value compared with the master's student condition would be expected. The same pattern of results was observed for the follow-up, although the regression coefficient for discipline problems was only significant when using one-tailed testing as preregistered, and the regression coefficient for the intervention condition was no longer significant. Additional exploratory analyses (see the online supplemental materials) did not provide evidence that the other investigated aspects of quality of delivery or student responsiveness as rated by the observers were associated with changes in utility value in the two intervention conditions. However, among the students' ratings of quality of delivery, individual students' perceptions of a supportive climate (at posttest), authenticity, as well as autonomy support emerged as positive predictors of changes in utility value. Autonomy support was also a significant predictor at the classroom level for the posttest. Furthermore, students' interest in the intervention positively predicted changes in utility value at both the individual and class levels. However, for the posttest, the predictive effect at the class level was further qualified by an interaction with the intervention condition, indicating that this predictive effect was limited to the teacher condition.

Exploratory Mediation Analyses

To better understand the psychological processes underlying the intervention, we conducted exploratory analyses in line with our logic model. We investigated whether effects of the intervention on utility value, cost, and achievement at T3 were mediated through motivational variables (i.e., growth mindsets,

expectancies, utility value) at T2 using cross-level mediation models (Pituch & Stapleton, 2012), in which we controlled for the mediator and outcome at the pretest as well as all other covariates that were included in the analyses for the main effects. The detailed results can be found in the online supplemental materials. For achievement, students' mindsets and utility value at T2 failed to explain achievement at T3 in these analyses (including the abovementioned set of covariates). Consequently, there were also no significant indirect effects for any of these potential mediators. Students' expectancies significantly predicted achievement at T3. For utility value as an outcome, we found small, positive cross-level indirect effects of the two intervention conditions through importance of effort (but no total indirect effects). For cost, we found small, negative cross-level indirect effects of the two intervention conditions through utility value and a tendency for a negative cross-level indirect effect of the master's student condition through importance of talent. Because these indirect effects of cost were negative, increases in utility value and decreases in importance of effort explained decreases in cost (if at all) and thus failed to explain the increases in cost that we found in the total effects.

Exploratory Moderation Analyses

Because students' expectancies were shown to be an important moderator of the effects of relevance interventions in previous research, we additionally explored such moderation effects. To do so, we added two cross-level interaction terms between the two intervention conditions and students' expectancies (i.e., self-concept or self-efficacy at pretest) into our multilevel regression analyses. Of the 80 interactions that we tested with self-concept and self-efficacy as moderating variables (20 outcomes \times intervention

Table 5

Results of Multilevel Regression Analyses Regressing Utility Value at Posttest and Follow-Up in the Intervention Conditions on Different Implementation Fidelity Measures

Variable	Utility value T2			Utility value T3		
	β	<i>SE</i>	<i>p</i>	β	<i>SE</i>	<i>p</i>
Utility value T1	.90	.10	<.001	.81	.11	<.001
Adherence	.16	.04	<.001	.12	.04	.001
Discipline problems	–.11	.03	.001	–.07	.04	.053
Intervention condition (1 = Master's student)	–.26	.10	.011	–.16	.10	.114
Adherence \times Intervention Condition	–.04	.10	.716	–.08	.10	.457
Discipline Problems \times Intervention Condition	.09	.06	.163	–.02	.07	.787

Note. T = time point. Class-level regression coefficients are taken out of two-level regression analyses (with standard errors adjusted for the nesting within schools) using only the students in the two intervention conditions. The models included a number of additional covariates at the student and the class level (i.e., gender, previous math grade, intrinsic value, cost, self-concept, self-efficacy, effort, teacher-rated effort, and math speed test at T1). For the sake of clarity, only the regression coefficients of the variables of interest are presented here.

conditions \times moderators), we found that six of the interaction terms were statistically significant at $p \leq .05$ (see the online supplemental materials). For self-concept as a moderating variable, we found statistically significant negative interactions on students' intrinsic value for the teacher condition at the follow-up and on students' self-concept for the master's student condition at the posttest, both indicating that students with high self-concepts tended to experience negative effects of the intervention, whereas there were no effects for students with low self-concepts. Furthermore, there was a significant positive interaction on students' effort for the teacher condition at the follow-up, indicating that negative effects of the intervention were found for students with low self-concept but not for those with high self-concept. For self-efficacy as a moderating variable, we found significant negative interactions on students' attainment value, importance of effort, and self-concept for the master's student condition at the posttest, indicating positive effects only for students with low self-efficacy (for attainment value and importance of effort) or negative effects only for students with high self-efficacy (for self-concept). Thus, with only one exception, the significant interaction terms were all negative and thus indicated more beneficial (or less harmful) effects of the intervention for students with low expectancies. However, because of the number of interactions tested and the exploratory nature of these analyses, these interactions should be interpreted with caution.

Discussion

With this study, we investigated the effectiveness of a relevance intervention in mathematics classrooms with a cluster-randomized trial in which we tested the potential scalability of this intervention, after its efficacy had been shown in a previous trial. To do so, the 90-min intervention was either delivered by trained master's students or the regular math teachers. Our study had four major findings regarding the effectiveness of the intervention. First, although we observed less positive and smaller effects than in the efficacy trial, we were able to replicate some of the positive effects of the intervention found previously, namely on utility value (in both conditions) and the math speed test (only in the master's student condition). Second, we observed positive effects of the intervention on students' growth mindsets (i.e., increased importance of effort and reduced importance of talent) when the intervention was delivered by master's students, identifying another potential active ingredient of the intervention. Third, we also observed some unintended and unexpected effects of the intervention, namely that the intervention increased perceived cost consistently across both intervention conditions and time points and undermined self-reported effort in one of these conditions in the long term. Fourth, we compared master's students and the regular math teachers in their effectiveness in delivering the intervention, finding only minor differences between the two conditions. We thus found that both master's students and math teachers can implement the intervention with positive effects on the target outcome (i.e., utility value), but the overall pattern of mixed effects also raises concerns regarding the scale up of the intervention.

In addition, we investigated implementation fidelity in the two intervention conditions, and found that, as expected, master's students showed a higher degree of adherence. Moreover, students perceived a higher quality of delivery and reported greater interest

when the intervention was delivered by master's students. Higher adherence and fewer discipline problems as rated by the observers as well as higher autonomy support and interest in the intervention as perceived by the students were related to more positive changes in utility value in the intervention conditions. We thereby identified important sources of variation in the effectiveness of the intervention across conditions, although the higher levels of adherence, perceived quality of delivery, and interest in the master's student condition did not translate into larger intervention effects. We discuss these findings further in the following sections.

Replicating Effects of the Efficacy Trial

Our major aim was to test whether the results of the previous efficacy trial could be replicated when the intervention is delivered by master's students and the regular math teachers. We found positive effects of the 90-min intervention on students' utility value until three months after the intervention in both intervention conditions, thereby replicating results from the previous efficacy trial (Brisson et al., 2017; Gaspard, Dicke, Flunger, Brisson, et al., 2015). Such findings should not be taken for granted as they show that relevance interventions can have long-lasting effects on students' utility value even when the intervention is implemented by their regular math teachers who only receive a short training, thereby enabling scaling up of the intervention. When comparing the results for the different utility facets, the largest effects were found for general utility and utility for job in line with the findings of the efficacy trial (Gaspard, Dicke, Flunger, Brisson, et al., 2015) and the focus of the intervention. In the master's student condition, we additionally found that students showed higher performance on a standardized achievement test (i.e., the speed test also used by Brisson et al., 2017) 3 months after the intervention, further attesting to the effectiveness of the intervention under this condition.

Regarding the practical implications, the effects of the intervention found in our study would be judged as small by conventional standards (Cohen, 1988). However, the intervention consisted of a 90-min session in math classrooms and can thus be considered a minimal intervention. Using recently proposed benchmarks for educational interventions, these effect sizes would be considered medium (Kraft, 2020). Given the design of our study (a large, pre-registered randomized trial with meaningful outcomes measured months after the intervention), the relatively low monetary cost of the intervention, and its reasonable potential for scalability as a scripted intervention, we conclude that even seemingly small effects such as the ones found in our study can be considered meaningful.

At the same time, the observed effects on utility value were smaller compared with the efficacy trial, and effects on more distal outcomes such as attainment and intrinsic value, expectancies, and teacher-rated effort (for which we only found a small effect in the teacher condition at posttest) could not be replicated. A lack of effects on students' expectancies, in particular, might have contributed to the limited effectiveness of the intervention overall. Prior research has found that relevance interventions can have positive effects on students' expectancies at least for some groups of students and that expectancies can mediate effects of relevance interventions on achievement (Brisson et al., 2017; Hulleman et al., 2017; Rosenzweig et al., 2020). Our exploratory mediation

analyses suggested that students' expectancies predicted their utility value and achievement at the follow-up, and, thus, larger effects of the intervention might have been expected if the intervention had been effective in fostering students' expectancies. Students' self-concept as well as their self-efficacy were found to be highly stable in this study, and the components of the intervention targeting students' expectancies might not have been strong enough to affect students' expectancies in the classroom context in the long term, especially if students do not perceive an increase in their performance (see also Brisson et al., 2017).

In general, such a reduced effectiveness compared with the efficacy study could have been expected because the intervention was delivered by a larger group of trained individuals (as compared with only five individuals who were involved in the development of the intervention) and implementation fidelity typically tends to decrease and show higher variation under such conditions (Durlak & DuPre, 2008; Hulleman & Cordray, 2009; O'Donnell, 2008; Weiss et al., 2014). Among the two groups who implemented the intervention in this effectiveness study, such differences could easily be expected for the math teachers. The master's students, however, share many similarities with the doctoral candidates who delivered the intervention in the efficacy study (e.g., similar age and educational background) and also received intensive training. The ninth-grade students could have viewed them as potential role models and seen them as experts from the university. This might also explain the differences between the interventions delivered by the master's students and the teachers in students' perceptions of quality of delivery and their interest in the intervention. Still, the master's students were not as involved in the development of the intervention and the design of the study as the doctoral candidates were. They were also a less selective group because they did not (yet) successfully apply to a doctoral program and the ninth-grade students might have seen them as less of an expert on the research topics presented in the intervention. In addition, small changes were made in the intervention materials compared with the efficacy study with the goal to optimize the materials. All of these differences between the efficacy and the effectiveness studies could potentially have contributed to smaller effects of the intervention. However, variation in implementation fidelity in the two intervention conditions (i.e., adherence, quality of delivery, and student responsiveness) predicted changes in utility value, thereby supporting the likely role that implementation fidelity played in that regard. Thus, even if training and deploying master's students to classrooms as part of their course work would be feasible on a larger scale, it is the lack of large effects that critically highlights the question of whether one should rely on master's students for this intervention.

Extending Outcomes: Effects on Students' Growth Mindsets

Extending the outcomes that were investigated in the efficacy study, we observed effects of the intervention on students' mindsets (i.e., an increase in importance of effort and a decrease in importance of talent) when the intervention was delivered by master's students. In line with the message of the intervention, this can be interpreted to mean that the students' mindsets became more growth-oriented and less fixed. Growth mindset interventions have shown positive effects on students' achievements, at least under specific conditions and for some groups of students (Blackwell et al., 2007; Sisk et al., 2018; Yeager et al., 2016,

2019). The first part of the intervention, in which students were presented with research results on the importance of effort for achievement in mathematics, might therefore be an active ingredient of the intervention. This first part of the intervention was included to buffer potential detrimental effects of the intervention for students with low expectancies (Durik et al., 2015). Because the different components of the intervention were not tested separately in this study and the efficacy study, their effects cannot be teased apart. Future research should therefore examine whether combining growth mindset and relevance interventions is more effective than applying only one of them.

Unexpected Effects of the Intervention on Cost and Self-Reported Effort

We also observed some unintended effects of the intervention, namely the increase in cost that was observed in both intervention conditions and at both posttest and follow-up and the decrease in self-reported effort that was found in the teacher condition at the follow-up. These effects were not observed in the efficacy trial (Gaspard, Dicke, Flunger, Brisson, et al., 2015, 2016) and were therefore not expected. They are also not in line with the negative correlation between utility value and cost and the positive correlation between utility value and effort, which were found in this study and previous research (e.g., Gaspard, Dicke, Flunger, Schreier, et al., 2015; Perez et al., 2019; Song et al., 2020; Trautwein et al., 2012).

Among the cost facets, our additional analyses suggested that the effects were most pronounced for emotional cost. These unintended effects on cost might be attributable to the slight changes in the intervention materials, which were made after the efficacy trial, such as the incorporation of take-home messages in the first part of the intervention that further emphasized the importance of effort. Alternatively, they could also be attributable to how the intervention was implemented by master's students and teachers in the classroom. Both master's students and teachers were instructed to frame the intervention in an autonomy-supportive way (i.e., students were supposed to give feedback about the intervention to improve it for future use and to provide their own perspective in the relevance-inducing tasks) to avoid reactance in the students or the feeling that they needed an intervention (Vansteenkiste et al., 2018; Yeager & Walton, 2011). However, it could be that not all master's students and teachers fully applied this framing and that subtle variations in the framing of the intervention can make the direct communication of utility value and the importance of effort for mathematics threatening to the students and lead them to perceive mathematics as requiring effort and creating anxiety (see also Canning et al., 2019). Exploratory mediation analyses suggested that positive effects on utility value and negative effects on importance of talent (in the master's student condition) rather contributed to reduced cost in comparison with the control condition. However, students' expectancies were observed to be important predictors of students' cost, and a lack of effects on expectancies might partially explain the observed effects on cost. These processes should be explored in more detail in future research.

The negative effect on self-reported effort in the teacher condition at the follow-up, which was not matched by a similar effect on teacher-reported effort, could potentially be explained by a change in students' standards. That is, the students might have

realized the importance of effort to perform well in mathematics. Instead of changing their effort, however, they might just have adjusted their own ratings of the effort they were investing as less than what would be ideally needed to perform up to their potential. Other studies have found that students do not necessarily alter their effort after realizing that more effort is needed as a result of intervention but rather adjust their expectations (Oreopoulos & Petronijevic, 2019). Similar processes might have occurred here. The teachers, on the other hand, who had delivered the intervention in the classroom, might have expected an increased effort in their students. In line with this expectation, they also rated their students' effort as higher at the posttest but then might have been disappointed with the seemingly small effect of the intervention and possibly even communicated this perceived lack of effort to their students.

Comparing the Delivery of the Intervention by Master's Students or Math Teachers

When comparing master's students and the regular math teachers in their effectiveness in delivering the intervention, we found only minor differences between the two conditions. This could be interpreted to mean that it is not necessary to deploy persons from university to the classroom to deliver the intervention, but that the teachers can do so successfully themselves. However, we also observed some differences in implementation fidelity with master's students, who bring a deeper theoretical background on the psychology of motivation and the requirements of randomized trials, and who had received extensive training, showing a higher degree of adherence as well as a higher quality of delivery (as rated by the students and partly also by the observers) and creating a higher level of interest in the intervention on the side of the participating students. The relative advantage in students' perceptions of quality of delivery and interest could potentially also be explained by novelty and role model effects (i.e., being taught by a young adult instead of the regular math teacher). These differences in implementation fidelity did not seem to transfer to differential effects on utility value, but they might explain the differences observed for effects on students' mindsets.

Although a high degree of adherence as well as students' perceptions of quality of delivery and their interest in the intervention contributed positively to the effectiveness of the intervention regarding utility value across the two intervention conditions, stronger effects of the master's student condition would have been expected given the levels of implementation fidelity. The effect of having a young adult as a potential role model talking about the usefulness of math instead of the regular math teacher thus did not have a larger effect in the longer run. It could be that the math teachers achieved the same goal through different means. For instance, it could be that the teachers were more convinced about the usefulness of mathematics and genuinely interested in their students realizing its relevance. Because of this interest and drawing on their expertise in terms of pedagogy and practice, they might have made adaptations that were beneficial, such as incorporating examples of the usefulness of mathematics tailored to their students' interests or the topics currently being covered in class. Although master's students and teachers used the same intervention materials, it could be that the teachers made additional connections with students' daily lives. This points to the ongoing

debate in the literature about the right mix of implementation fidelity and adaptation (Dane & Schneider, 1998; Durlak & DuPre, 2008; Van Daele et al., 2014). Research suggests that adaptations can be positive as long as the core components of the intervention are implemented as intended. It needs to be noted that adherence was high across both intervention conditions, which potentially was a precondition for the effects on utility value observed in both conditions. Another important question for future research is how teachers need to be trained to implement interventions with high fidelity and modify them in ways that keep the core components intact (Greene, 2015).

Strengths and Limitations

This study has several strengths that should increase confidence in the robustness of the reported results. First, we conducted a large cluster-randomized trial with adequate power to detect effects of the intervention at the classroom level. By randomizing classrooms and implementing the intervention at the classroom level, the risk of diffusion effects was reduced (Craven et al., 2001; Plewis & Hurry, 1998). Second, we preregistered our hypotheses, design, and analyses to increase transparency, to reduce our degrees of freedom in analyzing the data and thus minimize the potential bias attributable to *ex post* choices, and to thereby increase confidence in the validity of our findings (Wicherts et al., 2016). Third, we assessed a broad set of outcome measures, including self-reported motivation, teacher ratings, and standardized achievement tests to be able to uncover intended and unintended effects of the intervention on students' academic outcomes. Finally, we incorporated different aspects of implementation fidelity (rated from different perspectives) to be able to understand sources of variation in the effectiveness of the intervention.

However, our study also has several limitations that should be borne in mind when interpreting its findings. First, our sample was limited to ninth-grade students in the academic track in one region in Germany. As should be the case for well-crafted interventions, the intervention was developed with the target population in mind and included customized examples on the usefulness of mathematics. For example, because most students from the academic track go on to attend university, the intervention included information about university majors that required math skills. If the intervention was to be adapted for other tracks, it would probably need to focus more on examples of math-intensive vocational jobs, which are more often pursued by this population because they do not require a university entrance qualification but tend to be well-respected and well-paid in the German context (for an exemplary relevance intervention in this context, see Piesch et al., 2018). This context-specificity of relevance interventions also makes it challenging to replicate findings across contexts and populations and to provide evidence for mechanisms that generalize across those (Harackiewicz & Priniski, 2018; Yeager & Walton, 2011). Future research therefore needs to continue to investigate the effects of relevance interventions in different contexts and populations.

Second, although we aimed to test the intervention under realistic conditions, and both the master's students and teachers were able to deliver it with high levels of adherence as rated by the observers, the observation in the classroom could also have been a precondition of the high levels of fidelity in this study. The teachers (as well as the master's students) were informed that the goal was to test the

intervention as it is and that the observations were made for the purpose of investigating differences in the implementation between persons and classes. Research on teaching quality seems to indicate that teachers and students forget about observers after a few minutes (Praetorius et al., 2017). Still, future research is needed to test whether teachers would make more adaptations without such an observation and if this would lead to reduced effects of the intervention (for discussions about adaptation & fidelity, see Dane & Schneider, 1998; Greene, 2015; Van Daele et al., 2014).

Third, when interpreting the effects of the intervention that were found in this study, we referred to effects in comparison with the control condition when controlling for the initial levels of the outcomes and other covariates, which can be considered to be the most valid estimates of intervention effects from a methodological point of view (What Works Clearinghouse, 2020). However, these results are not necessarily aligned with the mean-level development over the school year within the conditions. In terms of the trajectory over time, we observed a decline in utility value in the control condition as has been found in previous research, including the efficacy trial (Gaspard, Dicke, Flunger, Brisson, et al., 2015; Watt, 2004). Instead of lifting utility value overall, the intervention thereby functioned as a buffer against this negative development (see also Hulleman et al., 2010, 2017).

Fourth, the relevance intervention implemented in this study consisted of different components and yielded some intended and some unintended effects. Moreover, although the logic model of the intervention was partially supported in that we found positive effects of the intervention on students' growth mindsets and utility value and students' expectancies predicted achievement as a distal outcome, we did not find effects of the intervention on students' expectancies and our exploratory mediation analyses could not explain the positive effects found on achievement in the master's student condition. More research is therefore needed to examine the mechanisms underlying the effects of relevance interventions and which components need to be included to benefit different groups of students. However, testing different combinations of intervention components in randomized field trials requires large sample sizes. It can also be challenging to assess process measures providing an insight into mechanisms in the field. Researchers might therefore want to return to the lab to examine such questions (for a similar conclusion, see Harackiewicz & Priniski, 2018). Furthermore, the exploratory analyses in which we investigated students' expectancies as moderators of the intervention effects provided only limited evidence for such moderation effects. In line with prior research using relevance-inducing reflection tasks (Hulleman & Harackiewicz, 2021), the moderation effects that we found mostly pointed toward more beneficial effects of the intervention for students with low expectancies. However, the overall pattern (i.e., only a few significant interactions) could also be interpreted to mean that the intervention had more or less similar effects for students with low and high expectancies. Future research might still want to conduct more comprehensive moderation analyses, possibly including latent profile analyses that can yield insights into profiles among a set of motivational variables and thereby also generate knowledge about the benefits and drawbacks of this type of intervention for subgroups of students (see Binning & Browman, 2020). Based on such knowledge about how individual students respond to motivation interventions, a potential approach for future research could involve targeting such inter-

ventions toward individual students on the basis of an assessment of their motivational needs instead of a universal prevention program such as the one applied in this study.

Fifth, the outcome measures included in our study also had some limitations. Students in Germany are not administered state-based standardized achievement tests in ninth grade, and we were therefore only able to use a standardized speed test and a self-developed curricular math test as achievement measures at the follow-up. Because teachers are free to decide the order in which they teach different topics during the school year, there was a high degree of missing data on the curricular math test, which undermined its reliability and the precision of the estimates of intervention effects on this outcome measure. Furthermore, the waitlist control design that we used in our study made it impossible to investigate long-term effects of the intervention. Because students in the academic track schools in our sample have no choice regarding the amount or level of math classes, we were also not able to investigate effects of the intervention on students' choices (for effects of the intervention on precursors of students' career choices, see Piesch et al., 2020).

Conclusion

To conclude, on the basis of a rigorous evaluation of the effectiveness of a short relevance intervention in the classroom, we found that both master's students and math teachers delivered the intervention with high levels of fidelity and with positive effects on the target outcome (i.e., utility value). This can be seen as an important step toward scaling up, especially considering that the teachers received a limited amount of training. However, these positive effects on utility value need to be weighed against the other undesirable effects of the intervention observed in this study, most notably an increase in students' perceived cost, as well as a lack of effects on some other outcomes (e.g., expectancies, teacher-rated effort). Furthermore, although finding effects of a 90-min intervention on outcomes measured up to 3 months later should not be taken for granted, the effects that were observed in this study can be considered small to medium (depending on the criteria applied; see Kraft, 2020). As such, it is not clear whether it would be worthwhile for teachers to engage in this intervention. Our analyses revealed some factors that contributed to larger effects of the intervention, including high adherence, a lack of discipline problems, and support for students' autonomy during the intervention. Future research should continue to examine the mechanisms underlying the effectiveness of these interventions and the conditions under which they can be successfully implemented in practice so that they can be successfully brought to scale.

References

- Autorengruppe Bildungsberichterstattung. (2018). *Bildung in Deutschland 2018: ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung* [Education in Germany 2018: An indicator-supported report with an analysis of effects and returns of education]. wbv Media GmbH & Co. KG. <https://doi.org/10.3278/6001820fw>
- Bandura, A. (1977). *Social learning theory*. Prentice Hall.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.
- Barron, K. E., & Hulleman, C. S. (2015). Expectancy-value-cost model of motivation. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (2nd ed., pp. 503–509). Elsevier. <https://doi.org/10.1016/B978-0-08-097086-8.26099-6>

- Binning, K. R., & Browman, A. S. (2020). Theoretical, ethical, and policy considerations for conducting social-psychological interventions to close educational achievement gaps. *Social Issues and Policy Review*, 14(1), 182–216. <https://doi.org/10.1111/sipr.12066>
- Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, 78(1), 246–263. <https://doi.org/10.1111/j.1467-8624.2007.00995.x>
- Bong, M., & Skaalvik, E. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, 15(1), 1–40. <https://doi.org/10.1023/A:1021302408382>
- Brisson, B. M., Dicke, A.-L., Gaspard, H., Häfner, I., Flunger, B., Nagengast, B., & Trautwein, U. (2017). Short intervention, sustained effects: Promoting students' math competence beliefs, effort, and achievement. *American Educational Research Journal*, 54(6), 1048–1078. <https://doi.org/10.3102/0002831217716084>
- Brisson, B. M., Hulleman, C. S., Häfner, I., Gaspard, H., Flunger, B., Dicke, A.-L., Trautwein, U., & Nagengast, B. (2020). Who sticks to the instructions—And does it matter? Antecedents and effects of students' responsiveness to a classroom-based motivation intervention. *Zeitschrift für Erziehungswissenschaft*, 23(1), 121–144. <https://doi.org/10.1007/s11618-019-00922-z>
- Brophy, J. (1999). Toward a model of the value aspects of motivation in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist*, 34(2), 75–85. https://doi.org/10.1207/s15326985ep3402_1
- Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*, 2(1), 49–68. <https://doi.org/10.1177/109442819921004>
- Canning, E. A., & Harackiewicz, J. M. (2015). Teach it, don't preach it: The differential effects of directly-communicated and self-generated utility-value information. *Motivation Science*, 1(1), 47–71. <https://doi.org/10.1037/mot0000015>
- Canning, E. A., & Harackiewicz, J. M. (2019). Utility value and intervention framing. In K. Ann Renninger & S. Hidi (Eds.), *The Cambridge handbook of motivation and learning* (pp. 645–662). Cambridge University Press. <https://doi.org/10.1017/9781316823279.027>
- Canning, E. A., Harackiewicz, J. M., Priniski, S. J., Hecht, C. A., Tibbetts, Y., & Hyde, J. S. (2018). Improving performance and retention in introductory biology with a utility-value intervention. *Journal of Educational Psychology*, 110(6), 834–849. <https://doi.org/10.1037/edu0000244>
- Canning, E. A., Priniski, S. J., & Harackiewicz, J. M. (2019). Unintended consequences of framing a utility-value intervention in two-year colleges. *Learning and Instruction*, 62, 37–48. <https://doi.org/10.1016/j.learninstruc.2019.05.001>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- Craven, R. G., Marsh, H. W., Debus, R. L., & Jayasinghe, U. (2001). Diffusion effects: Control group contamination threats to the validity of teacher-administered interventions. *Journal of Educational Psychology*, 93(3), 639–645. <https://doi.org/10.1037/0022-0663.93.3.639>
- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, 18(1), 23–45. [https://doi.org/10.1016/S0272-7358\(97\)00043-3](https://doi.org/10.1016/S0272-7358(97)00043-3)
- Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology*, 99(3), 597–610. <https://doi.org/10.1037/0022-0663.99.3.597>
- Durik, A. M., Hulleman, C. S., & Harackiewicz, J. M. (2015). One size fits some: Instructional enhancements to promote interest. In K. A. Renninger, M. Nieswandt, & S. Hidi (Eds.), *Interest in mathematics and science learning* (pp. 49–62). American Educational Research Association. https://doi.org/10.3102/978-0-935302-42-4_3
- Durik, A. M., Shechter, O. G., Noh, M., Rozek, C. S., & Harackiewicz, J. M. (2015). What if I can't? Success expectancies moderate the effects of utility value information on situational interest and performance. *Motivation and Emotion*, 39(1), 104–118. <https://doi.org/10.1007/s11031-014-9419-0>
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41(3–4), 327–350. <https://doi.org/10.1007/s10464-008-9165-0>
- Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 105–121). Guilford Press.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>
- Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 74–146). Freeman.
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2), 121–138. <https://doi.org/10.1037/1082-989X.12.2.121>
- Ennemoser, M., Krajewski, K., & Schmidt, S. (2011). Entwicklung und Bedeutung von Mengen-Zahlen-Kompetenzen und eines basalen Konventions- und Regelwissens in den Klassen 5 bis 9 [Development and importance of quantity-number competencies and basic knowledge of mathematical conventions and rules in grades 5 through 9]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 43(4), 228–242. <https://doi.org/10.1026/0049-8637/a000055>
- Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff, J., Lüdtke, O., Polikoff, M. S., Klusmann, U., & Trautwein, U. (2020). Don't blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*, 112(6), 1284–1302. <https://doi.org/10.1037/edu0000416>
- Flunger, B., Hollmann, L., Hornstra, L., & Murayama, K. (2020). *It's more about a lesson than a domain: Lesson-specific autonomy support, motivation, and engagement in math and a second language* [Manuscript submitted for publication]. Utrecht University.
- Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. G. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence*, 20(2), 507–537. <https://doi.org/10.1111/j.1532-7795.2010.00645.x>
- Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 International standard classification of occupations. *Social Science Research*, 25(3), 201–239. <https://doi.org/10.1006/ssre.1996.0010>
- Gaspard, H., & Lauermann, F. (2020). *Emotionally and motivationally supportive classrooms: A state-trait analysis of lesson- and classroom-specific variation in teacher- and student-reported teacher enthusiasm and student engagement* [Manuscript submitted for publication]. Hector Research Institute of Education Sciences and Psychology, University of Tübingen.
- Gaspard, H., Dicke, A.-L., Flunger, B., Brisson, B. M., Häfner, I., Nagengast, B., & Trautwein, U. (2015). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology*, 51(9), 1226–1240. <https://doi.org/10.1037/dev0000028>
- Gaspard, H., Dicke, A.-L., Flunger, B., Häfner, I., Brisson, B. M., Trautwein, U., & Nagengast, B. (2016). Side effects of motivational interventions?

- Effects of an intervention in math classrooms on motivation in verbal domains. *AERA Open*, 2(2), 1–14. <https://doi.org/10.1177/2332858416649168>
- Gaspard, H., Dicke, A.-L., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, 107(3), 663–677. <https://doi.org/10.1037/edu0000003>
- Gaspard, H., Häfner, I., Parrisius, C., Trautwein, U., & Nagengast, B. (2017). Assessing task values in five subjects during secondary school: Measurement structure and mean level differences across grade level, gender, and academic subject. *Contemporary Educational Psychology*, 48, 67–84. <https://doi.org/10.1016/j.cedpsych.2016.09.003>
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576. <https://doi.org/10.1146/annurev.psych.58.110405.085530>
- Greene, J. A. (2015). Serious challenges require serious scholarship: Integrating implementation science into the scholarly discourse. *Contemporary Educational Psychology*, 40, 112–120. <https://doi.org/10.1016/j.cedpsych.2014.10.007>
- Harackiewicz, J. M., & Priniski, S. J. (2018). Improving student outcomes in higher education: The science of targeted intervention. *Annual Review of Psychology*, 69, 409–435. <https://doi.org/10.1146/annurev-psych-122216-011725>
- Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, 111(5), 745–765. <https://doi.org/10.1037/pspp0000075>
- Harackiewicz, J. M., Hulleman, C. S., Rozek, C. S., Katz-Wise, S., & Hyde, J. S. (2010). *Parents' understanding of the utility value of STEM courses for high school students* [Paper Presentation]. 2010 Biennial Meeting of the Society for Research on Adolescence.
- Harackiewicz, J. M., Tibbetts, Y., Canning, E., & Hyde, J. S. (2014). Harnessing values to promote motivation in education. In S. A. Karabenick & T. Urdan (Eds.), *Advances in motivation and achievement: Vol. 18. Motivational interventions* (pp. 71–105). Emerald Group Publishing Limited. <https://doi.org/10.1108/S0749-742320140000018002>
- Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, 70(2), 151–179. <https://doi.org/10.3102/00346543070002151>
- Hulleman, C. S., & Barron, K. E. (2016). Motivation interventions in education: Bridging theory, research, and practice. In L. Corno & E. M. Anderman (Eds.), *Handbook of educational psychology* (3rd ed., pp. 160–171). Routledge, Taylor and Francis.
- Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, 2(1), 88–110. <https://doi.org/10.1080/19345740802539325>
- Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, 326(5958), 1410–1412. <https://doi.org/10.1126/science.1177067>
- Hulleman, C. S., & Harackiewicz, J. M. (2021). The utility-value intervention. In G. M. Walton & A. J. Crum (Eds.), *Handbook of wise interventions: How social psychology can help people change* (pp. 100–125). Guilford Press.
- Hulleman, C. S., Durik, A. M., Schweigert, S. B., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology*, 100(2), 398–416. <https://doi.org/10.1037/0022-0663.100.2.398>
- Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, 102(4), 880–895. <https://doi.org/10.1037/a0019506>
- Hulleman, C. S., Kosovich, J. J., Barron, K. E., & Daniel, D. B. (2017). Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology*, 109(3), 387–404. <https://doi.org/10.1037/edu0000146>
- Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development*, 73(2), 509–527. <https://doi.org/10.1111/1467-8624.00421>
- Keller, M. M., Hoy, A. W., Goetz, T., & Frenzel, A. C. (2016). Teacher enthusiasm: Reviewing and redefining a complex construct. *Educational Psychology Review*, 28(4), 743–769. <https://doi.org/10.1007/s10648-015-9354-y>
- Kim, J. S. (2019). Making every study count: Learning from replication failure to improve intervention research. *Educational Researcher*, 48(9), 599–607. <https://doi.org/10.3102/0013189X19891428>
- Kosovich, J. J., Hulleman, C. S., Phelps, J., & Lee, M. (2019). Improving algebra success with a utility-value intervention. *Journal of Developmental Education*, 42(2), 2–10.
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Kreber, C., McCune, V., & Klampfleiter, M. (2010). Formal and implicit conceptions of authenticity in teaching. *Teaching in Higher Education*, 15(4), 383–397. <https://doi.org/10.1080/13562517.2010.493348>
- Kronberger, A., & Weizenegger, T. (2009). *WADI - Wachhalten und Diagnostizieren von Grundkenntnissen und Grundfertigkeiten im Fach Mathematik Klassenstufe 9/10 Teil 1* [WADI - Keeping awake and diagnose basic knowledge and skills in the math domain Grade 9/10 Pt. 1]. Landesinstitut für Schulentwicklung.
- Kunter, M., Frenzel, A., Nagy, G., Baumert, J., & Pekrun, R. (2011). Teacher enthusiasm: Dimensionality and context specificity. *Contemporary Educational Psychology*, 36(4), 289–301. <https://doi.org/10.1016/j.cedpsych.2011.07.001>
- 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>
- Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente* [PISA 2000: Documentation of instruments]. Max-Planck-Institut für Bildungsforschung.
- Lauermann, F., Chow, A., & Eccles, J. S. (2015). Differential effects of adolescents' expectancy and value beliefs about math and English on math/science-related and human-services-related career plans. *International Journal of Gender, Science and Technology*, 7(2), 205–228.
- Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research*, 86(2), 602–640. <https://doi.org/10.3102/0034654315617832>
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, 19(6), 527–537. <https://doi.org/10.1016/j.learninstruc.2008.11.001>
- Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology*. British Psychological Society.
- Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, 44(6), 764–802. <https://doi.org/10.1080/00273170903333665>
- Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology*, 111(2), 331–353. <https://doi.org/10.1037/edu0000281>

- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2), 397–416. <https://doi.org/10.1111/j.1467-8624.2005.00853.x>
- McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, 22(1), 114–140. <https://doi.org/10.1037/met0000078>
- Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research*, 79(3), 1129–1167. <https://doi.org/10.3102/0034654309337522>
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.).
- Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T. T., & Trautwein, U. (2011). Who took the “x” out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science*, 22(8), 1058–1066. <https://doi.org/10.1177/0956797611415540>
- Noonan, R. (2017). STEM jobs: 2017 Update (ESA Issue Brief # 02-17). <http://www.esa.gov/reports/stem-jobs-2017-update>
- O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, 78(1), 33–84. <https://doi.org/10.3102/0034654307313793>
- Oreopoulos, P., & Petronijevic, U. (2019). *The remarkable unresponsiveness of college students to nudging and what we can learn from it*. EdWorkingPaper: 19–102. <http://www.edworkingpapers.com/ai19-102>
- Parrisius, C., Gaspard, H., Trautwein, U., & Nagengast, B. (2020). The transmission of values from math teachers to their ninth-grade students: Different mechanisms for different value dimensions? *Contemporary Educational Psychology*, 62, Article 101891. <https://doi.org/10.1016/j.cedpsych.2020.101891>
- Parrisius, C., Gaspard, H., Zitzmann, S., Trautwein, U., & Nagengast, B. (2020). *The “situative nature” of competence and value beliefs and the predictive power of autonomy support: A multilevel investigation of repeated observations* [Manuscript submitted for publication]. Hector Research Institute of Education Sciences and Psychology, University of Tübingen.
- Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology*, 106(1), 315–329. <https://doi.org/10.1037/a0034027>
- Perez, T., Dai, T., Kaplan, A., Cromley, J. G., Brooks, W. D., White, A. C., Mara, K. R., & Balsai, M. J. (2019). Interrelations among expectancies, task values, and perceived costs in undergraduate biology achievement. *Learning and Individual Differences*, 72, 26–38. <https://doi.org/10.1016/j.lindif.2019.04.001>
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119. <https://doi.org/10.3102/0013189X09332374>
- Piesch, H., Gaspard, H., Parrisius, C., Wille, E., & Nagengast, B. (2020). How can a relevance intervention in math support students' career choices? *Journal of Applied Developmental Psychology*, 71, Article 101185. <https://doi.org/10.1016/j.appdev.2020.101185>
- Piesch, H., Häfner, I., Gaspard, H., Flunger, B., Nagengast, B., & Harackiewicz, J. M. (2019). Helping parents support adolescents' career orientation: Effects of a parent-based utility-value intervention. *Unterrichtswissenschaft*, 47(3), 271–293. <https://doi.org/10.1007/s42010-018-0024-x>
- Pituch, K. A., & Stapleton, L. M. (2012). Distinguishing between cross- and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research*, 41(4), 630–670. <https://doi.org/10.1177/0049124112460380>
- Plewis, I., & Hurry, J. (1998). A multilevel perspective on the design and analysis of intervention studies. *Educational Research and Evaluation*, 4(1), 13–26. <https://doi.org/10.1076/edre.4.1.13.13014>
- Praetorius, A. K., McIntyre, N. A., & Klassen, R. M. (2017). Reactivity effects in video-based classroom research: An investigation using teacher and student questionnaires as well as teacher eye-tracking. *Zeitschrift für Erziehungswissenschaft*, 20(S1), 49–74. <https://doi.org/10.1007/s11618-017-0729-3>
- Priniski, S. J., Hecht, C. A., & Harackiewicz, J. M. (2018). Making learning personally meaningful: A new framework for relevance research. *Journal of Experimental Education*, 86(1), 11–29. <https://doi.org/10.1080/00220973.2017.1380589>
- Rakoczy, K., Buff, A., & Lipowsky, F. (2005). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie “Unterrichtsqualität, Lernverhalten und mathematisches Verständnis”. 1. Befragungsinstrumente* [Technical report of the Swiss-German video study “Instructional quality, learning behavior, and mathematical comprehension]. Gesellschaft zur Förderung Pädagogischer Forschung/Deutsches Institut für Internationale Pädagogische Forschung. <https://URN:urn:nbn:de:0111-opus-31060>
- Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (2006). *PISA 2003. Dokumentation der Erhebungsinstrumente* [PISA 2003. Documentation of assessment instruments]. Waxmann.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173–185. <https://doi.org/10.1037/1082-989X.2.2.173>
- Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., & Bloom, H. (2011). *Optimal design software for multi-level and longitudinal research (Version 3.01)* [Software]. www.wtgrantfoundation.org
- Rosenzweig, E. Q., & Wigfield, A. (2016). STEM motivation interventions for adolescents: A promising start, but further to go. *Educational Psychologist*, 51(2), 146–163. <https://doi.org/10.1080/00461520.2016.1154792>
- Rosenzweig, E. Q., Hulleman, C. S., Barron, K. E., Kosovich, J. J., Priniski, S. J., & Wigfield, A. (2019). Promises and pitfalls of adapting utility value interventions for online math courses. *Journal of Experimental Education*, 87(2), 332–352. <https://doi.org/10.1080/00220973.2018.1496059>
- Rosenzweig, E. Q., Wigfield, A., & Hulleman, C. S. (2020). More useful or not so bad? Examining the effects of utility value and cost reduction interventions in college physics. *Journal of Educational Psychology*, 112(1), 166–182. <https://doi.org/10.1037/edu0000370>
- Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61, Article 101860. <https://doi.org/10.1016/j.cedpsych.2020.101860>
- Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., & Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods*, 19(3), 317–333. <https://doi.org/10.1037/met0000013>
- Schmidt, S., Ennemoser, M., & Krajewski, K. (2013). *Deutscher Mathematiktest für 9. Klassen* [German mathematics test for Grade 9]. Hogrefe.
- Shin, D. D., Lee, M., Ha, J. E., Park, J. H., Ahn, H. S., Son, E., Chung, Y., & Bong, M. (2019). Science for all: Boosting the science motivation of elementary school students with utility value intervention. *Learning and Instruction*, 60, 104–116. <https://doi.org/10.1016/j.learninstruc.2018.12.003>
- Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology*, 42(1), 70–83. <https://doi.org/10.1037/0012-1649.42.1.70>
- Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-

- sets important to academic achievement? Two meta-analyses. *Psychological Science*, 29(4), 549–571. <https://doi.org/10.1177/0956797617739704>
- Song, J., Gaspard, H., Nagengast, B., & Trautwein, U. (2020). The Conscientiousness \times Interest Compensation (CONIC) model: Generalizability across domains, outcomes, and predictors. *Journal of Educational Psychology*, 112(2), 271–287. <https://doi.org/10.1037/edu0000379>
- Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S. (Eds.). (2019). *IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich* [IQB Educational Trend 2018: Mathematical and scientific competencies at the end of lower secondary school in the second state comparison]. Waxmann.
- Statistisches Bundesamt [Destatis]. (2018). *Bildungsstand der Bevölkerung: Ergebnisse des Mikrozensus 2017* [Educational level of the population: Results of the microcensus 2017].
- Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, 98(4), 788–806. <https://doi.org/10.1037/0022-0663.98.4.788>
- Trautwein, U., Lüdtke, O., Roberts, B. W., Schnyder, I., & Niggli, A. (2009). Different forces, same consequence: Conscientiousness and competence beliefs are independent predictors of academic effort and achievement. *Journal of Personality and Social Psychology*, 97(6), 1115–1128. <https://doi.org/10.1037/a0017048>
- Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy-value theory: A latent interaction modeling study. *Journal of Educational Psychology*, 104(3), 763–777. <https://doi.org/10.1037/a0027470>
- Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). National Foundation for Educational Research.
- Van Daele, T., Van Audenhove, C., Hermans, D., Van Den Bergh, O., & Van Den Broucke, S. (2014). Empowerment implementation: Enhancing fidelity and adaptation in a psycho-educational intervention. *Health Promotion International*, 29(2), 212–222. <https://doi.org/10.1093/heapro/das070>
- Vansteenkiste, M., Aelterman, N., De Mynck, G.-J., Haerens, L., Patall, E., & Reeve, J. (2018). Fostering personal meaning and self-relevance: A self-determination theory perspective on internalization. *Journal of Experimental Education*, 86(1), 30–49. <https://doi.org/10.1080/00220973.2017.1381067>
- Watt, H. M. G. (2004). Development of adolescents' self-perceptions, values, and task perceptions according to gender and domain in 7th-through 11th-grade Australian students. *Child Development*, 75(5), 1556–1574. <https://doi.org/10.1111/j.1467-8624.2004.00757.x>
- Watt, H. M. G., Shapka, J. D., Morris, Z. A., Durik, A. M., Keating, D. P., & Eccles, J. S. (2012). Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: A comparison of samples from Australia, Canada, and the United States. *Developmental Psychology*, 48(6), 1594–1611. <https://doi.org/10.1037/a0027838>
- Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, 33(3), 778–808. <https://doi.org/10.1002/pam.21760>
- What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook*. U.S. Department of Education, Institute of Education Sciences, National Center for Education and Regional Assistance. <https://doi.org/10.1037/e578392011-004>
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7, Article 1832. <https://doi.org/10.3389/fpsyg.2016.01832>
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68–81. <https://doi.org/10.1006/ceps.1999.1015>
- Wigfield, A., Rosenzweig, E. Q., & Eccles, J. S. (2017). Achievement values: Interactions, interventions, and future directions. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (2nd ed., pp. 116–134). Guilford Press.
- Wigfield, A., Tonks, S., & Klauda, S. T. (2016). Expectancy-value theory. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of motivation at school* (2nd ed., pp. 55–74). Routledge.
- Woolley, M. E., Rose, R. A., Orthner, D. K., Akos, P. T., & Jones-Sanpei, H. (2013). Advancing academic achievement through career relevance in the middle grades: A longitudinal evaluation of Career-Start. *American Educational Research Journal*, 50(6), 1309–1335. <https://doi.org/10.3102/0002831213488818>
- Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, 81(2), 267–301. <https://doi.org/10.3102/0034654311405999>
- Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774), 364–369. <https://doi.org/10.1038/s41586-019-1466-y>
- Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., Lee, H. Y., O'Brien, J., Flint, K., Roberts, A., Trott, J., Greene, D., Walton, G. M., & Dweck, C. S. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology*, 108(3), 374–391. <https://doi.org/10.1037/edu0000098>

Received March 6, 2020

Revision received November 20, 2020

Accepted December 20, 2020 ■