**Individualized Instruction – Conceptual and Empirical Examinations of Necessary Conditions for its Effectiveness**

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich 05

Psychologie und Sportwissenschaften

der Johann Wolfgang Goethe-Universität

in Frankfurt am Main

von

Leonard Tetzlaff

aus Frankfurt am Main

Frankfurt, 2022

vom Fachbereich 05 der

Johann Wolfgang Goethe - Universität als Dissertation angenommen.

Dekanin: Prof. Dr. Sonja Rohrmann

Gutachter: Prof. Dr. Garvin Brod

Prof. Dr. Gerhard Büttner

# Individualized Instruction – Conceptual and Empirical Examinations of Necessary Conditions for its Effectiveness

**Summary**

Individualization can be defined as the adaptation of instructional parameters to relevant characteristics of a specific learner. This definition raises several questions, however: Which characteristics are actually relevant? Which parameters of instruction need to be adjusted, and in which way, to positively interact with those characteristics? In a classroom context, additional questions arise: how can information about the relevant learner characteristics be delivered to the teacher? How can individualized instruction be delivered to each learner in a context that has originally been designed for whole-class instruction? By focusing on the measurement and modelling of learner characteristics and instructional adaptations, this dissertation aims to provide an insight into each of these issues.

This dissertation is divided into two parts. The first part is concerned with the theoretical (*Paper 1*) and statistical (*Paper 2*) modeling of learner characteristics in the context of individualized instruction. The second part is concerned with the measurement (*Paper 3*) and implementation (*Paper 4*) of individualized instruction in the classroom context.

*Paper 1* summarizes existing research on individualization from different research traditions. From this summary I derive the need for a dynamic conceptualization of learner characteristics (acknowledging that learners change during and in interaction with the learning process) and synthesize a dynamic framework that details the opportunities for individualization on three different timescales. *Paper 2* reports results from an exploratory study that investigated the potential benefits of utilizing person-centered analysis for the assessment of multivariate learner prerequisites and their interaction with instruction. We found that latent profiles over several reading related abilities could explain differential effectiveness of self-reported teaching foci in German third grade reading lessons. These findings indicate not just a need for stronger individualization of teaching but also an advantage of multivariate

conceptualizations of learner characteristics. Additionally, they show the utility of person-centered approaches for the investigation of such multivariate learner characteristics and their interaction with instruction.

In the second part, I investigate possible approaches to the implementation and measurement of individualization in a classroom context. *Paper 3* investigates whether teacher-, student- and observer perspectives converge when rating the amount of individualization present in regular classroom instruction. We found considerable agreement between the perspectives, indicating a common understanding of the construct at the classroom level as well as providing some evidence for the validity of the used measurement instruments. *Paper 4* replicates findings concerning the effectiveness of formative assessment procedures for fostering reading education, supplemented by a moderator analysis showing that only children with low performance at the beginning of the school-year profited from its implementation. This indicates that the information provided by formative assessment procedures helps teachers to identify struggling readers but does not seem to be utilized for adapting instruction to specific deficits of average or high performing children.

In sum, this dissertation contributes to research on individualized instruction by demonstrating necessary conditions for its effectiveness. It posits the need for a dynamic conceptualization of learner characteristics, demonstrates the advantage of multivariate learner profiles, and points out ways towards the successful implementation of individualized instruction in the classroom.

**Table of Contents**

# 1 Introduction

Individualizing instruction means adapting instruction to specific learners in their unique constellations of skills, experiences and prior knowledge structures. As such, it has been an important goal of educational research and practice for ages with first mentions of such a concept dating back to ca. 500 BC (Quintilian, trans. 1921). Especially in times of growing heterogeneity in classrooms (Corno, 2008; Decristan et al., 2017; Subban, 2006) and an increasing prevalence of technology-based learning (Bernacki et al., 2021), adapting instruction to individual learners' strengths and needs becomes increasingly important as well as increasingly feasible. The general effectiveness of individualized instruction has been repeatedly demonstrated at a class- or school-level in classroom instruction (Connor et al., 2007, 2018; Jung et al., 2018; Kingston & Nash, 2011; Pane et al., 2015; Stecker et al., 2005; Waxman et al., 1985) as well as in digital learning environments (Corbett, 2001; Ma et al., 2014; Steenbergen-Hu & Cooper, 2014). However, this prior research has mostly focused on the classroom level instead of the level of the individual learner. In addition, substantial heterogeneity of effect sizes indicates the existence of several factors that moderate the effectiveness of individualized instruction.

A first challenge that arises when trying to design individualized instruction is that successful individualization depends on correctly identifying and modelling the relevant characteristics of a specific learner. This identification poses several methodological challenges, especially if we conceptualize learner characteristics as dynamic and multivariate. Furthermore, in order to realize effective individualized instruction, teaching agents are in need of information about instructional adaptations that have been shown to increase learning success for learners with specific characteristics. The research tradition of "aptitude-treatment interactions" still lacks a solid foundation of reproducible results that would represent a reliable basis for such adaptations.

Finally, in the classroom context teachers need to find ways to deliver individualized instruction to specific learners without neglecting the rest of the class. This is a challenging task, given that teachers commonly face large classrooms with sometimes great student heterogeneity. Several studies could show that programs that support individualized instruction lead to increased learning gains at the classroom level (e.g. Connor et al., 2009; Waxman et al., 1985). However, more research is needed concerning the actual implementation in regular classroom practice.

In this dissertation, I gather evidence from previous research as well as my own studies in order to better understand how relevant learner characteristics and their interaction with instructional parameters can be modelled and how individualized instructional adaptations can be implemented in regular classroom instruction. I will start with an overview of the theoretical and empirical background this dissertation is based on. I will then briefly summarize the four manuscripts that together form this dissertation. I conclude by summarizing the most important findings, mentioning limitations as well as potential directions for future research.

## 2 Theoretical Background

In the following, I first summarize previous research on individualization, highlighting the commonalities between different approaches and compiling evidence for its effectiveness. I then introduce aptitude-treatment interactions (ATIs) as the most plausible explanation for the efficacy of individualization and map out several issues that this line of research is struggling with. In the last segment, I turn towards the implementation of individualized instruction in a classroom context and the additional difficulties this poses for practitioners and researchers.

## 2.1 Individualized Instruction

Human learners differ in a myriad of ways, including in their prior knowledge, intelligence, hair color, socioeconomic status, current affect, favorite music genre, current motivation, working memory capacity, and many more. Individualized or personalized learning approaches want to take these differences into account to optimize the fit of instruction to a specific learner. While earlier conceptualizations of individualized instruction certainly exist, the first psychological perspective on the issue can be traced back to Lev Vygotsky and his concept of the zone of proximal development (Vygotsky, 1930-1944/1978). This zone consists of all tasks or challenges that a learner is unable to accomplish without support but is able to accomplish *with* support. According to Vygotsky, optimal instruction should always be situated within this zone. The location, size and malleability of this zone for a given learning objective is defined by the personal characteristics of a specific learner. Not just prior knowledge but also cognitive characteristics such as intelligence or motivational factors such as interest can influence whether a certain challenge can be met with instructional support. Clearly, the zones of proximal development can be very different for different learners. Thus, instruction that addresses entire groups of learners at once runs the risk of being out of zone for at least some of the learners.

Besides these mainly theoretical considerations, first empirical evidence for the efficacy of individualized instruction can be traced back to Bloom's work on individual tutoring. Bloom and students found learning gains under individual one-on-one tutoring to be up to two standard deviations higher than in "regular instruction" (Bloom, 1984). The quest to scale up these effects to larger groups of learners became known as the "2-sigma-problem" (Barrows et al., 1986; Corbett, 2001). While future studies failed to replicate effect sizes of that magnitude, the general effect of one-on-one tutoring being the most effective form of learning remains (Vanlehn, 2011). The main difference between one-on-one tutoring and regular

instruction is that the one-on-one tutor has much more *opportunities* and much more *information* to adapt the instructional approach to individual learners (Bloom, 1984). This is true for the selection of goals and subgoals, the design of instructional units as well as assistance during the learning process (Lehman et al., 2008). Individualization of instruction can thus be conceptualized as scaling up the positive effects of one-on-one tutoring to larger groups without having to provide an individual tutor for each learner.

This endeavor can be seen as principally successful – across several different domains, different approaches to individualize instruction have been shown to have positive effects on learning gains. Intelligent tutoring systems, for example, are defined by their assessment of several specific learner characteristics and the subsequent adaptation of instruction (Nwana, 1990). Their effectiveness when compared to regular computer-assisted instruction (Ma et al., 2014; Steenbergen-Hu & Cooper, 2014) can thus be conceptualized as positive effects of individualization. But also in regular classroom instruction, formative assessment (e.g. Deno, 1990) and internal differentiation (e.g. Slavin, 1987) can both - among others - be conceptualized as individualization approaches (Dumont, 2019) and have been shown to have positive effects on learning gains (Allington, 1974; Jung et al., 2018; Slavin, 1987; Slavin & Karweit, 1985)

The common thread that weaves through all of these approaches - independent of context or domain - is that some form of assessment of learner characteristics is used to inform and adapt subsequent instruction. Individualization can thus be defined as the systematic adaptation of instructional parameters to the relevant characteristics of a specific learner (see *Paper 1*). In the literature, several different terms are used to describe this process. In the context of this dissertation, “individualized” and “personalized” instruction are understood as

synonyms and used in the broadest possible sense: any adaptation of instruction based on (perceived) learner characteristics to individual learners or groups of learners.

### 2.2 Aptitude-Treatment Interactions

In order for instructional adaptations to convey educational benefits, different instructional parameters need to interact with specific constellations of learner parameters. Without such interactions, some learners would learn better than others and some instructional approaches would be more effective than others. However, if there is no interaction between an adaptation and learners' characteristics, then there would be no benefit in adapting the instructional approach for specific learners. Interactions between instructional adaptations and learner characteristics have been conceptualized and examined under the term of aptitude-treatment interactions (ATIs).

ATIs have first been postulated by Cronbach (1957) who described them as a synthesis of the traditions of correlational (or interindividual) psychology and the tradition of experimental (or intraindividual) psychology. ATI research combines these two traditions by studying differential effects of experimental treatments for people at different points on spectra of interindividual differences. As mentioned above, these differential effects are necessary for individualization to have any benefit. Isolating and identifying specific interactions between learner characteristics (aptitudes) and instructional parameters (treatments), allows for optimal treatments to be selected for each learner. Aptitudes in this case are defined broadly as *any* learner characteristics that have or are presumed to have an effect on the response to a specific treatment (Cronbach, 1975).

In the following years, this paradigm has seen a surge of attention (e.g. Bracht, 1970; Snow, 1980, 1989; Tobias, 1978; Tobias & Redfield, 1980). Still, surprisingly few ATIs could be reliably demonstrated and replicated (Cronbach & Snow, 1977). This is puzzling, as they

*need* to exist in order to explain the positive effects of individualization. In addition, the face validity of the concept is also exceptionally high - whenever there is substantive heterogeneity in treatment effects, it is reasonable to assume that some characteristics of the learners are causing this heterogeneity. Several authors came to the same conclusion: ATIs *have* to exist – so people either have been looking in the wrong places (Driscoll, 1987; Tobias, 1989) or utilizing the wrong methods (Preacher & Sterba, 2019; Shapiro, 1975).

In the 40 years that have passed since then, several ATIs have been found (e.g. .Seufert et al., 2009; Suzuki & Dekeyser, 2017; Ziegler et al., 2021), most prominently the expertise reversal effect (e.g. Chen et al., 2017; Lee & Kalyuga, 2014; Tobias, 2010). But the general sentiment that the field is severely trailing behind the expectations of researchers and practitioners alike is still prevalent (Preacher & Sterba, 2019). A possible explanation for this lack of findings that is of particular relevance for this dissertation is the typical conceptualization and operationalization of an aptitude. Typically, aptitudes are conceptualized and operationalized as static, univariate constructs. This results in the operationalization as a single pretest measure, putting the focus on specific isolated variables to examine interactions with treatments.

### *2.2.1 Dynamic Aptitudes*

Learners and their aptitudes change during and in interaction with the instructional process. A completely static aptitude concept will fail to capture these changes, leading to wrong estimations of the interaction effects as the actual current aptitude of the learner might be different from the value that was measured at the beginning of the process. From this follows that a treatment that may provide an optimal fit at the beginning of the learning process is potentially mismatched for *some* of the learners at later stages of the same process (for example

depending on their rate of expertise development). Rey and Fischer ( 2013), for instance, demonstrated that even small gains in expertise (such as reading a text on the subject matter) can significantly alter the effectiveness of subsequent treatments.

Some aptitude concepts are static by definition, such as the concept of learning styles - the idea that learners fall into one of several distinct and stable categories that moderate the effectiveness of learning based on the mode of presentation or organization of the content to be learned. While this concept exhibits high face validity and quickly found widespread dissemination into practice (Wininger et al., 2019), current evidence mainly points against its effectiveness (Kirschner, 2017; Pashler et al., 2008). Learners may voice preferences concerning the mode of presentation or organization of learning materials, but their learning gains do not increase when their preferences are being met. Learning styles can thus be seen as an example of a larger group of individualization attempts that operate by sorting learners into distinct categories that are assumed to be stable over the course of the learning process. This group of attempts appears to be much less effective than adapting to dynamic characteristics such as prior knowledge (e.g. Rey & Fischer, 2013) or interest (e.g. Walkington, 2013).

Acknowledging learners' propensity to change and embracing it by assessing relevant characteristics at a high frequency could lead to more robust estimates of differential effects of treatments while also being able to probe the temporal dynamics of interventions (e.g. Breitwieser et al., 2021). This allows instruction to not just be adjusted to specific learners but to specific learners at specific points in time.

### *2.2.2 Multivariate Aptitudes*

Learners also don't just differ in single variables but they all bring their unique constellation of aptitudes into the learning process. It is possible that the value in one aptitude moderates the interaction of another with the treatment, such as a high amount of anxiety

preventing learners to utilize their prior knowledge. This concept has first been postulated by Snow (1987) under the name of aptitude-complexes: constellations of aptitudes that together influence treatment effects above and beyond the influence of each variable alone. This concept was further extended by Ackerman (2003), who noticed that specific cognitive, motivational, and even attitude aptitudes co-occur more often than others and called these constellations trait complexes. These possess a much higher ecological validity than the artificial complexes utilized by Snow and colleagues, but also potentially lessen the likelihood to find interactions with treatments due to their less extreme nature.

This phenomenon of multiple relevant learner characteristics interacting with each other as well as with the provided treatment has been likened to a "hall of mirrors that extends to infinity" by Cronbach (1975). This alludes to the impossibility of capturing every higher-order interaction. Inconsistent findings when analyzing specific interactions have also been explained as being due to unmodelled additional interactions. If the "true" ATIs are classified by interactions of several distinct variables, it is to be expected that inconsistent results are obtained when looking at only one of them (Cronbach, 1975). While this shows that multivariate aptitudes have been conceptualized several decades ago, most research in the field still concerned itself with the interaction of single aptitudes and treatments. This discrepancy can be partially explained by the methodological difficulties that come along with analyzing higher order interactions (Bauer & Shanahan, 2007).

Indeed, variable-centered approaches – a group of statistical methods that are focusing on the association between variables – inevitably run into overwhelming amounts of interpretational complexity when incorporating multiple higher-order interactions (Bauer & Shanahan, 2007; Cronbach, 1975). Besides this interpretational complexity, classic variable-centered approaches such as multiple regression also quickly become underpowered for

detecting higher-order interactions (Cronbach, 1975; McClelland & Judd, 1993) while also failing to capture nonlinear relations (Bauer & Cai, 2009).

### *2.2.4 Interim Conclusion: ATIs*

In sum, ATIs are still the only plausible explanation for the positive effects of individualization. While a few generalizable interactions have been established, the general conclusion Cronbach and Snow (1977) reached in their review still mostly holds: "No Aptitude X Treatment interactions are so well confirmed that they can be used directly as guides to instruction".

Multivariate learner characteristics have been increasingly analysed to explain variance in learning (e.g. Lonigan et al., 2018; Reinhold et al., 2020) but only very few studies actually looked at multivariate learner characteristics in interaction with instruction (e.g. Hofer et al., 2018; Hooper et al., 2006; Suzuki & Dekeyser, 2017). Similar things can be said about the dynamic modelling of aptitudes. As learners and their aptitudes change during and in interaction with the learning progress, so should the differential effectiveness of specific instructional approaches. While several studies looked at dynamic learner characteristics to explain learning success (e.g. Förster et al., 2022; Reinhold et al., 2020), and adaptations based on dynamic assessment have been shown to be successful (Jung et al., 2018), I am not aware of studies that explicitly modelled the interaction of dynamic learner characteristics with instruction in detail and with respect to the temporal dynamics.

## 2.3 Individualized Instruction in the Classroom

Concerning the implementation of individualized instruction in actual classroom practice, two separate things need to be accomplished. The first important consideration is that

structures need to be in place that allow individual students (or relatively homogenous groups of students) to receive instruction separate from the rest of the class.

In contrast to one-on-one tutoring where it is clear that attention is completely on one learner or intelligent tutoring systems that are in theory infinitely scalable, regular classroom instruction usually has only one teacher for 15-30 learners. Several different "systems" have been utilized to still enable teachers to target instruction at individual learners. Examples of those include: efficient grouping to target instruction at homogenous subgroups instead of individual students (internal differentiation; Allington, 1974; Constas & Sternberg, 2013; Slavin, 1987), designing instruction in a way that allows different students to engage with the same task at their own level (adaptive tasks; Bardy et al., 2021; Corno, 2008), or even empowering students to select and adapt their own instruction (self-regulated learning; Boekaerts & Corno, 2005; Paris & Paris, 2010).

The second important consideration is that the actual instructional input received by the learners needs to not only be different from that of the rest of the class but positively interact with some characteristics of these individual learners. To achieve this, teachers need information about learners' characteristics. While most of this information comes from the daily interaction of teachers with their students, research has shown that teachers are not always able to correctly assess relevant characteristics of their students: Whereas teachers are quite proficient at judging performance in their specific subject matter, the accuracy of their assessment decreases for other characteristics (Machts et al., 2016). This finding implies that the information gained from daily interactions should best be supported by some form of explicit diagnostic information. An especially promising approach of providing teachers with relevant information about individual learner characteristics is formative assessment (Deno, 1990).

### *2.3.1 Formative Assessment*

Formative assessment (also known as curriculum-based measurement or learning progress assessment) can be seen as an extension of the mastery learning concept (Bloom, 1968) with the goal to enable usage in more traditional grouped instructional settings. In mastery learning, a teaching agent sets several successive intermediate goals on which all students get regularly tested. When sufficient mastery of one intermediate goal is displayed, they progress to the next (Bloom, 1968). The main difference between the concepts is that in formative assessment, all students get tested on the same overarching learning goal instead of on their current intermediate goal (Fuchs, 2004). This allows the teacher to continuously monitor progress on a single scale and to adapt the instruction in case of stagnation. These adaptations don't have to be individualized - teachers can also use formative assessment data to identify specific trends in their class as a whole and adapt their whole-class instruction accordingly.

Meta-analyses have generally shown that formative assessment has positive effects on learning gains of students (Förster et al., 2018; Jung et al., 2018; Kingston & Nash, 2011; Lee et al., 2020; Stecker et al., 2005), with some indications that effects are larger for struggling readers (e.g. Jung et al., 2018). Besides the positive effects, a striking feature of the studies reported in the above-mentioned meta-analyses is the substantive heterogeneity of effect sizes. This implies that the positive effects of formative assessment are heavily dependent on moderating factors such as teacher experience or context (Kingston & Nash, 2011). A more detailed investigation into those mediating and moderating mechanisms is of utmost importance to utilize the full potential of formative assessment.

While formative assessment itself is only concerned with providing information, the generally assumed mechanism by which it conveys its positive effects on learning is that

teachers use that information to better adapt their instruction to individual learners (Brink et al., 2019; Cusi & Telloni, 2019; Jung et al., 2018; Kaftan et al., 2006; Yeh, 2010). While this assumption is intuitively plausible, I am not aware of any research that actually quantitatively investigated the relationship between formative assessment and individualized instruction. This is possibly due to difficulties associated with assessing individualized instruction in classroom education, which I will elaborate on in the next section.

### *2.3.2 Assessment of Individualized Instruction in the Classroom*

As classroom processes are constituted by an interplay of several actors in a closed system, it can be difficult to objectively and reliably assess them. Possible options for doing so include teacher reports, student reports and external observers, each associated with their own advantages and difficulties (Fauth et al., 2014; Kunter & Baumert, 2006).

When assessing individualized instruction in particular, these difficulties are complemented by additional complications: Due to the concept of individualized instruction being highly desirable, it is to be expected that teacher self-reports are biased towards reporting higher levels/more occurrence (Kopcha & Sullivan, 2007). This has been shown by Fraser (1982), who investigated teacher and student self-reports on individualized classroom environments with parallel scales. They found partial correspondence between teacher- and student self-reports but the teachers consistently rated the classroom environment as more individualized than their respective students.

Student ratings of individualized instruction are also potentially less reliable than those of other classroom processes (Lüdtke et al., 2006). The amount of individualized instruction individual students receive might vary based on the perceived need and students might not be able to correctly identify whether their classmates receive individualized instruction. This

implies that using common statistical models involving latent variables might not be appropriate as students might only rate the amount of individualized instruction they themselves received, not the amount generally present in the classroom, which would violate the assumptions behind such models (Rhemtulla et al., 2020). I am not aware of any studies that utilize external observers to quantitatively measure individualized instruction. An accurate assessment of within-classroom individualization at best reflects the scientific operationalization of the construct as well as the actual classroom processes as perceived by students and teachers. This a necessary prerequisite not just for testing claims of individualization as a mediator of the positive effects of formative assessment but also for investigating its prevalence and outcomes in regular classroom instruction, independent of supporting systems.

### *2.3.3 Interim Conclusion: Individualization in the Classroom*

In sum, while several distinct approaches exist to individualize instruction in classrooms, most of them are mainly concerned with providing opportunities for adaptations. The notable exception to this is formative assessment which is mainly concerned with providing relevant information to teachers in order to enable individualized instruction.

Most of these class-level approaches have been shown to increase learning gains when compared to regular classroom instruction, but the effect sizes show considerable heterogeneity – indicating a need for further research into their moderating or mediating factors. An increase in the amount of individually targeted instruction is usually assumed to be a mediator of these effects (Brink et al., 2019; Jung et al., 2018) but seldomly investigated. This is potentially due to the difficulties associated with assessing actual individualization in the classroom that have been described above.

## 3 Aims of This Dissertation

This dissertation can be broadly split in two different parts. The first part is concerned with theoretical and methodological considerations concerning individualized instruction. The second part is concerned with the implementation of individualized instruction in the classroom context. In the following, I will detail the main aims of each part.

*Theoretical Considerations*

1: My first aim is to broadly summarize and integrate the current state of research on individualization across different contexts and disciplines. The focus is on the dynamics of changing learner characteristics and instructional adaptations, since, as I will argue, a dynamic perspective on individualization is a necessary next step to move the field forward. Based on that summary, I propose a general framework of individualization that takes dynamics into account. I end by providing recommendations for future research on individualization.

2: The existence of differential effectiveness of treatments (or ATIs) is a necessary requirement for effective individualization, but the current evidence concerning such interactions is sparse. I posit that by looking at multiple relevant learner characteristics simultaneously, differential effectiveness of treatments, that would go unnoticed by just looking at univariate characteristics, can be identified.

2a: As the interaction of multiple learner characteristics with instruction can be hard to analyze with variable-centered approaches, I probe person-centered analyses as an alternative tool for identifying these multivariate learning prerequisites and their interaction with instruction.

*Implementing Individualized Instruction in the Classroom*

3: A prerequisite for research on individualized instruction in the classroom are reliable instruments to assess the amount of individualization present in any given classroom. Individualization looks quite different depending on the perspective - external observers utilize the scientific operationalization of the construct, teachers try to plan and implement it in their instruction, while students need to receive and utilize the individualized offer.

I investigate whether individualized instruction in regular classroom practice can be reliably assessed from teacher, student, and observer perspectives and whether these perspectives correlate with each other, indicating a shared understanding of the construct at the classroom level.

4: One of the most promising approaches of bringing individualized instruction to the classroom is formative assessment. Even though the general effectiveness of this approach has been repeatedly demonstrated, the actual mechanisms by which it operates and the factors that moderate its effectiveness are still underexplored. I would like to first replicate earlier findings that classes using a formative assessment program show, on average, greater learning gains than classes who don't. In addition to this replication, I also look at the initial performance level of the students, investigating whether certain subgroups of learners especially benefit from the program.

# 4 Summary of Papers

## 4.1 The iLearn Project

Three of the four papers in this thesis use data from the iLearn project. The main aim of the iLearn project was to investigate effects of formative assessment in the context of third grade German lessons and it was funded by the German ministry of education and research. The project was run in 2 cohorts, one in the year 2018/19, the other in the year 2019/20.

consisted of a pretest at the beginning of the schoolyear and a posttest after the summer holidays. At both of these occasions, all students took part in a pen and paper test battery, comprising measures of reading comprehension, spelling, and general intelligence administered to the whole class at once. A teacher-nominated subgroup of students additionally participated in a computer-based test battery comprising decoding, vocabulary, syntax comprehension and working memory capacity in individual sessions. To assess the teaching practice over the schoolyear, teachers filled out a short online questionnaire every three weeks. A subset of teachers also participated in classroom observations, which took place in the middle of the schoolyear.

We were able to recruit 77 teachers, 41 of which also used the formative assessment tool "quop" (Souvignier et al., 2021). The respective student sample comprised 668 students. Usage of "quop" entails all children taking a short online test every three weeks, the results of which are provided to teachers in a graphical form. Dependent on the specific research questions, we had to exclude some participants with missing data on relevant variables, leading to slightly differing sample sizes in the different analyses. For detailed information of the procedures and measures used, see the method sections of the respective papers (Appendix A).

**4.2 Paper 1:**

**Tetzlaff, L.,** Schmiedek, F., & Brod, G. (2021). Developing Personalized Education: A Dynamic Framework. *Educational Psychology Review*, *33*(3), 863-882.

In this review paper, we summarize and synthesize research on individualized instruction across three different research traditions – ATIs, classroom education, and digital

learning environments. We then explain how learner characteristics can vary across different timescales and differentiate between different "kinds of variance" – developmental processes, intervention-induced changes and short-term fluctuations. Concluding the paper, we merge these considerations into a dynamic framework of individualization across three timescales.

**Summary of existing research:** Looking at previous research conducted on individualized instruction, we found a clear advantage of dynamic, as opposed to static, approaches to modelling learner characteristics in order to inform instructional adaptations. This is exemplified by the dearth of reliable ATIs using static aptitude measures (Cronbach & Snow, 1977) (Bracht, 1970; Cronbach & Snow, 1977; Driscoll, 1987; Shapiro, 1975) and the prevalence of expertise reversal effects, which are by definition dynamic (Kalyuga, 2007; Khacharem et al., 2015; Rey & Fischer, 2013). The success of formative assessment in classroom contexts (Black & Wiliam, 1998; Jung et al., 2018; Kingston & Nash, 2011) and intelligent tutoring systems in digital learning environments (Corbett, 2001; Ma et al., 2014; Steenbergen-Hu & Cooper, 2014) can also be seen as a success of dynamic learner modelling as both approaches utilize it.

**Developing a dynamic framework:** Learners change on different timescales: from developmental processes that unfold over months or even years over intervention-induced changes that take place over weeks to short-term fluctuations that can occur over days or even moment-to-moment (Hertzog & Nesselroade, 2003). These different types of variance require different measurement approaches and provide opportunities for different instructional adaptations. If instruction is to be optimally adapted to specific learners, it needs to take these dynamics on different timescales into account. Based on these considerations, we construct a dynamic framework of personalized education, detailing the relevant assessment structures as well as potential for instructional adaptations on three separate timescales. These range from

the setting of appropriate learning goals at the macroscale over instructional design at the mesoscale to reacting to affective-motivational fluctuations on the microscale.

**Recommendations for future work:** Especially in the classroom context, there is a need for studies that illuminate in detail a) the amount of individualized practise and b) the actual fit between instructional adaptations and specific learner characteristics at specific points in time. In digital learning environments it would be desirable to have a more detailed account of the effect of specific adaptations on specific learners (at specific timepoints) instead of just comparing a system that adapts to one that doesn't. Concerning ATI research on the other hand, we recommend moving away from highly artificial settings where one specific aptitude interacts with one specific instructional parameter towards a more ecologically valid approach, both taking into account unique constellations of multiple dynamic aptitudes on the learner side as well as dynamic flexibility in treatments.

### 4.3 Paper 2:

**Tetzlaff, L.,** Edelsbrunner, P., Schmitterer, A., Hartmann, U., & Brod, G.: *A Person-Centered Approach to Modeling the Interactions Between Learner Characteristics and Instruction: Evidence for Differential Effectiveness of Reading Education*. Manuscript submitted for publication in *Journal of Educational Psychology*.

In this paper, we present a person-centered approach to the analysis of differential effectiveness of instructional parameters in German third grade reading instruction. We first use latent-profile analysis to group students based on their patterns of means on several reading-related abilities. We then investigate whether specific teacher selected instructional foci differentially affect students in these different groups.

**Background:** In order for individualized instruction to convey benefits, different instructional approaches need to show differential effectiveness across different learners. Learners, however, do not differ only in a single relevant characteristic but different learning prerequisites can interact with each other as well as with instructional parameters to influence learning gains (Cronbach, 1975). Such multivariate aptitudes and their interaction with treatments are difficult to analyze with variable-centered approaches, due to the exorbitant power requirements and interpretational complexity that go along with higher-order interactions in multiple regression models.

In the present work, the person-centered approach of latent profile analysis (Hickendorff et al., 2018) is used to examine the differential effectiveness of different instructional foci on the development of reading comprehension across different multivariate aptitude profiles. Person-centered approaches have the advantage of parsimoniously representing multivariate constellations of learning conditions that can be difficult to model and interpret with variable-centered approaches, while also allowing for nonlinear interactions (Bauer & Shanahan, 2007).

**Method:** Data from a longitudinal study (pre at the beginning of the school year/post at the end of the school year) on reading and reading-related skills of $N = 517$ students from Hesse and Lower Saxony during the third school year were analyzed. Reading prerequisites (decoding, syntax comprehension, vocabulary) were collected from students at pretest. Reading comprehension was collected from the students at both pre- and posttest. The teachers' (N = 49) reported their self-chosen teaching focus every three weeks, using a short online questionnaire. These reports were averaged over the school year.

We used latent profile analysis to segment 517 students into homogenous subgroups (latent profiles) according to their patterns of means across their decoding ability, syntax comprehension, vocabulary and reading comprehension. In a second step, a regression model

using the measurement error-correcting BCH approach (Asparouhov & Muthén, 2014) was used to test whether different instructional foci (vocabulary, advanced reading skills, and reading motivation) showed differential effectiveness for the different profiles.

**Results and their significance:** Based on a selection of fit criteria, we identified four profiles. Consistent with the simple view of reading, these could be labeled as "poor decoders" (30% of students), "poor comprehenders" (38%), "poor readers" (15%), and "good readers" (17%).

Concerning the interaction with instruction, an instructional focus on vocabulary over a school year primarily benefited "good readers" ($\beta = 0.33$, $p = .003$), at the expense of "poor comprehenders" ($\beta = -0.21$, $p = .051$) and "poor decoders" ($\beta = -0.22$, $p = .036$). In contrast, a focus on advanced reading skills, such as text comprehension, benefited "poor comprehenders" ($\beta = 0.22$, $p = .029$), at the expense of "good readers" ($\beta = -0.34$, $p = .003$).

These results suggest that there is a need for stronger individualization in regular classroom practice, because instruction targeting the whole class will always be a wasted opportunity for some subgroups of students, depending on their specific constellations of learner characteristics. We argue that in order to accurately capture individual differences in treatment response, multiple variables should be taken into account simultaneously. Person-centered analyses therefore provide a promising approach to identifying determinants of differential effectiveness of instruction.

### 4.4 Paper 3

**Tetzlaff, L.,** Hartmann, U., Dumont, H., & Brod, G.: *Assessing Individualized Instruction in the Classroom: Comparing Teacher, Student and Observer Perspectives.* Manuscript revised and resubmitted at *Learning & Instruction*.

In this paper, we investigated possible approaches for measuring individualized instruction in a classroom context. Utilizing teacher self-reports, student self-reports, and classroom observations, we probe the unique characteristics of the different approaches as well as the agreement between them.

**Background:** While individualization of instruction has been considered an important goal of pedagogical research and practice for years (Dockterman, 2018; Hess & Lipowsky, 2017), there is still a lack of instruments that allow for a reliable and valid assessment of the extent of individualization in regular classroom instruction. Measurement of classroom processes can be (and historically has been) approached from several distinct perspectives, each offering their own unique advantages and disadvantages (Kunter & Baumert, 2006; Lüdtke et al., 2006). In this study, self-report data from teachers and students were combined with in-situ observations to investigate the reliability of different approaches to measuring individualization as well as the agreement between them. Concerning individualized instruction, it is of special importance to find common ground between those perspectives: Only when there is an alignment between the scientific operationalization (as assessed by external observers), the implementation of teachers (as assessed via self-reports) and the actually experienced individualization by students (as assessed via self-reports) can theoretical claims about the effectiveness of individualization be empirically investigated in a classroom context. Studies on other classroom processes have found considerable agreement between external observers and students as well as teachers and students and little to no agreement between teachers and external observers (Fauth et al., 2014; Scherzinger & Wettstein, 2019).

**Method:** Data collection was conducted in third grade German reading classes in a total of 57 classes from 34 schools in Hesse, Germany. For the teacher perspective ($N$ = 57), we used parts of the DSAQ questionnaire (Prast et al., 2015) at the end of the school year and a

retrospective questionnaire repeated every three weeks. In situ observations were conducted once during the school year by trained observers using a standardized questionnaire. Student perspectives ($N$ = 621) were collected at the end of the school year, using a brief self-report questionnaire (Dumont, 2016). Individualized instruction was operationalized as at least one student working on a different task than the rest of the class at a specific point in time.

**Findings:** All three perspectives yielded reliable indicators of individualization, but not all agreed with each other. We found considerable agreement between students and observers ($r$ = .43, $p$ = .01), but neither students ($r$ = .03) nor observers ($r$ = .06) agreed with teachers' trait self-reports. Using retrospective teacher ratings given shortly after the time point of interest, we found that student ratings were significantly correlated with them ($r$ = .38, $p$ = .01). After correcting for response tendencies, the correlation between teacher and student ratings was even more pronounced ($r$ = .49, $p$ < .01). This is in line with previous research on other constructs relating to classroom instruction (Fauth et al., 2014; Kunter & Baumert, 2006; Scherzinger & Wettstein, 2019). The strong agreement of observers and students has been explained by both of them being external observers of the teacher (who is in control of the classroom instruction). The agreement between students and their teachers is commonly explained by their shared classroom history, while none of the two above mentioned mechanisms apply to the observer-teacher agreement (Fauth et al., 2014).

This implies that the construct of individualized instruction does exist at the classroom level shared between students and teachers. Furthermore, this conceptualization at least partially overlaps with the scientific operationalization as assessed by the external observers. These results pave the way for further studies that aim to empirically investigate the amount of individualized instruction present in regular classroom instruction, either as a cause of learning gains or as an outcome of an intervention.

**4.5 Paper 4:**

Schmitterer A., **Tetzlaff, L.,** Hasselhorn, M., & Brod, G: *Who benefits from Computerized Learning Progress Assessment in Reading Education? Evidence from a Two-Cohort Longitudinal Study.* Manuscript submitted for Publication in Journal of Computer Assisted Learning.

The goal of this paper was to investigate whether the usage of formative assessment tools improves learning gains in the domain of reading. To this end, we compared classes that used the formative assessment program "quop" with classes that didn't. We further investigated whether these effects are qualified by students' initial level of reading comprehension.

**Background:** Formative assessment procedures are one of the most promising approaches for increasing the amount and the effectiveness of individualization in regular classroom instruction. While several meta-analyses showed positive effects of formative assessment in general (e.g. Jung et al., 2018; Kingston & Nash, 2011), there exists a strong heterogeneity in the strength of these effects. This heterogeneity indicates a need for further investigation into factors that potentially moderate the positive effects of formative assessment. Several studies have looked at factors concerning the teacher (e.g. experience or social pressure; Schildkamp et al., 2020) or the specific program (e.g. computer-based vs. pen-and-paper; Kingston & Nash, 2011) but it is also conceivable that characteristics of the students, such as their initial performance, moderate the effects of formative assessment.

**Method:** Students ($N$ = 668) from 77 classes (41 of which used the formative assessment tool) were tested at the beginning of the school year as well as after the summer holidays. At both timepoints, their reading comprehension level was assessed via the ELFE II (Lenhard et al., 2017). We analyzed whether students' reading comprehension skills improved

more in classes with formative assessment than in control classes. Furthermore, we explored whether students' initial achievement level moderated the effect size of the formative assessment. We controlled for the fact that the study was conducted over two cohorts and in two different federal states.

**Findings:** Students in formative assessment classes on average showed better reading comprehension gains than in the control classes, indicating a general advantage of using formative assessment procedures. This effect was especially pronounced for students with low initial levels of reading comprehension, while no significant difference was shown for students one standard deviation above the mean at pretest. These findings imply that a possible mechanism of the positive effects is an increased focus on weak students. It is possible that the main use of the formative assessment information is to identify those students who need help, rather than inform *what* specific help they need. Another potential explanation of these differential effects is that teachers adapt their whole-class instruction in a way that primarily benefits low achievers.

## 5 Discussion

In the following paragraphs, I will summarize and evaluate the major results of this thesis – separated into substantive findings and methodological implications. I will first describe how these studies add to the current literature by furthering our theoretical understanding of the factors that constitute effective individualization as well as its implementation in classroom instruction. I will then continue by discussing the methodological implications of the studies, namely the potential of person-centered analysis for the modelling of multivariate learner characteristics and the peculiarities of using teacher self-reports for the assessment of classroom processes. Finally, I will describe the most important limitations of

this thesis and discuss potential future directions for research on individualized instruction that addresses those limitations and builds upon the presented findings.

## 5.1 Summary and Implications of Substantive Findings

### *5.1.1 Modelling of Learner Characteristics*

A dynamic measurement approach to individualization is necessary to successfully adapt to learners at specific points in the learning process. This claim is supported by studies showing the relative success of dynamic individualization endeavors as compared to static ones, as described in *Paper 1*. The advantage of formative assessment compared to purely summative measures for the development of readers who struggle at the beginning of the schoolyear (*Paper 4*) also adds to the literature that indicates benefits of dynamic assessment of learner characteristics (e.g. Jung et al., 2018; Kingston & Nash, 2011).

Another relevant finding for the modelling of learner characteristics is that multivariate learner profiles offer information beyond the isolated variables they are comprised of. This information can then be used to explain the differential effectiveness of instructional approaches. Especially in domains where the "outcome" to be learned is a complex skill (constituted by an interplay of several lower-level skills) such as reading comprehension (Hoover & Gough, 1990; Kendeou et al., 2009, 2016), multiple relevant learner characteristics interact with each other as well as with the instructional treatment. These complex interactions need to be taken into account in order to effectively adapt instruction to specific learners. *Paper 2* could show that multivariate aptitude conceptualizations explained differential effectiveness of instructional foci that might have been overlooked if only single variables and their interaction with instruction had been considered.

The findings of *Paper 4* suggest a similar conclusion. Although the class as a whole improved more if the formative assessment program was implemented, it was mostly learners

showing a low performance in the composite skill “reading comprehension” that differed from those in the control group. The focus teachers apparently placed on said students, might be due to the fact that the tool assessed reading comprehension as a whole, rather than its constituting factors. It is conceivable that information about multivariate constellations of skills could allow teachers to better adapt instruction to learners who struggle with a specific “subskill” albeit still showing average or better performance in the composite measure.

### *5.1.2 Promoting and Assessing Individualized Instruction in the Classroom*

Concerning classroom instruction, formative assessment can be seen as an effective tool for increasing learning gains, especially for children who struggle at the beginning of the schoolyear. These findings from *Paper 4* in the context of reading are in accordance with meta-analyses by Jung et al., (2018) as well as Kingston and Nash (2011), who found higher average effect sizes for students with intensive learning needs than in regular classes independent of domain. There are two possible explanations for these differential effects: teachers could use the formative assessment data to identify struggling learners in the class and then provide increased attention to them – this would lead to increased learning gains for learners who perform poorly at the beginning of the school year but not affecting those performing well. Another possible explanation is that teachers use the formative assessment data to adapt their whole-class instruction to primarily benefit struggling learners rather than targeting instruction at individual learners. If they were actually using the information to target instruction at *each* student specifics needs it would be reasonable to assume that the effect would be equal or potentially even stronger for high achieving students – in line with the Matthew effect (Duff et al., 2015; Onwuegbuzie et al., 2003). It is thus not clear whether teachers actually used the formative assessment data to adapt instruction to specific learners (*how* to focus on them) or

whether it mainly helped them identify struggling learners (*who* to focus on) whom teachers then simply gave more attention instead of specifically targeted instruction.

While individualization is a concept that is operationalized by researchers, administered by teachers and experienced by students, *Paper 3* showed that significant overlap between these perspectives exists, even when they are assessed with non-parallel instruments that focus on the aspects that are especially relevant for these specific perspectives. This overlap is of utmost importance for future research on the topic of individualized classroom instruction. Only when different actors are at least partially referencing the same situations when talking about individualized instruction can we actually draw conclusions that are relevant beyond the specific perspective used for measurement.

Establishing this also allows for further studies to investigate the actual amount of individualization as a mediating mechanism of formative assessment – a concept which has been proposed several times (Brink et al., 2019; Cusi & Telloni, 2019; Jung et al., 2018; Kaftan et al., 2006; Yeh, 2010), but to the best of my knowledge not yet investigated.

### 5.2 Methodological Implications

In addition to the substantive findings elaborated above, the results of the studies in this dissertation also carry methodological implications, both for the modelling of learner characteristics and for the assessment of instructional parameters.

When modelling multivariate learner characteristics, variable centered approaches such as multiple regression models often encounter the problem of higher-order interactions leading to power requirements that are hard to obtain as well as interpretational complexity of results (Bauer & Shanahan, 2007; Cronbach, 1975; Preacher & Sterba, 2018). *Paper 2* argued that Person-centered approaches (Bauer & Shanahan, 2007) such as latent profile analysis (e.g. Hickendorff et al., 2018) can be used to model multivariate aptitudes without encountering

these problems. Utilizing the so-called BCH-approach (Asparouhov & Muthén, 2014) to integrate the latent profiles into a secondary model allows for doing so without embezzling measurement error. By demonstrating interactions between multivariate learner profiles and treatment parameters that would have remained hidden in univariate analysis, *Paper 2* could show both the necessity of multivariate learner modelling in individualization research and the utility of person-centered approaches towards such endeavors.

Another methodological implication of this thesis is that when using teacher self-reports as indicators of classroom processes or for the assessment of instructional parameters, it is advisable to frame these self-reports in a way that clearly references specific past behavior instead of general tendencies to act in a certain way. *Paper 3* could show that this approach increases the correspondence with other perspectives (compared with trait-like measures of general behavior), presumably by lessening the impact of the socially desirable response tendencies (Kopcha & Sullivan, 2007). This is in line with other studies showing that when assessing classroom management, teacher self-reports showed higher correspondence with other perspectives when they were specifying a timeframe/specific situations as well as the classroom context (Clunies-Ross et al., 2008; Scherzinger & Wettstein, 2019). On a more general note, *Paper 2* could show that teacher self-reports on instructional practices can be used as an alternative to separate (randomized) treatment conditions in order to better capture the natural variability in teaching (at the cost of better experimental control). Taken together these findings offer several promising ways forward for research on ATIs that better reflects the complex and dynamic nature of instructional processes.

## 5.3 Limitations and Future Directions

In the following paragraphs, I will list some of the limitations that apply to all the studies in this dissertation. I will combine this with potential steps that could be taken to alleviate these limitations and further build upon the presented findings in future work.

### *5.3.1 Generalizability of Findings*

Since all of the studies in this dissertation are based on the iLearn Project, this brings with it some limitations concerning the generalizability of the findings. First and foremost, all of the studies were situated in the domain of reading, limiting the generalizability to other domains. Similarly, all of the students in the iLearn sample were German third grade students. Whether these findings generalize to other countries and age groups, needs to be tested empirically.

Concerning the teacher sample, since both participation in the study and usage of the "quop" program was entirely voluntary, it is reasonable to assume that our sample is at least partially influenced by positive biases of self-selection (Heckman, 1990). Meaning, teachers in our sample potentially show more behavior that is assumed to have positive impact on children (such as individualization) than the average teacher might. This could have an effect on the results of *Paper 3* and *Paper 4*. Future studies could aim to replicate our findings in other domains, contexts, and age-groups.

### *5.3.2 Learner-Centered Conceptualizations of Individualization*

All the studies in this dissertation utilized a very teacher-centric conceptualization of individualization. Other conceptualizations put a stronger focus on student participation in goal setting and task selection, banking on self-regulated learning as a part of individualization (Crosby & Fremont, 1960; McLoughlin & Lee, 2009). While learners have repeatedly been

shown not to select optimal tasks for themselves without guidance (Nugteren et al., 2018; Son & Metcalfe, 2000), there exist several studies that connect the practice of formative assessment with learner-driven individualization. In these studies formative assessment data is fed back to the students either directly or mediated through a teacher in order to allow learners to select tasks in accordance with their own strengths and deficits (Clark, 2012; Greene, 2020; Nicol & MacFarlane-Dick, 2007; Panadero et al., 2018).

Formative assessment has also repeatedly been linked to the development of self-regulated learning skills in students (Granberg et al., 2021; Xiao & Yang, 2019). This connection is interesting because self-regulated learning skills and associated metacognitive processes have also been conceptualized as a prerequisite for students to make effective choices in their own learning path (e.g. Dörr & Perels, 2019; Kuhn, 2016; Zhang & Zhang, 2019). This double-role of self-regulated learning skills as a mediator as well as an outcome of learner-driven individualization can be embraced by gradually shifting from a teacher-driven to a learner-driven model of goal setting and task selection (e.g. Corbalan et al., 2006; Salden et al., 2006).

As self-regulation related constructs show substantial intra-individual variance (e.g. Breitwieser & Brod, 2022), such student-centered perspectives on individualization should also benefit from a dynamic conceptualization. Such a dynamic conceptualization would include dynamic assessment of self-regulation with concurrent adaptations, taking into account findings from the ATI-literature, such as the expertise reversal effect for metacognitive prompts (Nückles et al., 2010, 2020). In a similar vein, it is conceivable that the amount of optimal guidance is dependent on a mixture of self-regulated learning skills, prior knowledge, cognitive capabilities, and thematic interest, with each variable influencing the effect of the others. In this case, a multivariate conceptualization would also be recommendable. Besides probing the

utility of dynamic and multivariate measurement for student-centered individualization, future studies could compare student-centered formative assessment to teacher-centered approaches. This would allow them to establish the relative strengths of both approaches as well as differentiate which parts of formative assessment practice actually benefit from teacher guidance.

### *5.3.3 Affective/Motivational Aptitudes and Outcomes*

Originally, ATI research often focused on affective and conative aptitudes, as well as personality traits (Cronbach & Snow, 1977; Snow, 1989, 1992). Especially the multivariate aptitude complexes postulated by Snow (1987) and trait complexes postulated by Ackerman, (2003) incorporated a mixture of cognitive, affective and conative variables. More recent work on individualized instruction could also show substantial positive effects on learning gains when the context of tasks was adapted to the specific interests of learners (Bernacki & Walkington, 2018; Walkington, 2013).

In contrast, the studies in this dissertation only looked at cognitive capabilities and prior performance as learning prerequisites and used learning gains as primary outcome measures. Future studies could incorporate affective-conative variables in multivariate learner models in line with the originally postulated aptitude complexes (Snow, 1987). It would also be interesting to look at affective-conative outcome variables, as a treatment that provides equal learning gains but higher enjoyment in specific learners can arguably be categorized as superior to an alternative that causes equal learning gains with less enjoyment in those learners.

### *5.3.4 Other Future Directions*

Another promising future direction would be to utilize formative assessment to provide feedback about multivariate learner profiles and their implication for instruction to teachers.

While several formative assessment programs already assess more than one variable (e.g. quop; Souvignier et al., 2021), they are usually fed back without any connection to each other. While this information can help teachers target instruction at the single measure children performed worst at, *Paper 2* could show that this does not necessarily lead to the best learning gains, if those variables meaningfully interact with each other. By integrating the individual assessment to comprehensive learner profiles, teachers can target instruction in a way that takes the whole learner in their specific constellation of strength and deficits into account, rather than just focusing on the single variable that seems to be weakest.

Similarly, research on formative assessment could be enriched by analyzing in detail the specific adaptations teachers take in response to specific patterns displayed in the formative assessment results as well as the resulting changes in the skills measured by formative assessment. This would require a very measurement-intensive design in order to correctly differentiate between stable interindividual differences, developmental trajectories, and intervention induced changes. A possible statistical approach for investigating such effects would be a random-intercept cross-lagged panel model (Hamaker et al., 2015).

Last but not least, it would be interesting to combine the assessment of individualized instruction in classrooms as demonstrated in *Paper 3* with a study on formative assessment to investigate whether the amount of individualization actually mediates the positive effects of formative assessment as postulated by Jung et al., (2018) among others.

### 5.4 Conclusion

This dissertation provides insights into relevant factors for each of the necessary steps for effective individualized instruction. For assessing learner characteristics, I argue for the necessity of dynamic and multivariate measurement in order to accurately capture learners in

their unique constellation of skills and their trajectories. For modelling the interaction of learner characteristics with instruction, I provide evidence for the advantages of utilizing person-centered, as opposed to variable-centered approaches. These findings provide a conceptual and methodological basis for future ATI research to more accurately model learners' aptitudes and their interaction with instruction.

For the implementation of individualized instruction in regular classroom practice, I demonstrated substantial within-class heterogeneity of the positive effects of formative assessment procedures. This provides some insight into the way formative assessment data is used by teachers as well as potential directions to improve the procedure. For the evaluation of individualized classroom instruction, I provided instruments that can reliably assess different aspects of individualized instruction from different perspectives while showing substantial overlap between them. These findings open up several directions for investigating the effects of individualized classroom instruction at the level of the individual learner.

Taken together, this thesis addresses several gaps in our understanding of successful individualization: from the modelling of learner characteristics over the fit of specific instructional parameters and specific learners to the assessment of actual individualization in regular classroom practice. It thus serves as a basis for future research to bridge the gap between theoretical considerations and actual classroom instruction.

## German Summary (Zusammenfassung)

Die Individualisierung von Unterricht wird seit langem als ein wichtiger Bestandteil effektiver Bildung angesehen (Corno, 2008; Dockterman, 2018). Individualisierung kann als die Anpassung von Unterrichtsparametern an relevante Merkmale bestimmter Lernenden definiert werden. Diese Definition wirft jedoch mehrere Fragen auf: Welche Merkmale sind tatsächlich relevant? Welche Unterrichtsparameter müssen auf welche Weise angepasst werden, um mit diesen Merkmalen positiv zu interagieren? Im Kontext des Klassenunterrichts stellen sich weitere Fragen: Wie können Informationen über die relevanten Merkmale der Lernenden an die Lehrkraft weitergegeben werden? Wie kann in einem Kontext, der hauptsächlich auf den Unterricht mit der ganzen Klasse ausgerichtet ist, jedem Lernenden ein individueller Unterricht erteilt werden? Diese Dissertation konzentriert sich auf die Messung und Modellierung von Lernendencharakteristika und Unterrichtsanpassungen und soll einen Einblick in jede dieser Fragen geben. Sie besteht aus vier Schriften und lässt sich grob in zwei Teile aufteilen:

Im ersten Teil werfe ich einen Blick auf zwei Aspekte der Messung und Modellierung relevanter Lernendencharakteristika - nämlich die Notwendigkeit einer dynamischen und multivariaten Modellierung. Ich argumentiere, dass dies notwendig ist, um Unterrichtsparameter so auszuwählen, dass der Lernzuwachs für eine*n bestimmte*n Lernende*n zu einem bestimmten Zeitpunkt im Lernprozess maximiert wird.

Die erste Schrift fasst die bestehende Forschung zur Individualisierung aus drei verschiedenen Forschungstraditionen – Unterrichtsforschung, Experimentalpsychologie und digitalem Lernen - zusammen. Aus dieser Zusammenfassung leite ich die Notwendigkeit einer dynamischen Konzeptualisierung von Lernendencharakteristika ab. Lernende verändern sich auf unterschiedlichen Zeitskalen: von Entwicklungsprozessen, die sich über Monate oder sogar

Jahre hinweg entfalten, über interventionsbedingte Veränderungen, die sich über Wochen hinweg vollziehen, bis hin zu kurzfristigen Schwankungen, die über Tage oder sogar von Augenblick zu Augenblick auftreten können (Hertzog & Nesselroade, 2003). Diese verschiedenen Arten der Varianz erfordern unterschiedliche Messansätze und bieten Möglichkeiten für unterschiedliche Unterrichtsanpassungen. Wenn der Unterricht optimal an bestimmte Lernende angepasst werden soll, muss diese Dynamik auf unterschiedlichen Zeitskalen berücksichtigt werden. Auf der Grundlage dieser Überlegungen konstruieren wir einen dynamischen Rahmen für individualisierten Unterricht, der die relevanten Messansätze sowie das Potenzial für Unterrichtsanpassungen auf drei verschiedenen Zeitskalen darstellt. Diese reichen von der Festlegung geeigneter Lernziele auf der Makroskala über die Unterrichtsgestaltung auf der Mesoskala bis hin zur Reaktion auf affektiv-motivationale Schwankungen auf der Mikroskala.

Die zweite Schrift berichtet die Ergebnisse einer explorativen Studie, die den potenziellen Nutzen von personenzentrierten Analysen für die Erfassung multivariater Lernvoraussetzungen und deren Interaktion mit Unterricht untersuchte. Dabei wurden Daten einer Längsschnitterhebung (prä am Anfang des Schuljahres/post am Ende des Schuljahres) zu Lese- und leserelevanten Fähigkeiten von $N$ = 517 Schüler*innen im Laufe des dritten Schuljahres analysiert. Als Lernvoraussetzungen wurden bei den Schüler*innen am Prätest Dekodieren, Syntaxverständnis, Wortschatz und als Lernziel (sowohl Prä- als auch Posttest) das Leseverständnis erhoben. Die gewählten Unterrichtsschwerpunkte der Lehrkräfte ($N$ = 49) wurden alle 3 Wochen mithilfe eines kurzen online-Fragebogens als Selbstbericht erfasst und über das Schuljahr gemittelt. Wir fanden heraus, dass latente Profile – ermittelt über mehrere lesebezogene Fähigkeiten – die differentielle Wirksamkeit von selbstberichteten Unterrichtsschwerpunkten im deutschen Leseunterricht der dritten Klasse erklären können. Dies weist einerseits auf die Notwendigkeit einer stärkeren Individualisierung dieses

Unterrichts hin, andererseits illustriert es den Vorteil multivariater Konzeptualisierungen sowie den Nutzen personenzentrierter Ansätze für die Untersuchung multivariater Lernendencharakteristika und ihrer Interaktion mit Unterrichtsparametern. Im zweiten Teil untersuche ich Fragen zur Umsetzung von Individualisierung im Klassenraum. Erfolgreiche Individualisierung im Klassenraum hat die zusätzliche Schwierigkeit, dass nicht nur Wissen über relevante Lernendencharakteristika und deren Interaktion mit Unterrichtsparametern erforderlich ist, sondern auch Raum und Gelegenheit gefunden werden muss, um einzelne Schüler*innen individualisiert zu unterrichten, ohne den Rest der Klasse zu beeinträchtigen.

In der dritten Schrift wird untersucht, ob Lehrkraft-, Schüler*innen- und Beobachtendenperspektiven bei der Bewertung des Ausmaßes der Individualisierung im regulären Unterricht übereinstimmen. Dies ist eine wichtige Voraussetzung für weitere Forschung zur Individualisierung im Unterricht. Nur wenn eine Übereinstimmung zwischen der wissenschaftlichen Operationalisierung (wie sie von externen Beobachtern verwendet wird), der tatsächlichen Umsetzung des Konzepts durch die Lehrkräfte (wie sie anhand von Selbstberichten eingeschätzt wird) und der tatsächlich erlebten Individualisierung durch die Schüler*innen (wie sie anhand von Selbstberichten eingeschätzt wird) besteht, können theoretische Behauptungen über die Wirksamkeit von Individualisierung im Unterrichtskontext empirisch untersucht werden. Zu diesem Zweck wurden in insgesamt 57 Klassen aus 34 hessischen Grundschulen Daten über den Deutschunterricht in der dritten Klasse erhoben. Für die Lehrkraftperspektive ($N$ = 57) wurden Teile des DSAQ-Fragebogens (Prast et al., 2015) am Ende des Schuljahres und ein retrospektiver Fragebogen, der alle drei Wochen wiederholt wurde, verwendet. In-situ-Beobachtungen wurden einmal während des Schuljahres von geschulten Beobachtenden anhand eines standardisierten Fragebogens durchgeführt. Die Schüler*innenperspektiven ($N$ = 621) wurden am Ende des Schuljahres mit

einem kurzen Selbstauskunftsfragebogen (Dumont, 2016) erhoben. Individualisierter Unterricht wurde operationalisiert als Situation in der mindestens ein*e Schüler*in an einer anderen Aufgabe arbeitete als der Rest der Klasse.

Alle drei Perspektiven ergaben zuverlässige Indikatoren für Individualisierung, aber nicht alle stimmten miteinander überein. Wir fanden eine beträchtliche Übereinstimmung zwischen Schüler*innen und Beobachtenden aber weder Schüler*innen noch Beobachtende stimmten mit den Selbsteinschätzungen der Lehrkräfte überein. Bei der Verwendung retrospektiver Lehrkraftbewertungen, die kurz nach dem interessierenden Zeitpunkt abgegeben wurden, zeigte sich, dass die Schüler*innenbewertungen signifikant mit diesen korreliert waren. Die gefundene Übereinstimmung zwischen den Perspektiven deutet auf ein gemeinsames Verständnis des Konstrukts auf Klassenebene hin und liefert einige Belege für die Validität der verwendeten Messinstrumente.

In der vierten Schrift werden Ergebnisse zu positiven Effekten von Lernverlaufsdiagnostik auf den Lernzuwachs repliziert, ergänzt durch eine Moderatorenanalyse, die zeigt, dass vor allem Kinder mit schwachen Leistungen zu Beginn des Schuljahres von der Implementation des Verfahrens profitieren. Dazu wurden Schüler*innen ($N$ = 668) aus 77 Klassen (von denen 41 die Lernverlaufsdiagnostik verwendeten) zu Beginn des Schuljahres und nach den Sommerferien getestet. Zu beiden Zeitpunkten wurde ihr Leseverständnis mit dem ELFE II (Lenhard et al., 2017) gemessen. Die Schülerinnen und Schüler in den Lernverlaufsdiagnostik-Klassen zeigten im Durchschnitt bessere Fortschritte im Leseverständnis als in den Kontrollklassen, was auf einen generellen Vorteil des Einsatzes von Lernverlaufsdiagnostik hinweist. Dieser Effekt war besonders ausgeprägt bei Schüler*innen mit niedrigem Ausgangsniveau im Leseverständnis, während bei Schüler*innen, die beim Prätest eine Standardabweichung über dem Mittelwert lagen, kein signifikanter Unterschied festgestellt wurde. Dies deutet darauf hin, dass die Informationen,

die die Lernverlaufsdiagnostik liefert, vor allem dazu beitragen, Leseschwächen zu erkennen und Kinder mit Schwächen individuell zu fördern, aber anscheinend nicht dazu genutzt werden, den Unterricht an die spezifischen Defizite von Kindern mit durchschnittlichen oder hohen Leistungen anzupassen.

Zusammengenommen bestätigen diese Befunde die Notwendigkeit von individualisiertem Unterricht für die Maximierung der Lernzuwächse einzelner Schüler*innen. In einem Feld, in dem die einschlägige Forschung über mehrere verschiedene Disziplinen und Analyseebenen verteilt ist, dient die vorliegende Dissertation als wichtige Brücke zwischen konzeptionellen Überlegungen und dem tatsächlichen Unterricht im Klassenzimmer. In Bezug auf die Methodik zeigt sie auf, dass eine wirksame Individualisierung im besten Fall auf einer dynamischen und multivariaten Diagnostik beruhen sollte. Schließlich zeigt sie den Nutzen von personenzentrierten Analyseansätzen für die Beantwortung von Fragen nach der differentiellen Wirksamkeit von Unterrichtsansätzen bei multivariaten Lernendenprofilen.

Zukünftige Forschung könnte diese Befunde aufgreifen um den Mehrwert multivariater Lernendenprofile in anderen Domänen als Lesen zu bestätigen und mit einem dynamischen Modellierungsansatz zu verknüpfen. Auch wäre es interessant, die Rolle von individualisiertem Unterricht als Mediator der positiven Effekte von Lernverlaufsdiagnostik empirisch zu untersuchen. Die vorliegende Arbeit liefert für beide Fragestellungen methodische und konzeptuelle Grundlagen.

## References

Ackerman, P. L. (2003). Aptitude complexes and trait complexes. *Educational Psychologist*, *38*(2), 85–93. https://doi.org/10.1207/S15326985EP3802_3

Allington, R. (1974). Differentiating Instruction to Improve Comprehension in Middle School Content Areas. *Paper presented at the Annual Meeting of the International Reading Association* (19th, New Orleans, May 1-4, 1974)

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. S*tructural Equation Modeling: A Multidisciplinary Journal*, 21, 329-341.

Bardy, T., Holzäpfel, L., & Leuders, T. (2021). Adaptive Tasks as a Differentiation Strategy in the Mathematics Classroom: Features from Research and Teachers' Views. *Mathematics Teacher Education and Development*, *23*, 26–53.

Barrows, H. S., Myers, A., Williams, R. G., & Moticka, E. J. (1986). Large group problem-based learning: A possible solution for the "2 sigma problem." *Medical Teacher*, *8*(4), 325–331. https://doi.org/10.3109/01421598609028991

Bauer, D. J., & Cai, L. (2009). Consequences of Unmodeled Nonlinear Effects in Multilevel Models. *Journal of Educational and Behavioral Statistics*, *34*(1), 97–114. https://doi.org/10.3102/1076998607310504

Bauer, D. J., & Shanahan, M. J. (2007). Modeling complex interactions: Person-centered and variable-centered approaches. *Modeling contextual effects in longitudinal studies*, *21*, 255-83.

Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). A Systematic Review of Research on Personalized Learning : Personalized by Whom , to What , How , and for What Purpose ( s )? *Educational Psychology Review.*

Bernacki, M. L., & Walkington, C. (2018). The role of situational interest in personalized learning. *Journal of Educational Psychology*, *110*(6), 864–881. https://doi.org/10.1037/EDU0000250

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, *1*(2), 1–12. https://doi.org/10.1021/ed063p318

Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as

Effective as One-to-One Tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Boekaerts, M., & Corno, L. (2005). Self-Regulation in the Classroom: A Perspective on Assessment and Intervention. *Applied Psychology*, *54*(2), 199–231. https://doi.org/10.1111/J.1464-0597.2005.00205.X

Bracht, G. H. (1970). Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, *40*(5), 627–645. https://doi.org/10.3102/00346543040005627

Breitwieser, J., & Brod, G. (2022). The interplay of motivation and volitional control in predicting the achievement of learning goals: An intraindividual perspective. *Journal of Educational Psychology*. Advance Online Publication.

Breitwieser, J., Neubauer, A., Schmiedek, F., & Brod, G. (2021). Self-Regulation prompts promote the achievement of learning goals – but only briefly: Uncovering hidden dynamics in the effects of a psychological intervention. *Learning and Instruction.* Advance Online Publication.

Brink, M., Bartz, D. E., & And Bartz, D. E. (2019). Effective Use of Formative Assessment by High School Teachers. *Practical Assessment, Research, and Evaluation*, *22*(1), 8. https://doi.org/https://doi.org/10.7275/p86s-zc41

Chen, O., Kalyuga, S., & Sweller, J. (2017). The Expertise Reversal Effect is a Variant of the More General Element Interactivity Effect. *Educational Psychology Review*, *29*(2), 393–405. https://doi.org/10.1007/s10648-016-9359-1

Clark, I. (2012). Formative Assessment: Assessment Is for Self-regulated Learning. *Educational Psychology Review*, *24*(2), 205–249. https://doi.org/10.1007/S10648-011-9191-6/TABLES/1

Clunies-Ross, P., Little, E., & Kienhuis, M. (2008). Self-reported and actual use of proactive and reactive classroom management strategies and their relationship with teacher stress and student behaviour. *Educational Psychology*, *28*(6), 693–710. https://doi.org/10.1080/01443410802206700

Connor, C. M., Mazzocco, M. M. M., Kurz, T., Crowe, E. C., Tighe, E. L., Wood, T. S., & Morrison, F. J. (2018). Using assessment to individualize early mathematics instruction. *Journal of School Psychology*, *66*, 97–113. https://doi.org/10.1016/j.jsp.2017.04.005

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science*, *315*(5811), 464–465.

https://doi.org/10.1126/science.1134513

Connor, C. M., Piasta, S. B., Glasney, S., Schatschneider, C., Fishman, B. J., Underwood, P. S., & Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child-by-instruction interactions on students' literacy. *Child Development*, *80*(1), 77–100.

Constas, M. A., & Sternberg, R. J. (2013). Translating theory and research into educational practice: Developments in content domains, large-scale reform, and intellectual capacity. In *Translating Theory and Research Into Educational Practice: Developments in Content Domains, Large-Scale Reform, and Intellectual Capacity*. Taylor and Francis. https://doi.org/10.4324/9780203726556

Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2006). Towards a personalized task selection model with shared instructional control. *Instructional Science*, *34*(5), 399–422. https://doi.org/10.1007/s11251-005-5774-2

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2109*, 137–147. https://doi.org/10.1007/3-540-44566-8_14

Corno, L. (2008). On Teaching Adaptively. *EDUCATIONAL PSYCHOLOGIST*, *43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist*, *12*(11), 671.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116–127. https://doi.org/10.1037/h0076829

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Crosby, G., & Fremont, H. (1960). Individualized algebra. *The Mathematics Teacher*, *53*(2), 109–112. https://doi.org/10.2307/27956078

Cusi, A., & Telloni, A. (2019, February). The role of formative assessment in fostering individualized teaching at university level. In *Eleventh Congress of the European Society for Research in Mathematics Education* (No. 16). Freudenthal Group; Freudenthal Institute; ERME.

Decristan, J., Fauth, B., Kunter, M., Büttner, G., & Klieme, E. (2017). The interplay between class heterogeneity and teaching quality in primary school. *International Journal of Educational Research*, *86*, 109-121.

Deno, S. L. (1990). Individual differences and individual difference: The essential difference

of special education. *The Journal of Special Education*, *24*(2), 160–173. https://doi.org/10.1177/002246699002400205

Dockterman, D. (2018). Insights from 200+ years of personalized learning. *npj Science of Learning*, *3*(1), 1-6.

Dörr, L., & Perels, F. (2019). Improving metacognitive abilities as an important prerequisite for self-regulated learning in preschool children. *International Electronic Journal of Elementary Education*, *11*(5), 449-459.

Driscoll, M. P. (1987). Aptitude-treatment interaction research revisited. *Paper presented at the annual convention of the association for educational communications and technology, february.* Atlanta, GA, USA

Duff, D., Bruce Tomblin, J., & Catts, H. (2015). The Influence of Reading on Vocabulary Growth: A Case for a Matthew Effect. *Journal of Speech, Language, and Hearing Research*, *58*(3), 853–864. https://doi.org/10.1044/2015_JSLHR-L-13-0310

Dumont, H. (2016). Die empirische Untersuchung von individueller Förderung als Perspektive für die Unterrichtsqualitätsforschung. In *Bedingungen und Effekte guten Unterrichts* (pp. 107–116).

Dumont, H. (2019). Neuer Schlauch für alten Wein? Eine konzeptuelle Betrachtung von individueller Förderung im Unterricht. *Zeitschrift für Erziehungswissenschaft, 22*(2), 249-277. https://doi.org/10.1007/s11618-018-0840-0

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg. *Zeitschrift Für Pädagogische Psychologie*, *28*(3), 127–137. https://doi.org/10.1024/1010-0652/a000129

Förster, N., Erichsen, M., & Forthmann, B. (2022). Measuring reading progress in second grade: Psychometric properties of the quop-L2 test series. *European Journal of Psychological Assessment*

Förster, N., Kawohl, E., & Souvignier, E. (2018). Short- and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension. *Learning and Instruction*, *56*(April), 98–109. https://doi.org/10.1016/j.learninstruc.2018.04.009

Fraser, B. J. (1982). Individualized Classroom Environment Questionnaire. *Evaluation News, 3*(2), 72-73. https://doi.org/10.1177/109821408200300221

Fuchs, L. S. (2004). The Past, Present, and Future of Curriculum-Based Measurement Research. *School Psychology Review*, *33*(2), 188–192.

Granberg, C., Palm, T., & Palmberg, B. (2021). A case study of a formative assessment practice and the effects on students' self-regulated learning. *Studies in Educational Evaluation*, *68*, 100955. https://doi.org/10.1016/J.STUEDUC.2020.100955

Greene, J. A. (2020). Building upon synergies among self-regulated learning and formative assessment research and practice. *Assessment in Education: Principles, Policy & Practice,* 27(4), 463-476.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Heckman, J. J. (1990). Selection Bias and Self-selection. *Econometrics*, 201–224. https://doi.org/10.1007/978-1-349-20570-7_29

Hertzog, C., & Nesselroade, J. R. (2003). Assessing Psychological Change in Adulthood: An Overview of Methodological Issues. *Psychology and Aging*, *18*(4), 639–657. https://doi.org/10.1037/0882-7974.18.4.639

Hess, M., & Lipowsky, F. (2017). Lernen individualisieren und Unterrichtsqualität verbessern. In *Individualisierung im Grundschulunterricht* (pp. 23–31). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-15565-0_3

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences*, *66*, 4–15. https://doi.org/10.1016/J.LINDIF.2017.11.001

Hofer, S. I., Schumacher, R., Rubin, H., & Stern, E. (2018). Enhancing physics learning with cognitively activating instruction: A quasi-experimental classroom intervention study. *Journal of Educational Psychology*, *110*(8), 1175–1191. https://doi.org/10.1037/EDU0000266

Hooper, S. R., Wakely, M. B., de Kruif, R. E., & Swartz, C. W. (2006). Aptitude–treatment interactions revisited: Effect of metacognitive intervention on subtypes of written expression in elementary school students. *Developmental Neuropsychology*, *29*(1), 217-241

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

Jung, P.-G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of Data-Based Individualization for Students with Intensive Learning Needs: A Meta-Analysis.

*Learning Disabilities Research & Practice*, *33*(3), 144–155. https://doi.org/10.1111/ldrp.12172

Kaftan, J. M., Buck, G. A., & Haack, A. (2006). Using Formative Assessments to Individualize Instruction and Promote Learning. *Middle School Journal*, *37*(4), 44–49. https://doi.org/10.1080/00940771.2006.11461545

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension: Core components and processes. *Policy Insights from the Behavioral and Brain Sciences*, 3(1), 62-69.

Kendeou, P., Savage, R., & van den Broek, P. (2009). Revisiting the simple view of reading. *British Journal of Educational Psychology*, *79*(2), 353–370. https://doi.org/10.1348/978185408X369020

Khacharem, A., Zoudji, B., & Kalyuga, S. (2015). Expertise reversal for different forms of instructional designs in dynamic visual representations. *British Journal of Educational Technology*, *46*(4), 756–767. https://doi.org/10.1111/bjet.12167

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers and Education*, *106*, 166–171. https://doi.org/10.1016/j.compedu.2016.12.006

Kopcha, T. J., & Sullivan, H. (2007). Self-presentation bias in surveys of teachers' educational technology practices. *Educational Technology Research and Development*, *55*(6), 627–646. https://doi.org/10.1007/s11423-006-9011-8

Kuhn, D. (2016). Learning is the key twenty-first century skill. *Learning: Research and Practice, 2(2)*, 88-99.

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

Lee, C. H., & Kalyuga, S. (2014). Expertise reversal effect and its instructional implications. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning in*

*education: Infusing psychological science into the curriculum.* (pp. 32–44). Society for the Teaching of Psychology.

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The Effectiveness and Features of Formative Assessment in US K-12 Education: A Systematic Review. *Applied Measurement in Education*, *33*(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

Lehman, B., Matthews, M., D'Mello, S., & Person, N. (2008). What Are You Feeling? Investigating Student Affective States During Expert Human Tutoring Sessions. In *Intelligent Tutoring Systems* (pp. 50–59). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-69132-7_10

Lenhard, W., Lenhard, A., & Schneider, W. (2017). *ELFE II – Ein Leseverständnistest für Erst- bis Siebtklässler*. Göttingen: Hogrefe.

Lonigan, C. J., Goodrich, J. M., & Farver, J. M. (2018). Identifying differences in early literacy skills across subgroups of language-minority children: A latent profile analysis. *Developmental psychology*, *54*(4), 6

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research*, *9*(3), 215–230. https://doi.org/10.1007/s10984-006-9014-8

Ma, W., Nesbit, J. C., & Liu, Q. (2014). Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis. *Journal of Educational Psychology*, *106*(4), 901–918. https://doi.org/10.1037/a0037123.supp

Machts, N., Kaiser, J., Schmidt, F. T. C., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review*, *19*, 85–103. https://doi.org/10.1016/j.edurev.2016.06.003

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*(2), 376–390. https://doi.org/10.1037/0033-2909.114.2.376

McLoughlin, C., & Lee, M. J. W. (2009). Personalised learning spaces and self-regulated learning: Global examples of effective pedagogy. *ASCILITE 2009 - The Australasian Society for Computers in Learning in Tertiary Education*, 639–645.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational measurement: issues and practice, 37(1)*, 21-34.

Nückles, M., Hübner, S., Dümer, S., & Renkl, A. (2010). Expertise reversal effects in writing-

to-learn. *Instructional Science*, *38*(3), 237–258. https://doi.org/10.1007/s11251-009-9106-9

Nückles, M., Roelle, J., Glogger-Frey, I., Waldeyer, J., & Renkl, A. (2020). The Self-Regulation-View in Writing-to-Learn: Using Journal Writing to Optimize Cognitive Load in Self-Regulated Learning. *Educational Psychology Review*, *32*(4), 1089–1126. https://doi.org/10.1007/S10648-020-09541-1/TABLES/3

Nugteren, M. L., Jarodzka, H., Kester, L., & Van Merriënboer, J. J. G. (2018). Self-regulation of secondary school students: self-assessments are inaccurate and insufficiently used for learning-task selection. *Instructional Science*, *46*(3), 357–381. https://doi.org/10.1007/s11251-018-9448-2

Nwana, H. S. (1990). Intelligent Tutoring Systems: an overview. In *Artificial Intelligence Review* (Vol. 4).

Onwuegbuzie, A. J., Collins, K. M. T., & Elbedour, S. (2003). Aptitude by treatment interactions and Matthew effects in graduate-level cooperative-learning groups. *The Journal of Educational Research*, *96*(4), 217–230. https://doi.org/10.1080/00220670309598811

Panadero, E., Andrade, H., & Brookhart, S. (2018). Fusing self-regulated learning and formative assessment: a roadmap of where we are, how we got here, and where we are going. *Australian Educational Researcher*, *45*(1), 13–31. https://doi.org/10.1007/S13384-018-0258-Y/FIGURES/1

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology, 50(1)*, 128– 138. https://doi.org/10.1111/bjet.12592

Paris, S. G., & Paris, A. H. (2003). Classroom applications of research on self-regulated learning. In *Educational psychologist* (pp. 89-101). Routledge.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles concepts and evidence. *Psychological Science in the Public Interest, Supplement*, *9*(3), 105–119. https://doi.org/10.1111/j.1539-6053.2009.01038.x

Prast, E. J., van de Weijer, E., Kroesbergen, E. H., & van Luit, J. E. H. (2015). Readiness-based differentiation in primary school mathematics : Expert recommendations and teacher self-assessment. *Frontline Learning Research*, *3*(2), 90–116. https://doi.org/https://doi.org/10.14786/flr.v3i2.163

Preacher, K. J., & Sterba, S. K. (2019). Aptitude-by-Treatment Interactions in Research on Educational Interventions. *Exceptional Children*, *85*(2), 248-264. http://dx.doi.org/10.1177/0014402918802803

Reinhold, F., Strohmaier, A., Hoch, S., Reiss, K., Böheim, R., & Seidel, T. (2020). Process data from electronic textbooks indicate students' classroom engagement. *Learning and individual differences, 83*, 101934.

Rey, G. D., & Fischer, A. (2013). The expertise reversal effect concerning instructional explanations. *Instructional Science*, *41*(2), 407–429. https://doi.org/10.1007/s11251-012-9237-2

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/MET0000220

Salden, R. J. C. M., Paas, F., & Van Merrienboer, J. J. G. (2006). Personalised Adaptive Task Selection in Air Traffic Control. *Learning and Instruction*, *16*, 350–362.

Scherzinger, M., & Wettstein, A. (2019). Classroom disruptions, the teacher–student relationship and classroom management from the perspective of teachers, students and external observers: a multimethod approach. *Learning Environments Research*, *22*(1), 101–116. https://doi.org/10.1007/s10984-018-9269-x

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research, 103*, 101602.

Seufert, T., Schütze, M., & Brünken, R. (2009). Memory characteristics and modality in multimedia learning: An aptitude–treatment–interaction study. *Learning and Instruction*, *19*(1), 28–42. https://doi.org/10.1016/J.LEARNINSTRUC.2008.01.002

Shapiro, K. R. (1975). An overview of problems encountered in aptitude-treatment interaction (ATI) research for instruction. *Educational Communication and Technology Journal*, *23*(2), 227–241. https://doi.org/10.1007/BF02768380

Slavin, R. E. (1987). Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis. *Review of Educational Research*, *57*(3), 293–336. https://doi.org/10.3102/00346543057003293

Slavin, R. E., & Karweit, N. L. (1985). Effects of Whole Class, Ability Grouped, and Individualized Instruction on Mathematics Achievement. *American Educational Research Journal*, *22*(3), 351–367. https://doi.org/10.3102/00028312022003351

Snow R. E. *Aptitudes and instructional methods: Research on individual differences in*

*learning-related processes Final report 1975–1979*, Aptitude Research Project, School of Education, Stanford University, Sept. 1980.

Snow, R. E. (1987). Aptitude complexes. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and instruction: Vol. 3. Conative and affective process analyses* (pp. 13-59). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Snow, R. E. (1989). Cognitive-conative aptitude interactions in learning. In R. Kanfer & R. Cudeck (Eds.), *Abilities, motivation, and methodology: The Minnesota Symposium on Learning and Individual Differences.* (pp. 435–474). Lawrence Erlbaum Associates, Inc.

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational psychologist*, *27*(1), 5-32.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and Control Strategies in Study-Time Allocation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(1), 204–221. https://doi.org/10.1037/0278-7393.26.1.204

Souvignier, E., Förster, N., Hebbecker, K., & Schütze, B. (2021). quop: An Effective Web-Based Approach to Monitor Student Learning Progress in Reading and Mathematics in Entire Classrooms. In S. Jornitz & A. Wilmers (Eds.), *International Perspectives on School Settings, Education Policy and Digital Strategies. A Transatlantic Discourse in Education Research* (pp. 283–298). Leverkusen: Budrich.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*(8), 795–819. https://doi.org/10.1002/pits.20113

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*(2), 331–347. https://doi.org/10.1037/a0034752

Subban, P. (2006). Differentiated instruction: A research basis. *International education journal*, *7*(7), 935-947.

Suzuki, Y., & Dekeyser, R. (2017). Exploratory research on second language practice distribution: An Aptitude × Treatment interaction. *Applied Psycholinguistics*, *38*(1), 27–56. https://doi.org/10.1017/S0142716416000084

Tobias, S. 1978. *Interaction between achievement and instructional method.* Paper presented at the meeting of the American Educational Research Association. March 1978, Toronto, Canada.

Tobias, S. (1989). Another look at research on the adaptation of instruction to student characteristics. *Educational Psychologist*, *24*(3), 213–227. https://doi.org/10.1207/s15326985ep2403_1

Tobias, S. (2010). The expertise reversal effect and aptitude treatment interaction research (Commentary). *Instructional Science*, *38*(3), 309–314. https://doi.org/10.1007/s11251-009-9103-z

Tobias, Sigmund. & Redfield, Robert. & City Univ. of New York, NY. (1980). Anxiety, Prior Achievement, and Instructional Support. [Washington, D.C.] : Distributed by ERIC Clearinghouse, https://eric.ed.gov/?id=ED190594

Vanlehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, *46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*(4), 932–945. https://doi.org/10.1037/a0031882

Waxman, H. C., Wang, M. C., Anderson, K. A., & Walberg, H. J. (1985). Adaptive Education and Student Outcomes: A Quantitative Synthesis. *The Journal of Educational Research, 78*(4), 228-236. https://doi.org/10.1080/00220671.1985.10885607

Wininger, S. R., Redifer, J. L., Norman, A. D., & Ryle, M. K. (2019). Prevalence of Learning Styles in Educational Psychology and Introduction to Education Textbooks: A Content Analysis. *Psychology Learning & Teaching, 18*(3), 221-243. https://doi.org/10.1177/1475725719830301

Xiao, Y., & Yang, M. (2019). Formative assessment and self-regulated learning: How formative assessment supports students' self-regulation in English language learning. *System*, *81*, 39–49. https://doi.org/10.1016/J.SYSTEM.2019.01.004

Yeh, S. S. (2010). Understanding and addressing the achievement gap through individualized instruction and formative assessment. *Assessment in Education: Principles, Policy & Practice, 17*(2), 169-182. https://doi.org/10.1080/09695941003694466

Zhang, D., & Zhang, L. J. (2019). Metacognition and Self-Regulated Learning (SRL) in Second/Foreign Language Teaching. *Second Handbook of English Language Teaching*, 883-897. https://doi.org/10.1007/978-3-030-02899-2_47

Ziegler, E., Edelsbrunner, P. A., & Stern, E. (2021). The benefit of combining teacher-direction with contrasted presentation of algebra principles. *European Journal of Psychology of Education, 36*(1), 187-218. https://doi.org/10.1007/s10212-020-00468-3

# Appendix

## A Original Manuscripts

### *Paper 1*

**Tetzlaff, L.,** Schmiedek, F., & Brod, G. (2021). Developing Personalized Education: A Dynamic Framework. *Educational Psychology Review*, *33*(3), 863-882.

# Developing Personalized Education: A Dynamic Framework

Leonard Tetzlaff, Florian Schmiedek, & Garvin Brod

Author Note

Leonard Tetzlaff, DIPF | Leibniz Institute for Research and Information in Education

Florian Schmiedek, DIPF | Leibniz Institute for Research and Information in Education, Goethe Universität Frankfurt

Garvin Brod, DIPF | Leibniz Institute for Research and Information in Education, Goethe Universität Frankfurt

Correspondence concerning this article should be addressed to Leonard Tetzlaff, DIPF | Leibniz Institute for Research and Information in Education,

Rostocker Straße 6, 60323 Frankfurt, email: tetzlaff@dipf.de

**Abstract**

Personalized education – the systematic adaptation of instruction to individual learners – has been a long-striven goal. We review research on personalized education that has been conducted in the laboratory, in the classroom, and in digital learning environments. Across all learning environments, we find that personalization is most successful when relevant learner characteristics are measured repeatedly during the learning process and when these data are used to adapt instruction in a systematic way. Building on these observations, we propose a novel, dynamic framework of personalization that conceptualizes learners as dynamic entities that change during and in interaction with the instructional process. As these dynamics manifest on different timescales, so do the opportunities for instructional adaptations – ranging from setting appropriate learning goals at the macroscale to reacting to affective-motivational fluctuations at the microscale. We argue that instructional design needs to take these dynamics into account in order to adapt to a specific learner at a specific point in time. Finally, we provide some examples of successful, dynamic adaptations and discuss future directions that arise from a dynamic conceptualization of personalization.

Keywords: Aptitude-Treatment-Interaction, Personalization, Individualization, Formative Assessment, Intelligent Tutoring Systems

## Declarations

**Funding**

Research was supported by the Stiftung Mercator GmbH. GB was supported by a Jacobs Foundation Research Fellowship.

**Conflicts of interest/Competing interests**

Not Applicable

**Ethics approval**

Not Applicable

**Consent to participate**

Not Applicable

**Consent for publication**

Not Applicable

**Availability of data and material**

Not Applicable

**Code availability**

Not Applicable

## Developing Personalized Education: A Dynamic Framework

The personalization of education has been a desired goal in educational science and practice for more than 200 years. Educators and policymakers alike are putting their hopes in personalization as a panacea for achievement gaps, lack of student motivation, and more effective instruction in general. Broadly construed, personalized education refers to the adaptation of instruction to a specific learner and is juxtaposed with "traditional" instruction that is targeted at entire groups of learners. By changing the mode, content, or rate of instruction in accordance with some characteristic of the learner, it is suggested that individual shortcomings of learners can be addressed and their resources leveraged (Dockterman, 2018).

A key argument for the efficacy of personalization can be drawn from empirical demonstrations that learning gains in one-on-one tutoring are up to two standard deviations higher than in conventional classroom instruction (Bloom, 1984). This phenomenon and the subsequent desire to scale up the relevant instructional components to larger groups of learners became known as the 2-sigma-problem in educational psychology (e.g., Barrows, Myers, Williams, & Moticka, 1986; Corbett, 2001). Although later studies reported less extreme effect size differences (e.g., Vanlehn, 2011), scaling up the benefits of one-on-one tutoring to larger groups of learners has remained one of the driving forces behind research on personalization.

Bloom (1984) explained the considerable effect of one-on-one tutoring in his studies by arguing that a personal tutor is better able to assess the individual characteristics of the specific learner and to select appropriate instructional methods and materials – such as identifying the zone of proximal development (Rieber & Carton, 1988) and choosing tasks that are located within it. Other benefits of tutoring include fluent adaptations of the instructional method during the tutoring process as well as dynamic reactions to fluctuations in affective or motivational states of the learner. (Lehman, Matthews, D'Mello, & Person, 2008)

More recently, Bloom's explanations have been supported by the emergence of Intelligent Tutoring Systems (ITS), which have been shown to greatly increase learning gains when compared to regular instruction (Ma, Nesbit, & Liu, 2014; Steenbergen-Hu & Cooper, 2014). ITS have two defining features: (1) student modelling, that is, the assessment of several specific learner characteristics through direct measures (e.g., correct or incorrect responses to tasks) or indirect measures (e.g., logfiles of clicking behavior) and (2) the subsequent adaptation of instruction. The success of these tutors can therefore be conceptualized as a success of personalized education. The benefits of personalization can not only be seen in digital environments, however. Even in a conventional classroom context, there is mounting evidence for increased learning gains through personalized instruction across a wide range of settings and subjects (Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; Connor et al., 2009; Jung, McMaster, Kunkel, Shin, & Stecker, 2018; Slavin & Karweit, 1985; Stecker, Fuchs, & Fuchs, 2005; Waxman et al., 1985).

This paper is not intended as a comprehensive review of the expansive literature on personalized education. Instead, we aim to scope out a new direction that adopts a dynamic, person-centered perspective on the subject while still maintaining a systematic and data-based approach. By taking intraindividual dynamics into account, it is possible to adapt instruction not only to a specific learner but also to that learner at a specific point in time. To achieve this, we first look at three different environments in which personalized education has been studied: laboratory research, digital learning environments, and traditional classroom environments. We then highlight commonalities that underlie effective personalization across all three environments – dynamic assessment and subsequent data-driven adaptations of instruction and/or assistance. We conclude that, for personalized education to be effective, dynamic student modeling is needed. A student model is considered dynamic if it accounts for potential

changes in relevant learner characteristics that may occur during the instructional process. Teaching agents need knowledge on which learner characteristics are relevant for the learning process at different timescales and on the different ways in which these can vary – between individuals, within individuals over time, and in response to interventions.

In the last part of this manuscript, we present a dynamic framework of personalized education that offers a systematic classification of different learner dynamics over different timescales, as well as a broad characterization of the corresponding instructional adaptations. Since our focus lies on the dynamic modelling of learner characteristics, we only briefly touch upon these adaptations by providing some promising examples as well as some references for further reading.

**What is Personalized Education?**

We define personalized education as the data-based adjustment of any aspect of instructional practice to relevant characteristics of a specific learner. Relevant learner characteristics are defined as all variables that explain (or are assumed to explain) variance in learning outcomes. By instruction, we mean any interaction between learning and teaching agent that has (or is assumed to have) direct or indirect relevance for the learning process.

In the personalized education literature, there are several related terms that are used sometimes interchangeably, sometimes carrying slightly different connotations. For the purpose of this article we are using ‘adaptive’ as an umbrella term for all educational approaches that adjust some aspect of instruction based on a measured or predicted characteristic of a learner or group of learners. We are using ‘personalization’ synonymous with ‘individualization’, meaning that the adjustment of instructional practice is targeted at a specific learner and thus implying some form of assessment or modelling. This is in contrast to ‘differentiation’, which we use for any practices that adjust instruction to different groups of learners. Figure 1 details the loop that we deem necessary for effective personalization.

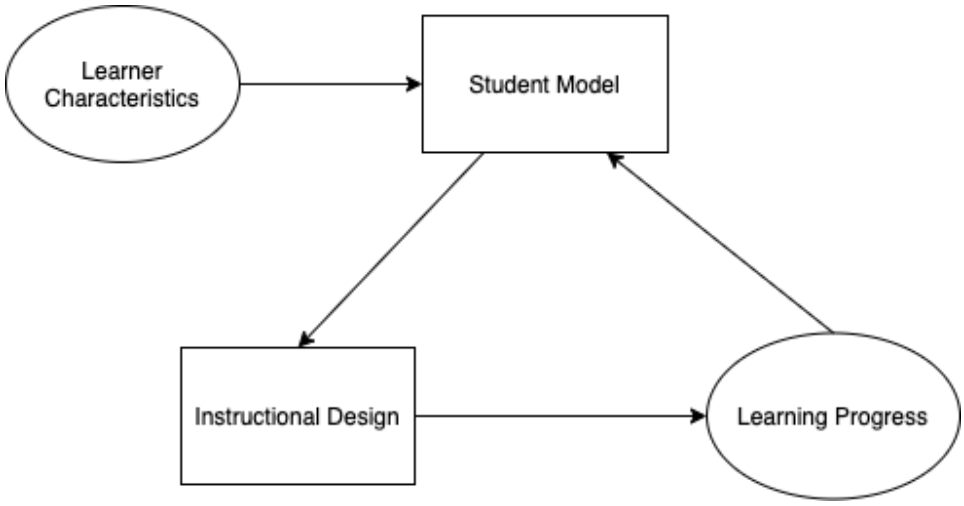



*Figure 1. General Personalization Loop. Ellipses correspond to the learning agent, boxes correspond to the teaching agent*

This loop consists of the following sequence of steps:

Step 1 – Initial Assessment of Learner Characteristics: Identifying those learner characteristics that are relevant for the specific learning process and assessing them in order to establish a student model

Step 2 – Instructional Design: Designing an instructional unit that forms or facilitates the next step towards the overarching learning goal.

Step 3 – Progress Assessment: Using the information from task performance and embedded assessment to update the student model based on the progress in the to-be-learned material.

These data- or system-driven personalization endeavors represent one end of a continuum. The other end of this continuum puts the focus on learner participation in goal setting and task selection, allowing learners to personalize their own learning path. The predominantly learner-driven approach is quite prevalent in educational science and teacher education (e.g., Crosby & Fremont, 1960) as well as in e-learning programs (McLoughlin & Lee, 2009). The assumption behind these learner-driven approaches is that learners will

generally know best what is best for them. Psychological research on metacognition, however, shows that this is not necessarily the case and learners do not always select the most appropriate tasks (Nugteren, Jarodzka, Kester, & Van Merriënboer, 2018; Son & Metcalfe, 2000). We posit that simply shifting control to the learner is not sufficient for personalized instruction. Rather, the amount of learner control should be carefully selected in accordance with the learning prerequisites as well as with the learning goals. For an exemplary model of such a dynamic allocation of control see (Corbalan, Kester, & Van Merriënboer, 2006).

## Existing Research on Personalized Education

In the following three sections, we will briefly review existing research in the field of personalized education. We will begin with research on aptitude-treatment interactions, which form the basis for effective personalization. We will then examine different approaches to personalize learning in a classroom setting, before moving on to personalized education in digital learning environments.

### Aptitude-Treatment Interactions

The main paradigm under which psychology has studied personalization is called aptitude-treatment interactions (ATI). The concept was established by Lee Cronbach, who saw in it the synthesis of correlational (aptitudes) and experimental (treatment) psychology (Cronbach, 1957). Using methods of correlational psychology, interindividual differences in relevant characteristics (aptitudes) are assessed and used to group people with similar values together. Then, relying on experimental psychology, people from these groups get randomly assigned to different treatments. If a disordinal interaction is found (Group A learns best under Treatment A, Group B under Treatment B), there is evidence for the efficacy of providing learners with different treatments, based on that particular aptitude. The existence of these interactions is a necessary prerequisite for any form of personalization to show direct effects on learning. Without the existence of ATIs, some learners may learn better than others and

some instructional parameters may foster learning better than others but there would be no advantage of adapting instruction to specific learners.

Over the following 40 years, a lot of research was carried out using this paradigm – with remarkably sparse robust results (see Tobias, 1989). While a few disordinal interactions between aptitude measures and different treatments have been found (see Cronbach, 1975), the vast majority of ATI studies found no or only ordinal interactions (both groups learn better under Treatment A, but the difference between treatments is bigger for one group). Cronbach and Snow's (1977) exhaustive review of the early literature on ATI studies concluded that "no aptitude by treatment interactions are so well confirmed that they can be used directly as guides to instruction“ (page 492). Several more recent reviews also reached the same conclusion, speculating on different reasons for this apparent failure, including a focus on laboratory experiments (Shapiro, 1975), factorially complex aptitude measures (Bracht, 1968), a focus on the surface structure of treatment (Tobias, 1989), and low specificity of to-be-learned content (Driscoll, 1987). Besides these mainly conceptual shortcomings, there exists also a series of methodological concerns that may lead to a reduced prevalence of demonstrable ATIs, chief among them a disregard for multilevel structure of data, a lack of statistical power (Preacher & Sterba, 2018), as well as a focus on linear modelling, which may lead to biased or false negative results when the true relationships between aptitudes and treatments are nonlinear (Bauer & Cai, 2009; Dumas, McNeish, & Greene, 2020).

Of note, none of the reviews on ATIs have reached the conclusion that they simply do not, or only in very rare circumstances, exist. The concept has such a high face validity that it seemed more reasonable to assume that researchers just had not yet looked in the right places (Tobias, 1989), or in the correct way (Shapiro, 1975). The demonstrated efficacy of data-based

individualization also strongly implies some kind of ATIs existing; no other mechanism has been proposed so far to be responsible for these effects.

A special case of ATI is the expertise reversal effect (Kalyuga, 2007). The expertise reversal effect is present if a certain instructional parameter leads to increased learning gain in novices, but decreased learning gains in experts. This effect is particularly interesting in the context of this paper as it highlights the need for a dynamic conceptualization of aptitudes. Instructional parameters that prove effective at the beginning of a learning process (low expertise) can actively impede learning as expertise grows. Even an intervention as simple as reading a short text can drastically alter the effectiveness of subsequent instruction (Rey & Fischer, 2013).

***Interim Conclusion: Aptitude-Treatment Interactions***

If ATIs exist, why are they so hard to find even when researchers are actively looking for them? We believe that the aptitude concept used in most ATI studies, which stems from differential/correlational psychology, does not suffice to answer questions about differential effectiveness of treatments. Instead, a dynamic perspective is needed for the following reason. By design, ATI research focuses on average differences between groups of students and from there tries to draw conclusions about the learning processes of individual students. Since learning processes are always intraindividual processes, trying to approach them by analyzing interindividual differences is suboptimal. Many different combinations of long-term (e.g., maturational and environmental) and short-term (e.g., affective-motivational) processes can lead to the same value on a scale of interindividual differences (Borsboom, Kievit, Cervone, & Hood, 2009) but can indicate completely different instructional practices (Bracht, 1970). Researchers usually operationalize aptitudes via single measurement points. Learners and their specific aptitudes vary considerably in their general stability, their developmental trajectories, and their responsiveness to instruction. An initial measure of, for example, metacognitive skills

can capture learners at the upper or lower end of their intraindividual distributions, at the beginning or the end of a developmental process, and directly before or after an intervention that completely changes the value. Research in developmental psychology, however, suggest that learners and their aptitudes are dynamic entities that a) change over time, b) are sensitive to different interventions, and c) fluctuate (for a similar distinction, see Nesselroade, 1991).

**Personalized Learning in Digital Learning Environments**

One of the most common forms of personalization in digital learning environments is the adaptation of instructional materials to fit the 'learning style' of the learner (Kumar & Ahuja, 2020; Truong, 2016; Yang, Hwang, & Yang, 2013). Despite its widespread prevalence (not just in e-learning) and some empirical studies reporting increased learning gains through consideration of the individual learning styles, the validity of the concept as well as the robustness of the evidence has been heavily disputed (Kirschner, 2017; Pashler, McDaniel, Rohrer, & Bjork, 2008). Other personalization strategies include adapting to the users 'intelligence profile', 'media preferences', prior knowledge, or motivation level. These adaptations are usually based on a single initial assessment of the characteristic in question which is then used to sort the learner in one of several discrete groups. (Essalmi, Ayed, Jemni, Graf, & Kinshuk, 2015).

Despite those personalization strategies, a meta-analysis on the effectiveness of e-learning programs for nurses found no benefit compared to regular instruction (Lahti, Hätönen, & Välimäki, 2014). Similarly, Sitzmann, Kraiger, Stewart, & Wisher (2006) found no advantage of e-learning over classroom instruction in their meta-analysis, as long as the same instructional methods were used in both conditions.

In contrast, the research tradition of intelligent tutoring systems (ITS), a subfield of e-learning, have taken a much more dynamic approach to personalization. ITS are, by definition,

computer programs that model learners' psychological states to provide personalized instruction (Ma et al., 2014). These so-called student models allow personalization over and above adjusting the difficulty of the next task based on the performance in the current one or assigning the user specific content based on static pretest measures. Several meta-analyses have shown the effectiveness of these systems across different domains and contexts (Corbett, 2001; Ma et al., 2014; Steenbergen-Hu & Cooper, 2014), leading to the conclusion that dynamic student modelling and subsequent adaptations are an effective mechanism for promoting learning gains.

While this line of research does not allow statements about isolated interaction effects of specific treatment variables with specific aptitudes, it does provide some empirical evidence regarding the efficacy of adapting to specific characteristics. Characteristics that have been successfully adapted to, over and above prior knowledge, include metacognitive skills (Azevedo, Witherspoon, Chauncey, Burkett, & Fike, 2009), current affect (D'Mello, Olney, Williams, & Hays, 2012; Lehman et al., 2013), and motivation (Walkington, 2013; for a comprehensive overview of different adaptations in ITS see Aleven, McLaughlin, Glenn, & Koedinger, 2017).

***Interim Conclusion: Personalization in Digital Learning Environments***

The discrepancy between the null effects found for many forms of e-learning and the convincing evidence for the efficacy of ITS further strengthens the point that the success of ITS can not just be traced back to them being computer-based (and thus flexible and delocalized). Instead, it seems plausible to conclude that they are caused by the dynamic assessment and subsequent instructional adaptations that set ITS apart from other forms of e-learning. While most e-learning systems claim some form of 'individualization' or 'personalization' of content based on some form of pretest, the lack of significant effects on

learning gains of these adaptations suggests that dynamic assessment likely is a necessary precursor for effective personalization.

To conclude, while digital learning environments offer the potential for new ways to personalize instruction, the empirical evidence indicates that just because something is personalized, it does not mean that it automatically fosters learning. Adapting to learner characteristics that are not strongly connected to learning processes (such as learning styles) or using static modelling as a basis for adaptations can be seen as potential culprits for ineffective personalization attempts. In contrast, using dynamic modelling to assess and adapt to relevant learner characteristics can lead to learning gains only rivalled by one-on-one human tutoring (Corbett, 2001; Vanlehn, 2011).

**Personalized Learning in the Classroom**

The most basic approach to data-based personalization in classroom contexts is ability grouping – the grouping of students with similar ability (usually measured once at the start of the program) either in different classes or within a class in order to present different materials or progress content at a different pace for the separate groups (Slavin, 1987). While still a far cry from true personalization, the appeal of these methods lies in their practicability. Administering a single test to measure ability in broad categories is already part of most school systems and providing specific instruction to 2-3 different groups of students is much less daunting a task than doing so for 20-30 individual students. In their field study on individualized mathematics instruction, (Slavin & Karweit, 1985) found a clear benefit of ability grouping vs. conventional whole-class teaching as well as a clear benefit of a completely personalized model (Team Assisted Individualization) vs. ability grouping on student achievement.

Formative Assessment, also known as learning progress assessment or curriculum-based measurement is the most widespread approach to systematically personalize education in classrooms. Black and William (2009, p. 5) have put forward the following definition of Formative Assessment: “Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited.” This stands in contrast to summative assessment, which is not meant to directly inform further instruction but rather provide a summary of the knowledge or skill level of the learner at the end of a predefined period (Harlen & James, 1997).

The concept of Formative Assessment originated in the field of special education, and was thought of as an advancement upon Bloom’s mastery learning (Deno, 1990; Fuchs, 2004; Wesson, King, & Deno, 1984). In the classical mastery learning approach (Bloom, 1968), students are repeatedly tested on the content they are currently trying to learn – upon reaching a certain proficiency they advance to the next content (and are tested on that). In the Formative Assessment approach, the students regularly complete parallel tests on the content they should have mastered by the end of the year/semester. This allows the teaching agent to continuously monitor progress on a single scale and to adapt the instruction in case of stagnation. In the beginning of the 21$^{st}$ century, this concept began to gain a lot more traction internationally and in general educational sciences and the evidence for its effectiveness grew (Black & Wiliam, 2009; Förster & Souvignier, 2014; Klauer, 2011; Stecker, Fuchs, & Fuchs, 2005; Waxman et al., 1985).

Even though different Formative Assessment procedures vary significantly in several parameters (type of feedback, learner or teacher driven, one- or multidimensional etc.), they all incorporate some dynamic assessment of learning progress on an individual basis and they all

seem to have at least some positive effects on learning, compared with a business-as-usual control group. This has been shown in several meta-analyses reporting effect sizes of $d = 0.32$ in regular classrooms (Kingston & Nash, 2011), and higher ones for students with special educational needs (Jung et al., 2018).

### *Interim Conclusion: Personalization in the Classroom*

The effectiveness of Formative Assessment procedures compared to regular classroom instruction also highlights the advantages of dynamic modelling through repeated measurements during the learning process. Usually, formative assessment is only used to track (multidimensional) domain knowledge, but using similar techniques to also measure progress in characteristics such as metacognitive skills or strategy knowledge to identify and address shortcomings could be a worthwhile endeavor.

A challenge that Formative Assessment poses to scientists researching personalized education is that the actual instructional adaptations are usually left up to the practitioners. Their dynamic nature makes it quite hard to reliably assess or prescribe them. While the above-mentioned success of these practices shows that many teachers are able to draw meaningful conclusions from the assessment data, the absence of information concerning the instructional adaptations still poses a significant obstacle to furthering our understanding about personalization in detail.

## Synthesis

An overall conclusion that can be drawn from the research discussed thus far is that personalization seems to be more successful when it takes the dynamic nature of learning processes into account. Dynamic means that the constituting factors of successful learning can change during and in interaction with the instructional process.

Evidence for this conclusion can be drawn from the surprising lack of clear ATIs using static aptitude measures as well as from the success stories of ITS and Formative Assessment, both of which use dynamic assessment procedures to create and update student models, allowing the teaching agents to continuously adapt their instruction to a developing learner. We argue that the success of these practices is a direct consequence of this dynamic approach to student modelling.

Generally speaking, a student model is any abstract representation of a learner that is being held by a teaching agent (Holt, Dubs, Jones, & Greer, 1994). These student models can be formal, such as the placement on a distribution of test scores, or informal, such as a teacher believing someone to be a fast learner, as well as high-level, such as an aggregated grade over a whole school year in a specific subject, or low-level, such as a specific mistake a student made twice in a row. Static student models get established once, usually to compare the student either to a comparable sample or to a specific criterium. Their underlying conceptualization is deterministic, that is, knowledge about the learning prerequisites as well as the specific instructional parameters is deemed sufficient to determine learning progress over a longer period of time. While we do not want to dispute that this is theoretically possible when knowing *all* relevant prerequisites, it does not seem to be a realistic proposition. Dynamic modeling deals with this lack of information by leaving room for differing individual trajectories. Repeated measurements can be used to correct mistaken assumptions and better determine future learning.

Not only do static characterizations potentially lead to an aptitude-treatment mismatch, they can also deprive the learner of the opportunity to acquire the lacking aptitudes. An example of this can be seen in a study in which learning gains increased for low-engagement students when presented with material which did *not* correspond to their preferred learning style (Kelly & Tangney, 2006). Another (hypothetical) example would be a teaching agent

measuring the metacognitive skills of a learner and concluding from a low value that the learner needs a lot of explicit feedback and guidance in task selection. This in turn drastically reduces the learning opportunities for the student to actually improve in judging their own learning and selecting appropriate tasks – a phenomenon known as part of the assistance dilemma (see Koedinger, Pavlik, Mclaren, & Aleven, 2008).

A dynamic modeling approach is not a new invention in research on learning. Developmental psychology has been using so-called microgenetic methods (consisting of highly frequent measurements during times of interesting developmental processes) since the 1920s to better understand the development of cognitive competencies in early childhood (Catán, 1986). Recently, there has also been a rise of studies employing measurement-intensive longitudinal designs and recognizing the potential of within-person analyses as well as dynamic measurement models in educational research (Dumas et al., 2020; Murayama et al., 2017). Even for presumably stable traits, such as intelligence, dynamic testing procedures have been shown to produce educationally relevant information beyond that produced by static tests (Resing, de Jong, Bosma, & Tunteler, 2009; Vogelaar, Resing, & Stad, 2020). The underlying assumption behind dynamic assessment is that learners change during and in interaction with the instructional process. If the characteristics of learners were stable entities that completely predicted learning outcomes under specific treatment conditions (as assumed in the early days of ATI research), there would be no need for dynamic modeling and thus no measurable advantage in employing it. Since the evidence clearly points to increased learning gains as a result of dynamic modeling (and subsequent adaptations), we will now turn to the different

ways learners and their characteristics can change, as well as the educational relevance of these changes.

## Learner Dynamics

Leaning on the conceptualization of Hertzog and Nesselroade, 2003, we propose that there are three main ways in which relevant characteristics of a learner can vary: along an

individual developmental trajectory, in response to an intervention, as well as in short-term fluctuations.

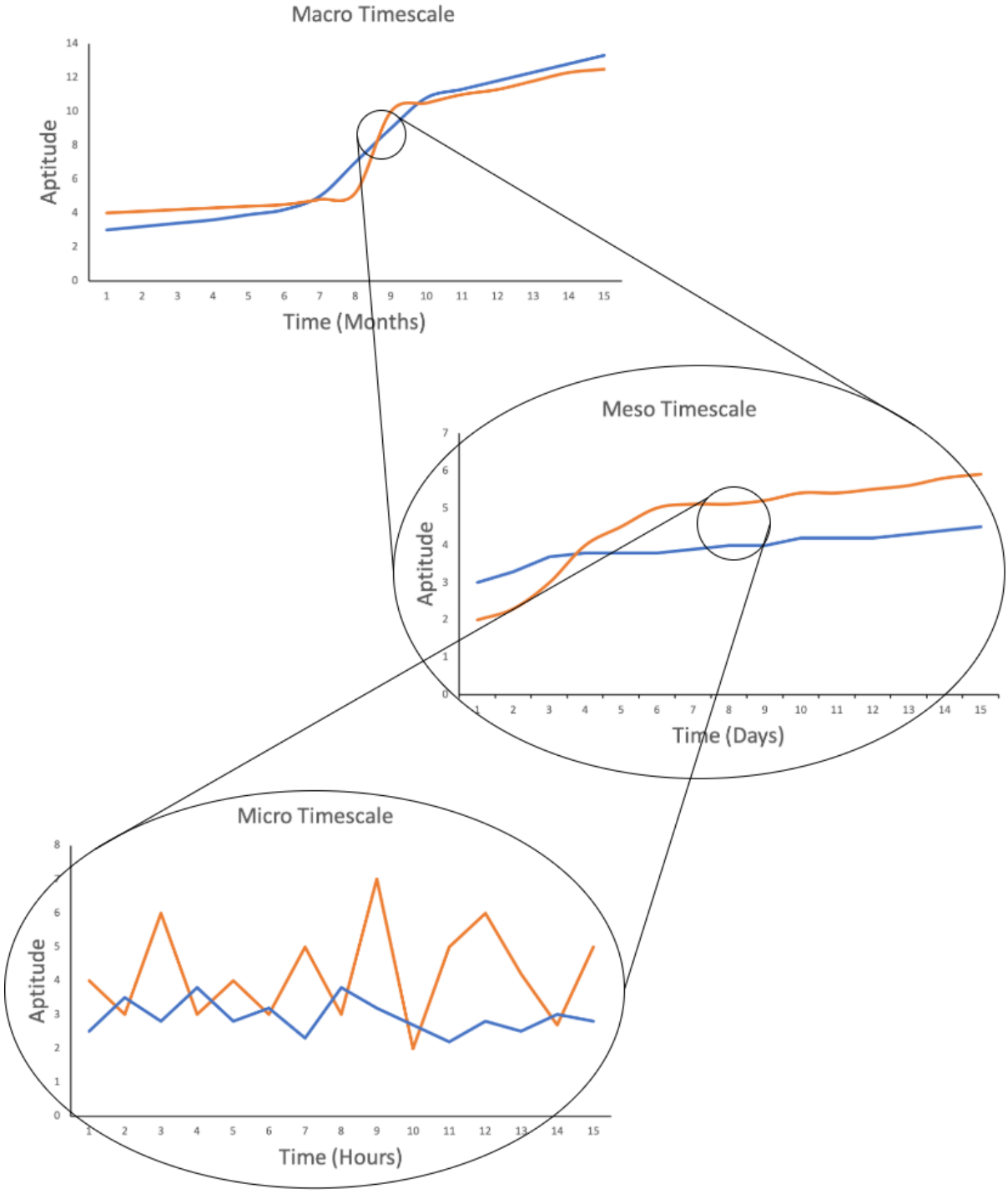



Figure 2: Development of two fictional learners in a single aptitude over the course of months, days and hours. Figure 2: Development of two fictional learners in a single aptitude over the course of months, days and hours.

Figure 2 depicts the fictional change of a single aptitude of two learners over time. This aptitude could, for example, be learners' metacognitive control skills. In the uppermost graph we can see the development over several months. Both learners show a gradual increase that changes in steepness over time with certain periods of more pronounced development. We can also see that there are clear differences between the learners in the measured level of the aptitude at most time points.

As soon as we zoom in on a particular point in time and look at the development from day to day, we can see that the value for the aptitude of both learners also shows a systematic trend – this can be caused by instructional input or some other intervention, such as changes in the environment of the learners. Learners react quite differently to the same instruction and already simple interventions can have far-reaching consequences for the development of specific aptitudes. These changes operate on a much smaller timescale than the developmental processes outlined above (and need to be accompanied by regular assessment to be correctly modeled).

By zooming in even further and looking at the processes within a specific day, we can see that the value for the aptitude of both learners fluctuates. Even though one of them shows a higher average performance, on specific tasks he or she may perform far below the other. We can also see that the amplitude and frequency of these fluctuations differs between people. A high amplitude in intraindividual fluctuations of relevant aptitudes can lead to very unstable performance patterns and indicates a need for instructional adaptations.

As can be seen in Figure 2, obtaining an aptitude measure at a single point in time may be influenced by all three dynamics, making it difficult to infer appropriate instructional adaptations. By using repeated measures at different timescales, the teaching agent can identify if a performance is typical, if an intervention was successful for a specific learner, or if a learner

needs additional assistance on a specific day. Particularly if dynamics at the different time levels are nonlinear (as in Figure 2), a sufficiently dense temporal resolution of measurements is necessary to capture them. In the following sections, we will have a closer look at how knowledge about learner dynamics on different timescales can be used to inform instructional decision making.

**Learner Dynamics on the Macroscale**

We define the macroscale as the timescale of months to years. The main driver behind learner dynamics on the macroscale are developmental processes. Developmental processes are changes in relevant learner characteristics that are part of the regular development of students. These can be caused by maturation of brain structures, common environmental influences (such as the onset of schooling) as well as potential interactions between them. The attainment of mastery in a certain domain can also be conceptualized as a developmental process. These developmental trajectories differ from person to person (and from characteristic to characteristic) in their intercepts, slopes, and general shapes.

The performance on working memory tasks is a prime example of a developmental trend. It is increasing rapidly roughly up to the age of nine for simple tasks and roughly up to the age of thirteen for complex tasks (Luciana, Conklin, Hooper, & Yarger, 2005). Other characteristics show flatter developmental trajectories. This is the case for most affective-motivational factors, which remain relatively stable across the lifespan despite showing remarkable short-term variability (e.g., Röcke & Brose, 2013).

In a learning context, the macroscale is the scale of higher-order goals, such as mastery and skill acquisition. The most obvious example of instructional decision making on the macro scale is the grade-based school system. In most educational systems, students get sorted into groups according to their age, which then get assigned to specific curricula that are assumed to

be suitable for that specific age group. The underlying theory behind this grouping is that most relevant differences between learners are developmental differences and that the shape of the trajectories is relative consistent across learners. Some assumed cases of accelerated or protracted development can easily be addressed by assigning children to slightly higher or lower age groups (Dockterman, 2018). The decision as to which skill or content to master is often out of control of the single teaching agent, but the specific individual learning goal and the optimal learning path towards that goal still need to be determined.

Most digital learning environments also try to guide learners to a specific, preset learning goal and only come into play after the to-be-mastered content has been selected.
In laboratory settings, there also exists evidence for differential effectiveness of treatments based on the age of the learner (see Breitwieser & Brod, 2020).

**Learner Dynamics on the Mesoscale**

We define the mesoscale as the timescale of days to weeks. The main driver behind learner dynamics on the mesoscale are intervention-induced changes. Intervention-induced changes describe any changes in relevant learner characteristics that result directly from an intervention. Under a broad definition, every instructional unit can be conceptualized as an intervention intended to modify the domain knowledge of the learner. In a more specific sense, the fact that some characteristics respond to targeted small-scale interventions opens up leverage points for teaching agents. Instead of adapting instruction to a specific characteristic, teaching agents can chose to modify it to have a better basis for subsequent instruction. A prime example of a characteristic that shows strong intervention-induced changes is the strategy knowledge of learners (e.g. Ryan, Short, & Weed, 2008). With short strategy trainings, learners can expand their repertoire of available learning strategies, which can lead to increased learning gains at the domain level.

In a learning context, the mesoscale is the scale of instructional units – bundles of tasks, explanations, examples etc. that can be processed in one session. Each instructional unit should present the next logical step towards the overarching learning goal and the difficulty should be adapted to the learning progress of the particular student. If a particular skill or knowledge component that would be required to proceed towards the learning goal is found to be missing, an instructional unit targeting that component should be presented.

Most ITS track multidimensional domain knowledge in order to generate appropriate instructional units (Nwana, 1990), but there are also several examples of small-scale interventions that are targeted at other characteristics that are identified as relevant for learning, such as metacognitive skills (Aleven, Mclaren, Roll, & Koedinger, 2006; D'Mello, Olney, Williams, & Hays, 2012) or epistemic emotions (Lehman et al., 2013). Formative assessment is also primarily operating on the mesoscale – the learning progress caused by the previous instruction gets measured in order to better inform subsequent instruction. This includes simple adaptations of difficulty as well as addressing specific gaps in knowledge or skills of individual learners. Finally, there is a long tradition of laboratory research showing the potential of utilizing the malleability of characteristics such as metacognition (Eslami Sharbabaki H, 2013) or strategy knowledge (Ryan et al., 2008) to increase domain-level learning gains.

**Learner Dynamics on the Microscale**

We define the microscale as the timescale of minutes to hours. The main driver behind learner dynamics on the microscale are short-term fluctuations in relevant characteristics. An obvious example of a characteristic fluctuating in value is the affective state of a learner. The way we feel can change from moment to moment. But even characteristics that are traditionally assumed to be stable traits, such as working memory capacity, have been shown to manifest substantial intraindividual variance, not just from day to day but even from moment to moment

(Dirk & Schmiedek, 2016) These fluctuations happen over larger timescales as well, but their relevance for educational decision making mainly lies in the microscale.

This relevance is partially shown in classroom education, where the concept of assessing and modifying students affective and motivational states on a day-to-day (or even moment-to-moment) basis forms part of what has been called the „supportive climate“ dimension of good teaching (Fauth, Decristan, Rieser, Klieme, & Büttner, 2014). Teaching that fosters a supportive climate has been linked to increased student engagement and achievement (Reyes, Brackett, Rivers, White, & Salovey, 2012).

There is also a growing base of research attempting to automate affect detection in classrooms via facial recognition systems (Bosch et al., 2016; Dragon et al., 2008). Studies on human one-on-one tutoring have likewise shown that expert tutors monitor the affective states of their tutees and engage in pedagogical moves such as off-topic conversation or positive feedback to counteract significant negative affect (Lehman et al., 2008).

Additionally, there are several examples of ITS fine-tuning some part of their content on a moment-by-moment basis, based on intraindividual fluctuations in affective, cognitive, or process variables. GazeTutor is using eye-tracking to detect boredom and disengagement in learners and tries to reengage them via dialog/animation and has been shown to increase learning gains compared to an equivalent system without affect modeling (D’Mello et al., 2012). Help Tutor und Meta-tutor are tracking difficulties in metacognitive monitoring/control that the learner exhibits (such as inefficient help seeking) and offering prompts aimed at improving these behaviors (Aleven, Roll, McLaren, & Koedinger, 2016; Azevedo et al., 2009). Most ITS are also providing feedback during task processing that adapts to the specific errors and/or the solution path the student has chosen (Koedinger, Brunskill, Baker, McLaughlin, & Stamper, 2013; Vanlehn, 2011).

## A Dynamic Framework of Personalized Education

Taking into account these different learner dynamics and their relevance for learning processes on the different timescales, Figure 3 provides an updated version of Figure 1 and highlights the relevant instructional decision-making processes and opportunities for adaptations at each of the different timescales.

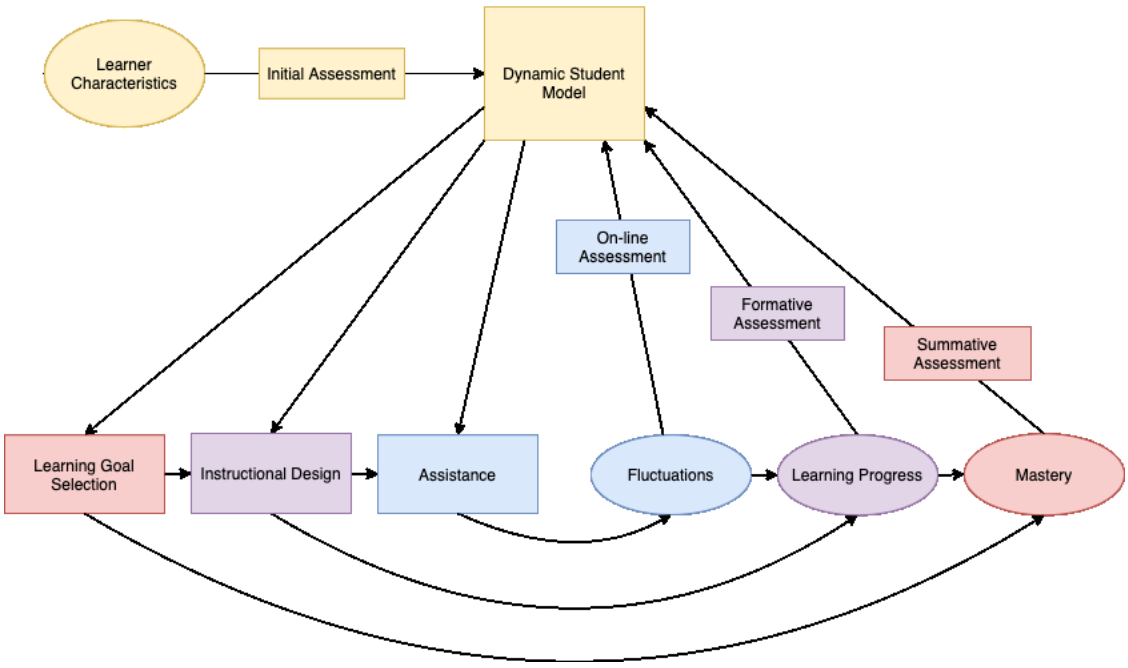



Figure 3: A dynamic framework of personalization on three different timescales. Ellipses correspond to the learning agent, boxes to the teaching agent.

In order for personalized learning to be effective, the general learning prerequisites (which are influenced by the individual developmental trajectory and the current age) over all relevant characteristics should be assessed as they inform the instructional decision making on

all timescales. Information from this initial assessment can be supplemented and adjusted by repeated measurements throughout the learning process.

The elements colored in red show the macroscale of personalization. The main instructional decision to be made on that scale is the selection of an appropriate higher-order learning goal. Progress towards mastery of that goal can be continuously measured and in case of stagnation, instructional change can be implemented. This cycle of assessing the learner prerequisites, setting a reasonable learning goal, and employing summative assessment practices to check whether mastery was achieved (which influences the learning goals for the next cycle) is the backbone of personalized instruction at the macroscale. As described above, decisions on the macro level are often predetermined by context and thus difficult to truly personalize. Nevertheless, adapting to student characteristics on the macroscale was historically one of the first steps from homogenous ability grouping towards truly personalized instruction (Dockterman, 2018; Lee & Park, 2008)

Designing an instructional unit that falls in the zone of proximal development of the learner and fits their individual learning prerequisites is the main instructional adaptation of relevance on the mesoscale (colored in purple). As posited in Bloom's mastery learning approach (Bloom, 1968), the teaching agent needs to measure the success of the instructional unit before proceeding with the next one. Upon completion of the unit, there needs to be some assessment of the learning gains and subsequent selection/design of the next unit (located in the zone of proximal development and presenting a logical next step on the way to the high-level learning goal). This cycle of presenting an instructional unit, measuring the learning progress with formative assessment procedures, and then using that information to design the next instructional unit is the main way to personalize instruction at the mesoscale. An integral part of designing personalized instructional units is "efficient" task selection. Personalizing task selection based on predicted efficiency as well as learner preference has been shown to

increase training and transfer performance, respectively, when compared to yoked control groups (Salden, Paas, & Van Merrienboer, 2006).

If we take a closer look at the processes within an instructional unit (colored in blue), we can see that the main way that a teaching agent can adapt on this scale is by selectively giving or withholding assistance. We define assistance as any action a teaching agent takes that facilitates progress in the current task (e.g. error specific feedback, scaffolding, hints). It is important to note here that quicker or easier task progress does not necessarily translate to increased learning gains. Studies on the assistance dilemma (Koedinger et al., 2008) generally imply an inverted-U shaped relationship between task difficulty (after assistance) and learning progress, where too little assistance can leave the learner unable to make progress on the task, and too much assistance does not require the learner to engage in the cognitive processes necessary for deep processing (and thus robust learning). Studies on the expertise reversal effect (Kalyuga, Ayres, Chandler, & Sweller, 2003) also imply that, generally speaking, extensive assistance should be provided if a task is new and difficult and then should be gradually reduced as the learner gains expertise in that specific task. Besides this general trend, assistance should also be given or withheld reactively, depending on fluctuations in relevant characteristics. This means that if a learner is experiencing frustration, it might be advisable to increase the amount of assistance for that specific task-step, regardless of the general amount of expertise displayed. Assessing fluctuations in task performance or affective-motivational factors 'on-line' (parallel to the task progress) and reacting by giving or withholding assistance is the main personalization lever at the microscale.

These adaptations of provided assistance can take many different forms, ranging from affective-motivational (e.g. D'Mello, Olney, Williams, & Hays, 2012) over metacognitive support (e.g. Azevedo, Witherspoon, Chauncey, Burkett, & Fike, 2009) to the provision of hints

or error specific feedback (Koedinger, Brunskill, Baker, McLaughlin, & Stamper, 2013; Vanlehn, 2011).

## Conclusion

This article summarized the key findings from three mostly distinct research traditions on personalized education, synthesizing them into a comprehensive framework of personalized instruction and highlighting the need for dynamic assessment. While there are examples of successful personalization based on relatively stable pretest measures, empirical evidence as well as conceptual considerations strongly point towards an advantage of dynamic modelling, at least for those characteristics that show substantial intraindividual variance.

Assessing such characteristics at a high frequency throughout the learning process provides a variety of relevant information. It allows to separate individual characteristics at the macro, meso, and micro levels. This way, estimates of presumably stable trait characteristics (e.g., aptitudes) may be measured with increasing precision as more observations are collected. Also, individual differences in characteristics of observed learning curves, like learning rates or asymptotes, may be parameterized, estimated, and used as prognostic information for further learning processes that follow. Furthermore, the amount of intraindividual variability around average levels or trends may provide useful information. For example, sustained strong variability in task performance can give hints to instructors that the performance bottleneck lies in a highly volatile characteristic, such as affect, motivation, or metacognitive control, rather than in a stable (or monotonously increasing) characteristic such as domain knowledge. Finally, information on how different relevant variables that show such variation are coupled (i.e., correlated at the within-person level) within learners across time may be of diagnostic value. For example, Neubauer, Dirk, & Schmiedek (2019) report that within-child fluctuations (within and across days) in working memory performance are coupled with different dimensions of affect for different groups of children. Inferring such learner characteristics directly from

process data may aid the on-line adaptation of learning circumstances to individual learners' needs.

Other fields already lead the way towards dynamic modeling. In the field of clinical psychology there has been a similar push towards dynamic intraindividual patient models instead of basing personalization attempts on interindividual difference scores (Fisher & Boswell, 2016). These allow a much better fit of the treatment to the needs of the patient, as well as an easier adaptation of treatment parameters to changes in the process. We argue that a dynamic conceptualization is also needed to bring the science of personalized education (and ATI) forward.

This dynamic conceptualization undoubtedly brings with it an additional load for teaching agents. They not only need to regularly assess relevant parameters but they also have to use this information to inform subsequent instructional decisions. This load can be partially constrained by knowledge about which characteristics can be reasonably expected to vary over which timescale and the educational relevance of this variance. The presented framework serves as a starting point for such considerations. By mapping out the decision space for teaching agents, we identified relevant kinds of learner parameters on each timescale and provided some rough classification of the different levels of instructional practice that can be adapted: goal setting, design of instructional units/task selection, and assistance. The proposed framework further constrains the selection of both learner parameters and instructional parameters to those that are actually relevant on the specific timescale. It also provides a frame of reference for the localization of future research questions regarding personalized education by systematically differentiating between different kinds of learner dynamics, the learner characteristics they apply to, as well as the instructional levers that can be manipulated.

There has been substantial progress in research on personalized education in recent years, not just towards more precise measurement and conceptualization of aptitudes, but also towards a systematic classification of instructional adaptations. Nevertheless, we are still a long way off from being able to reliably describe adaptations at different timescales based on learner characteristics. In most Formative Assessment studies, the instructional adaptations are left up to the practitioners, providing almost no mechanistical information regarding causes of the observed benefit. Most classical ATI studies only define the treatment in very broad categories (learner vs. teacher driven, high vs. low structure material) and thus fail to account for the complex nature of instructional practice. Most ITS studies are designed to only evaluate a complete 'package' of adaptations (e.g., a system that tracks and interacts with affect vs. one that does not), providing evidence for or against the usage of that system but containing little information about specific adaptations. Future research needs to better isolate specific treatment variables in order to study their effects on specific learners at specific points in the learning process. Only then can we move to a truly evidence-based practice of personalized education, be it in the classroom, the laboratory, or in a digital learning environment.

## References


Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2017). Instruction based on adaptive learning technologies. *Handbook of Research on Learning and Instruction*, 522–560. https://doi.org/10.4324/9781315736419.ch24

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help Helps, but only so Much: Research on Help Seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education*, *26*, 205–223. https://doi.org/10.1007/s40593-015-0089-1

Azevedo, R., Witherspoon, A., Chauncey, A., Burkett, C., & Fike, A. (2009). MetaTutor: A MetaCognitive tool for enhancing self-regulated learning. *AAAI Fall Symposium - Technical Report*, *FS-09-02*, 14–19.

Barrows, H. S., Myers, A., Williams, R. G., & Moticka, E. J. (1986). Large group problem-based learning: A possible solution for the "2 sigma problem." *Medical Teacher*, *8*, 325–331. https://doi.org/10.3109/01421598609028991

Bauer, D. J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics*, *34*, 97–114. https://doi.org/10.3102/1076998607310504

Black, P., & Wiliam, D. (2009). *Developing the theory of formative assessment*. 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bloom, B. (1968). Learning for mastery. *Evaluation Comment*, *1*, 1–12. https://doi.org/10.1021/ed063p318

Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, *13*, 4–16. https://doi.org/10.3102/0013189X013006004

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In *Dynamic Process Methodology in the Social and Developmental Sciences* (pp. 67–97). https://doi.org/10.1007/978-0-387-95922-1_4

Bosch, N., D'Mello, S. K., Baker, R. S., Ocumpaugh, J., Shute, V., Ventura, M., … Zhao, W. (2016). Detecting student emotions in computer-enabled classrooms. *IJCAI International Joint Conference on Artificial Intelligence*, *2016-Januar*, 4125–4129.

Bracht, G. H. (1970). Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, *40*, 627–645. https://doi.org/10.3102/00346543040005627

Breitwieser, J., & Brod, G. (2020). Cognitive Prerequisites for Generative Learning: Why Some Learning Strategies Are More Effective Than Others. *Child Development*, cdev.13393. https://doi.org/10.1111/cdev.13393

Catán, L. (1986). The Dynamic Display of Process: Historical Development and Contemporary Uses of the Microgenetic Method. *Human Development*, *29*, 252–263. https://doi.org/10.1159/000273062

Connor, C. M. D., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science*, *315*, 464–465. https://doi.org/10.1126/science.1134513

Connor, C. M., Piasta, S. B., Glasney, S., Schatschneider, C., Fishman, B. J., Underwood, P. S., & Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child-by-instruction interactions on students' literacy. *Child Development*, *80*, 77–100.

Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2006). Towards a personalized task selection model with shared instructional control. *Instructional Science*, *34*, 399–422. https://doi.org/10.1007/s11251-005-5774-2

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2109*, 137–147. https://doi.org/10.1007/3-540-44566-8_14

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684. https://doi.org/10.1037/h0043943

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116–127. https://doi.org/10.1037/h0076829

Crosby, G., & Fremont, H. (1960). Individualized algebra. *The Mathematics Teacher*, *53*, 109–112. https://doi.org/10.2307/27956078

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human Computer Studies*, *70*, 377–398. https://doi.org/10.1016/j.ijhcs.2012.01.004

Deno, S. L. (1990). Individual Differences and Individual Difference. *The Journal of Special Education*, *24*, 160–173. https://doi.org/10.1177/002246699002400205

Dirk, J., & Schmiedek, F. (2016). Fluctuations in elementary school children's working memory performance in the school context. *Journal of Educational Psychology*, *108*,

722–739. https://doi.org/10.1037/edu0000076

Dockterman, D. (2018). Insights from 200+ years of personalized learning. *Npj Science of Learning*, *3*, 1–6. https://doi.org/10.1038/s41539-018-0033-x

Dragon, T., Arroyo, I., Woolf, B. P., Burleson, W., el Kaliouby, R., & Eydgahi, H. (2008). Viewing Student Affect and Learning through Classroom Observation and Physical Sensors. In *LNCS* (Vol. 5091, pp. 29–39). https://doi.org/10.1007/978-3-540-69132-7_8

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical–psychometric paradigm for modern educational psychology. *Educational Psychologist*, *55*, 1–18. https://doi.org/10.1080/00461520.2020.1744150

Eslami Sharbabaki H, H. V. (2013). The Effect of Metacognitive Strategy Training on Social Skills and Problem - Solving Performance. *Journal of Psychology & Psychotherapy*, *03*, 4. https://doi.org/10.4172/2161-0487.1000121

Essalmi, F., Ayed, L. J. Ben, Jemni, M., Graf, S., & Kinshuk. (2015). Generalized metrics for the analysis of E-learning personalization strategies. *Computers in Human Behavior*, *48*, 310–322. https://doi.org/10.1016/j.chb.2014.12.050

Fisher, A. J., & Boswell, J. F. (2016). Enhancing the Personalization of Psychotherapy With Dynamic Assessment and Modeling. *Assessment*, *23*, 496–506. https://doi.org/10.1177/1073191116638735

Förster, N., & Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learning and Instruction*, *32*, 91–100. https://doi.org/10.1016/j.learninstruc.2014.02.002

Fuchs, L. S. (2004). The Past, Present, and Future of Curriculum-Based Measurement Research. *School Psychology Review*, *33*, 188–192.

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *International Journal of Phytoremediation*, *21*, 365–379. https://doi.org/10.1080/0969594970040304

Hertzog, C., & Nesselroade, J. R. (2003). Assessing Psychological Change in Adulthood: An Overview of Methodological Issues. *Psychology and Aging*, *18*, 639–657. https://doi.org/10.1037/0882-7974.18.4.639

Holt, P., Dubs, S., Jones, M., & Greer, J. (1994). The State of Student Modelling. In *Student Modelling: The Key to Individualized Knowledge-Based Instruction* (pp. 3–35). https://doi.org/10.1007/978-3-662-03037-0_1

Jung, P.-G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of Data-Based Individualization for Students with Intensive Learning Needs: A Meta-Analysis. *Learning Disabilities Research & Practice*, *33*, 144–155. https://doi.org/10.1111/ldrp.12172

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*, 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, Vol. 38, pp. 23–31. https://doi.org/10.1207/S15326985EP3801_4

Kelly, D., & Tangney, B. (2006). Adapting to intelligence profile in an adaptive educational system. *Interacting with Computers*, *18*, 385–409. https://doi.org/10.1016/j.intcom.2005.11.009

Kingston, N., & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice*, *30*, 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers and Education*, *106*, 166–171. https://doi.org/10.1016/j.compedu.2016.12.006

Klauer, K. J. (2011). Lernverlaufsdiagnostik – Konzept, Schwierigkeiten und Möglichkeiten. *Empirische Sonderpädagogik*, 207–224.

Koedinger, K. R., Brunskill, E., Baker, R. S. J. D., McLaughlin, E. A., & Stamper, J. (2013). New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, *34*, 37–41. https://doi.org/10.1609/aimag.v34i3.2484

Koedinger, K. R., Pavlik, P., McLaren, B. M., & Aleven, V. (2008). Is It Better to Give Than to Receive? The Assistance Dilemma as a Fundamental Unsolved Problem in the Cognitive Science of Learning and Instruction. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2155–2160.

Kumar, A., & Ahuja, N. J. (2020). An adaptive framework of learner model using learner characteristics for intelligent tutoring systems. *Advances in Intelligent Systems and Computing*, *989*, 425–433. https://doi.org/10.1007/978-981-13-8618-3_45

Lahti, M., Hätönen, H., & Välimäki, M. (2014, January). Impact of e-learning on nurses' and student nurses knowledge, skills, and satisfaction: A systematic review and meta-analysis. *International Journal of Nursing Studies*, Vol. 51, pp. 136–149. https://doi.org/10.1016/j.ijnurstu.2012.12.017

Lehman, B., Matthews, M., D'Mello, S., & Person, N. (2008). What Are You Feeling? Investigating Student Affective States During Expert Human Tutoring Sessions. In *Intelligent Tutoring Systems* (pp. 50–59). https://doi.org/10.1007/978-3-540-69132-7_10

Lehman, B., Mello, S. D. ', Strain, A., Mills, C., Gross, M., Dobbins, A., … Graesser, A. (2013). Inducing and Tracking Confusion with Contradictions during Complex Learning. *International Journal of Artificial Intelligence in Education*, *22*, 85–105. https://doi.org/10.3233/JAI-130025

Luciana, M., Conklin, H. M., Hooper, C. J., & Yarger, R. S. (2005). The Development of Nonverbal Working Memory and Executive Control Processes in Adolescents. *Child Development*, *76*, 697–712. https://doi.org/10.1111/j.1467-8624.2005.00872.x

Ma, W., Nesbit, J. C., & Liu, Q. (2014). Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis. *Journal of Educational Psychology*, *106*, 901–918. https://doi.org/10.1037/a0037123.supp

McLoughlin, C., & Lee, M. J. W. (2009). Personalised learning spaces and self-regulated learning: Global examples of effective pedagogy. *ASCILITE 2009 - The Australasian Society for Computers in Learning in Tertiary Education*, 639–645.

Murayama, K., Goetz, T., Malmberg, L.-E., Pekrun, R., Tanaka, A., & Martin, A.-J. (2017). Within-person analysis in educational psychology: Importance and illustrations. In *British Journal of Educational Psychology Monograph Series II: Psychological Aspects of Education --- Current Trends: The role of competence beliefs in teaching and learning* (Vol. 12, pp. 71–87).

Nesselroade, J. R. (1991). The warp and the woof of the developmental fabric. In *R. M. Downs, L. S. Liben, & D. S. Palermo (Eds.), Visions of aesthetics, the environment & development: The legacy of Joachim F. Wohlwill* (pp. 213–240).

Neubauer, A. B., Dirk, J., & Schmiedek, F. (2019). Momentary working memory performance is coupled with different dimensions of affect for different children: A mixture model analysis of ambulatory assessment data. *Developmental Psychology*, *55*, 754–766. https://doi.org/10.1037/dev0000668

Nugteren, M. L., Jarodzka, H., Kester, L., & Van Merriënboer, J. J. G. (2018). Self-regulation of secondary school students: self-assessments are inaccurate and insufficiently used for learning-task selection. *Instructional Science*, *46*, 357–381. https://doi.org/10.1007/s11251-018-9448-2

Nwana, H. S. (1990). Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, *4*, 251–277. https://doi.org/10.1007/BF00168958

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles concepts and evidence. *Psychological Science in the Public Interest, Supplement*, *9*, 105–119. https://doi.org/10.1111/j.1539-6053.2009.01038.x

Resing, W. C. M., de Jong, F. M., Bosma, T., & Tunteler, E. (2009). Learning During Dynamic Testing: Variability in Strategy Use by Indigenous and Ethnic Minority Children. *Journal of Cognitive Education and Psychology*, *8*, 22–37. https://doi.org/10.1891/1945-8959.8.1.22

Rey, G. D., & Fischer, A. (2013). The expertise reversal effect concerning instructional explanations. *Instructional Science*, *41*, 407–429. https://doi.org/10.1007/s11251-012-9237-2

Rieber, R. W., & Carton, A. S. (1988). The Collected Works of L. S. Vygotsky. Boston, MA: Springer

Röcke, C., & Brose, A. (2013). Intraindividual Variability and Stability of Affect and Well-Being. *GeroPsych*, *26*, 185–199. https://doi.org/10.1024/1662-9647/a000094

Ryan, E. B., Short, E. J., & Weed, K. A. (2008). The Role of Cognitive Strategy Training in Improving the Academic Performance of Learning Disabled Children. *Journal of Learning Disabilities*, *19*, 521–529. https://doi.org/10.1177/002221948601900902

Salden, R. J. C. M., Paas, F., & Van Merrienboer, J. J. G. (2006). Personalised Adaptive Task Selection in Air Traffic Control. *Learning and Instruction*, *16*, 350–362. https://doi.org/10.1016/j.learninstruc.2006.07.007

Slavin, R. E. (1987). Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis. *Review of Educational Research*, *57*, 293–336. https://doi.org/10.3102/00346543057003293

Slavin, R. E., & Karweit, N. L. (1985). Effects of Whole Class, Ability Grouped, and Individualized Instruction on Mathematics Achievement. In *American Educational Research Journal Fall* (Vol. 22). http://journals.sagepub.com/doi/pdf/10.3102/00028312022003351

Son, L. K., & Metcalfe, J. (2000). Metacognitive and Control Strategies in Study-Time Allocation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*, 204–221. https://doi.org/10.1037/0278-7393.26.1.204

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

819. https://doi.org/10.1002/pits.20113

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*, 331–347. https://doi.org/10.1037/a0034752

Suzmann, T., Kraiger, K., Stewart, D., & Wisher, R. (2006). The comparative effectiveness of web-based and classroom instruction: A meta-analysis. *Personnel Psychology*, *59*, 623–664. https://doi.org/10.1111/j.1744-6570.2006.00049.x

Tobias, S. (1989). Another Look at Research on the Adaptation of Instruction to Students Characteristics. *Educational Psychologist*, *24*, 213–227. https://doi.org/10.1207/s15326985ep2403_1

Truong, H. M. (2016). Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities. *Computers in Human Behavior*, *55*, 1185–1193. https://doi.org/10.1016/j.chb.2015.02.014

Vanlehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, *46*, 197–221. https://doi.org/10.1080/00461520.2011.611369

Vogelaar, B., Resing, W. C. M., & Stad, F. E. (2020). Dynamic Testing of Children's Solving of Analogies: Differences in Potential for Learning of Gifted and Average-Ability Children. *Journal of Cognitive Education and Psychology*, *19*, 43–64. https://doi.org/10.1891/jcep-d-19-00042

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*, 932–945. https://doi.org/10.1037/a0031882

Waxman, H. C., Wang, M. C., Anderson, K. A., Herbert, J., Waxman, C., Wang, M. C., & Anderson, K. A. (1985). *Adaptive Education and Student Outcomes : A Quantitative Synthesis*. *78*, 228–236.

Wesson, C. L., King, R. P., & Deno, S. L. (1984). Direct and frequent measurement of student performance: If it's good for us, why don't we do it? *Learning Disability Quarterly*, *7*, 45–48. https://doi.org/10.2307/1510260

Yang, T.-C., Hwang, G.-J., & Yang, S. J.-H. (2013). Development of an Adaptive Learning System with Multiple Perspectives based on Students' Learning Styles and Cognitive Styles. *Journal of Educational Technology & Society*, Vol. 16, pp. 185–200.

https://doi.org/10.2307/jeductechsoci.16.4.185

***Paper 2***

**Tetzlaff, L.,** Edelsbrunner, P., Schmitterer, A., Hartmann, U., & Brod, G.: *A Person-Centered Approach to Modeling the Interactions Between Learner Characteristics and Instruction: Evidence for Differential Effectiveness of Reading Education.* Manuscript submitted for publication in *Journal of Educational Psychology*.

# A Person-Centered Approach to Modeling the Interactions Between Learner Characteristics and Instruction: Evidence for Differential Effectiveness of Reading Education

Leonard Tetzlaff[1], Peter Edelsbrunner[2], Alexandra Schmitterer[1], Ulrike Hartmann[1], Garvin Brod[1,3]

[1]DIPF | Leibniz Institute for Research and Information in Education

Rostocker Straße 6, 60323 Frankfurt, Germany

[2]ETH Zürich

Rämistrasse 101, 8092 Zurich, Switzerland

[3]Goethe Universität Frankfurt

Theodor-W.-Adorno-Platz 1, 60323 Frankfurt, Germany

## Author Note

Correspondence concerning this article should be addressed to:

Leonard Tetzlaff,
DIPF | Leibniz Institute for Research and Information in Education,
Rostocker Straße 6, 60323 Frankfurt, email: tetzlaff@dipf.de

**Declarations of Interest: None**

Parts of the results presented in this manuscript will also be presented at the GEBF Conference 2022.

All data and analysis code have been made publicly available and can be accessed at https://osf.io/a97gv/?view_only=73f3249ac61f4b618415d3116e1b164f.

## Abstract

The differential effectiveness of instructional approaches has been difficult to demonstrate for learners who differ across multiple dimensions of learning prerequisites. In the present study, we explored a person-centered approach to examining the differential effectiveness of reading instruction. In $N = 517$ German third-grade students, latent profile analysis identified four subgroups that differed across multiple characteristics consistent with the *simple view of reading*: poor decoders, poor comprehenders, poor readers, and good readers. Over a school year, different instructional foci showed differential effectiveness for students in these different profiles. An instructional focus on vocabulary primarily benefited good readers at the expense of poor decoders and poor comprehenders, while a focus on advanced reading abilities benefitted poor comprehenders at the expense of poor decoders and good readers. These findings demonstrate the utility of a person-centered approach to studying differential effectiveness and point to the need for more individualized reading instruction. We argue for the broad applicability of this approach to both differential effectiveness research as well as dynamic approaches to individualized instruction. Implications for adaptive teaching in reading instruction are also discussed.


*Keywords*: differential effectiveness, person-centered analysis, simple view of reading, aptitude-treatment interactions, latent profile analysis

## Educational Impact and Implications Statement

This study demonstrates the utility of person-centered analyses for investigating the effects of different instructional foci on the learning gains of specific learners. This will allow educators to better differentiate for which learners a specific intervention or instructional approach is most promising. The results of this study also indicate the need for more individualized reading instruction and give some indications as to what this might look like.

# A Person-Centered Approach to Modeling the Interactions Between Learner Characteristics and Instruction: Evidence for Differential Effectiveness of Reading Education

When teachers engage in differentiated instruction, they often begin by intuitively grouping students based on multiple characteristics that they consider relevant prerequisites for learning. For example, they might view some students as smart but lazy and others as motivated learners with weak self-regulation. Based on their experience and intuitive diagnosis, they use such categories as a heuristic for adapting their teaching to learners' individual needs (Corno, 2008).

In contrast to this common educational practice in which educators consider multiple characteristics for each student, most research investigating the differential effectiveness of educational interventions focuses on only one characteristic at a time (e.g., intelligence or prior knowledge; (Kalyuga, 2007, Ziegler et al., 2021). This limitation in the approaches used in research might be caused by analytic difficulties arising when trying to model interactions between multiple learner characteristics and treatments.

In the present study, we propose and use latent profile analysis (Hickendorff et al., 2018) as an approach to investigating the differential effectiveness of instructional foci based on multiple learner characteristics. We employ this approach to model the differential effectiveness of instructional parameters for students with different aptitude profiles. We situate this study within the area of reading instruction, because this field already has a well-developed model for multivariate learning prerequisites and their relation to educational success, as assessed by reading comprehension (Hoover & Gough, 1990; Hoover & Tunmer, 2018).

In this introduction, we begin with a brief review of previous research on individualized instruction and then address the need for and difficulties involved in modeling multivariate learner characteristics. We then introduce latent profile analysis as a person-centered approach to modeling multivariate learning prerequisites. Turning to the domain of reading instruction, we discuss the simple view of reading (Hoover & Gough, 1990), the model we use to label and interpret profiles of multivariate reading abilities. We finish this section by introducing the specifics of our study, which focused on third-grade reading instruction.

**Individualized Instruction**

Individualized instruction, the adaptation of instruction to the needs of specific learners or groups of learners, has long been regarded as an important aspect of teaching practice (Corno, 2008; Dockterman, 2018; Tetzlaff et al., 2021). Demographic changes and recently established educational policies, for example, regarding inclusive education, have increased student heterogeneity in classrooms in various countries (Corno, 2008; Decristan et al., 2017; Subban, 2006). Due to this increasing heterogeneity, individualized instruction has been a topic of increasing interest in recent times.

The promises of individualization are associated with the impressive effects of one-on-one in-person tutoring (Bloom, 1984; Vanlehn, 2011). Endeavors to individualize education can thus be understood as ways to scale up the effects of one-on-one tutoring to larger groups of learners, without having to provide a human tutor for each learner. This is difficult, especially in regular classrooms where instruction is supposed to address large groups of learners at the same time.

This problem has been addressed by several different approaches, including adaptive teaching (Corno, 2008), differentiated instruction (Constas & Sternberg, 2013), and formative assessment (Deno, 1990). These approaches all have one thing in common: The systematic

adaptation of instructional parameters based on the relevant characteristics of specific learners (Tetzlaff et al., 2021). By considering individual learner characteristics such as prior knowledge, cognitive capabilities, and affective/motivational traits and states, teaching agents adapt instructional parameters to achieve optimal fit and maximize potential learning gains across a group of students with heterogeneous preconditions (Grimm et al., 2021).

In order for these instructional adaptations to meet the aim of individualized instruction, instructional parameters need to show differential effectiveness for different learners. Differential effectiveness has been defined by Hunt (1975) as follows: "To consider the differential effectiveness of an educational approach (…) is not simply to point out a few persons to whom the principle does not apply (…). Rather than ask whether one educational approach is generally better than another, one asks, 'Given this kind of person, which of these approaches is more effective for a given objective?'" (Hunt, 1975).

When considering differences in effectiveness for different kinds of individuals depending on the learning objective, the concept of differential effectiveness can be seen as a variation of the aptitude-treatment-interaction (ATI) paradigm (Cronbach, 1957). Without these interactions of learner characteristics (aptitudes) and instructional parameters (treatments), some learners would learn better than others and some instructional approaches would be more effective than others, but adapting the instructional approach to specific learners would have no effect.

**The Challenge of Modeling Multivariate Learner Characteristics**

The most frequent approach in research on differential effectiveness (or ATIs) is to look at the interaction of one specific learner characteristic with different treatments (e.g., Bracht, 1970; Kalyuga, 2007; Seufert et al., 2009). For example, a well-known effect identified with this approach is expertise reversal: Learners with lower prior knowledge tend to benefit from stronger instructional guidance, whereas for learners with higher prior knowledge, the same

amount of guidance can be unnecessary or distracting (Jiang et al., 2018; Kalyuga, 2007). Similarly, it has been found that learners with lower general reasoning ability can benefit from stronger teacher-guidance, while learners with higher general reasoning ability can benefit from stronger self-guidance (Ziegler et al., 2020). Similar interactions have been found between working memory and the effects of conceptual versus fluency activities during instruction (Fuchs et al., 2014).

These studies all have one thing in common: They identify a single learner characteristic that might be of high importance for successful learning in a particular scenario. They then examine interactions of this learner characteristic with different educational interventions. However, most learners don't just differ in one relevant attribute. Rather, different learner characteristics can interact with one another, leading to different learning outcomes than would be the case for each characteristic on its own (Hooper et al., 2016; Lonigan et al., 2018; Reinhold et al., 2020). This phenomenon has been described under the name of aptitude complexes (Snow et al., 1987) or trait complexes (Ackerman, 2003). In order to correctly model interactions with treatment parameters, these multivariate learner characteristics should be taken into account.

Such multivariate interactions of learner characteristics might occur in various contexts (for some theoretical examples, see Cronbach, 1975) and may have the potential to provide informative insights for individualized instruction, however, it is difficult to find informative and reliable ways to statistically model such effects. This is due to the potential for a large number of different higher-order interactions when such interactions occur between multiple variables. In regular statistical models such as multiple regressions, this can easily lead to third-, fourth-, or even higher-order interactions in a regression model (Bauer & Shanahan, 2007) .One problem with such models is that they quickly become almost impossible to interpret

because, in an almost endlessly complex manner, the interpretation of any effect will always be qualified by another higher-order effect (Cronbach, 1975). Eventually, a large number of effects emerge, which all depend on each other, leaving researchers with a messy picture about what is going on (Bauer & Shanahan, 2007).

Reviewing early research on ATIs, Cronbach (1975) already identified potential higher order interactions as a problem, likening them to a hall of mirrors, which refers to the opaque picture created when entering a hall with mirrors that face each other. This hall of mirrors suffers not only from interpretational complexity but also other issues, such as finding models with sufficient statistical power to find effects in such complex data situations (Cronbach, 1975), and the nature of interactions that might not always be linear (Bauer & Shanahan, 2007), necessitating an even more complex model. Overall, whereas multivariate learning prerequisites are a topic of great interest for educational researchers, these methodological challenges make it difficult to adequately examine such prerequisites and their interactions with educational interventions. To the best of our knowledge, even recent best-practice recommendations for the examination of aptitude-treatment interactions do not tackle the issue of how to best model multivariate aptitudes and overcome these obstacles (Preacher & Sterba, 2019).

**Person-Centered Approaches to Multivariate Learning Prerequisites**

In the present study, we propose a solution to the methodological issues in the hall of mirrors: a person-centered approach to the investigation of differential effectiveness involving multivariate learning prerequisites. Person-centered analysis provides a possible avenue of investigation into these processes by moving the focus of analysis from the interactions of single variables to entire persons and their characteristic constellation of learning prerequisites. The specific approach that we propose and use in the present study is that of latent profile analysis (Hickendorff et al., 2017). In a latent profile analysis, learners are grouped according

to their constellations of mean patterns across multiple variables. Thus, the different profiles represent groups of learners who systematically differ in their multivariate learner characteristics. By grouping learners into profiles using this method and then comparing the effects of educational interventions between those groups, their differential effectiveness across different patterns of multivariate learning prerequisites can be examined. Overall, this approach has the advantage of being able to group many learners into just a few distinct categories. This feature helps to achieve high statistical power while capturing multivariate and potentially nonlinear information across multiple learning prerequisites in a much more concise manner than a regression analysis with higher-order interactions (Bauer & Shanahan, 2007).

Indeed, recent research has started applying latent profile analysis and similar methods to model aptitudes: Hooper et al. (2016) identified profiles of learners who differed in their language, problem solving, attention, and self-monitoring characteristics. Learners with different profiles showed differential development in writing skills during an intervention. With a similar approach, Lonigan et al. (2018) found that language-minority children's profiles of proficiency in their first and second language predicted their development of early literacy skills in preschool. Reinhold et al. (2020) identified subgroups of sixth graders with different engagement profiles that were systematically related to their development in mathematics achievement. Finally, Grimm et al. (2021) applied latent profile analysis on outcome variables instead of aptitude variables to model the differential effects of experimental conditions on third graders' development of multiple reasoning skills.

These fruitful applications of latent profile analysis indicate that the community of researchers interested in multivariate learning prerequisites has started acknowledging the potential of such person-centered approaches. However, one characteristic that these studies have in common is that, while they did investigate how learner profiles relate to learning

outcomes, they did not investigate the actual differential effectiveness of educational interventions. In addition to directly influencing the learning outcome, multivariate learner characteristics can affect the effects of specific instructional parameters. Only when we understand this interaction, can we speak of differential effectiveness and use the data as a foundation for providing different treatments to different learners (Hunt, 1975). We are not aware of prior research taking this person-centered approach to investigating differential effectiveness as we define it here. There are, however, studies finding differential effectiveness of treatments for subgroups of learners that have been grouped with other approaches. An example of this is a study by Hofer et al., (2018) in which intelligence and gender interacted in their effect on the efficacy of cognitively activating instruction in physics. Female students with intelligence estimates in the highest sample-based quartile benefitted the most from an intervention implementing various means of cognitive activation in comparison to a business-as-usual approach. In this study, we use latent profile analysis to examine the differential effectiveness of different reading interventions for different groups of learners who systematically differ in their prerequisite skills for reading according to the well-known model of the simple view of reading.

**The Simple View of Reading**

Reading comprehension is a complex skill that is constituted by an interplay of several different components (Kendeou et al., 2016). One of the most prominent theories to explain how these components relate and how they interact to form the construct of reading comprehension is the simple view of reading (SVR, Hoover & Gough, 1990). According to that theory, reading comprehension can be modeled as the product of linguistic comprehension (LC) and decoding (D) abilities. Decoding (D) is defined as efficient word recognition: the ability to quickly access the appropriate entry in the mental lexicon after seeing a written word (Hoover & Gough, 1990). It covers both the fluency as well as the accuracy of the decoding

process and is usually measured with pseudoword reading tasks. Linguistic comprehension (LC) is defined as the ability to take lexical information (i.e., semantic information at the word level) and derive sentence and discourse interpretations (Hoover & Gough, 1990). It is often conceptualized as listening comprehension and assessed through the retelling of a text that has been read aloud by another person (Hoover & Tunmer, 2018), but it has also been assessed through vocabulary knowledge (e.g., Singer & Crouse, 1981; Tunmer & Chapman, 2012) or syntax comprehension (Tilstra et al.,2009).

The multiplicative nature of the model implies a difficulty in compensating for deficits in one of the two abilities with improved performance in the other. Evidence of this relation has been found in multiple alphabetic languages (see Hjetland et al., 2020 for a comprehensive review). This two-dimensional conceptualization means that readers fall into one of four quadrants of reading comprehension: *Good readers* have good word reading/decoding skills, complemented by good comprehension abilities, *poor readers* lack in both abilities, while *poor decoders* as well as *poor comprehenders* show good performance in one of the two components, combined with poor performance in the other.

We rely on the SVR here not because we believe it accurately describes all aspects of reading performance (see Castles et al., 2017 for a discussion of its accuracy), but because it serves as a good basis for considering both theoretically grounded multivariate aptitude profiles, as well as their interaction with instruction. For the present study, it is important to note that the SVR implies specific predictions regarding optimal reading instruction for learners with different preconditions. A straightforward deduction would be that poor comprehenders benefit from instruction that specifically targets comprehension while poor decoders would benefit from instruction that specifically targets decoding. This is supported by a meta-analysis by Galuschka et al., (2014) showing that, in general, interventions that target

children's specific deficits are more effective in alleviating their reading difficulties than more general approaches. On the other hand, there are also theories of dyslexia (and supportive evidence) that postulate a deficit in phonological working memory as the root cause, in which case decoding-focused instruction might fail to help those children (e.g. Alt et al., 2021; Menghini et al., 2011; Smith-Spark & Fisk, 2007). In the present study, we examine how the instructional emphasis that teachers put on different aspects of reading instruction meets the actual needs of students with these different constellations of learner characteristics and thereby helps them improve their reading comprehension.

**Instructional Foci in Third-Grade Reading Lessons**

In the present study, we examine the differential effectiveness of reading instruction foci in third-grade classes within the context of German elementary schooling. The curriculum for third-grade students in the federal German state of Hesse (the location of our study) generally puts a strong emphasis on strengthening students' reading motivation, teaching advanced reading abilities such as passage comprehension and summarizing texts, and fostering vocabulary acquisition—for example, by finding synonyms of words or using previously unknown words in exemplary sentences (Hessian Ministry of Education, 2021). Which of these aspects is emphasized at which point in time, and how much time is invested in each of these aspects, will naturally vary from teacher to teacher. We hypothesize that the instructional foci teachers choose will differentially affect students based on their individual learning prerequisites. While these instructional foci are not identical to specific training of these reading-related abilities, we still assume that it is useful to refer to findings from related training studies to inform our hypotheses about their differential effectiveness.

Fostering reading motivation can take many forms, for example, encouraging learners to seek out and read literature based on their own interests. Reading motivation is positively related to reading comprehension (Kuşdemir & Bulut, 2018), and longitudinal studies indicate

a reciprocal relation between reading comprehension and reading motivation in second and third grades (e.g. Schiefele et al., 2016). This strongly implies that, at least for a specific subset of students—those who possess the necessary skills to read texts without instructional support—fostering reading motivation could lead to increased reading comprehension over time. An assumed mediating mechanism of this relationship is the increased frequency of reading in out-of-school contexts in highly motivated readers (Guthrie et al., 1999). In this case, it is likely that positive effects of interventions focusing on reading motivation appear with a considerable delay.

Advanced reading abilities comprise several different techniques dealing with passage comprehension, such as highlighting important aspects, writing short summary sentences for specific passages, and rephrasing content in one's own words. As advanced reading abilities can be addressed by a broad spectrum of instructional approaches, it is difficult to derive specific predictions from the training literature. It is reasonable to assume that a focus on comprehension skills mainly benefits those children who specifically struggle with comprehension, as opposed to those struggling with decoding, or with both.

While vocabulary training has repeatedly been shown to increase vocabulary size/word knowledge (e.g., Segers & Verhoeven, 2003), a transfer to reading comprehension abilities seems to be limited (Mezynski, 1983). This is especially interesting given the strong correlation between vocabulary size and reading comprehension (Carroll, 1993; Freebody & Anderson, 1983). These results are not completely unambiguous though. A meta-analysis of these transfer effects by (Elleman et al., 2009) found significant variability between studies, mostly dependent on the type of vocabulary measure used, but also in relation to students' learning prerequisites. This variability might point to the efficacy of vocabulary training for learners with specific learning prerequisites, but not for others. Therefore, investigating the differential

effectiveness of that transfer is of particular interest: Are there children that make progress in reading comprehension as a result of vocabulary training, and how are their learning prerequisites constituted?

**The Present Study**

In the present study, we employed the person-centered approach of latent profile analysis to examine the differential effectiveness of different instructional foci for third graders' development of reading comprehension. Specifically, we posed the following two questions:

1) Which latent profiles of reading abilities exist within a group of third-grade learners? We employed latent profile analysis to examine whether meaningful differences in profiles regarding decoding and comprehension abilities exist among third graders.

2) Do different instructional foci show differential effectiveness across readers with different profiles over the course of one school year? To examine this question, we assessed teachers' foci throughout the course of one school year. By instructional foci, we refer to the emphasis that they put on different aspects of reading instruction (i.e., vocabulary, advanced reading abilities, and reading motivation). Based on the rationales outlined above, we hypothesized that a focus on reading motivation will benefit good readers—those who already show good decoding and word reading skills—while we left the research question of which learners benefit from vocabulary training or advanced reading training open.

## Method

The current study was carried out as part of a larger research project running from 2018–2020 and was approved by the Ethics Committee of DIPF. Data were collected in two cohorts, one in the school year 2018/2019 in the state of Hesse in Germany, the other in the school year 2019/2020 in the states of Hesse and Lower Saxony. Each cohort completed a pretest at the beginning of the school year and a posttest before the summer break. All tests

were administered by research assistants, apart from the posttests in the school year 2019/2020 that had to be administered by the respective teachers due to pandemic-related school lockdowns. Since the tests were easy to administer, we do not expect that this caused a significant reduction in data quality. The teachers and their students participated in the study on a voluntary basis and did not receive any compensation. As the recruitment was done in cooperation with the ministry of education of Hesse, we did not put an upper limit on the sample size. A simulation study by Nylund et al. (2007) indicates that a sample size of at least 500 is recommended to identify the correct number of profiles in a latent profile analysis.

**Transparency and Openness**

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and we follow JARS (Kazak, 2018). All data, analysis code, and research materials are available at [https://osf.io/a97gv/?view_only=73f3249ac61f4b618415d3116e1b164f]. Data were analyzed using R, version 4.0.2 (R Core Team, 2020) and the package MplusAutomation, version 0.78-3 (Hallquist & Wiley, 2018) as well as Mplus 8 (Muthén & Muthén, 2017). As it is an exploratory study, its design and its analysis were not pre-registered.

**Sample**

We relied on two different samples for the analysis: The whole sample of the present study was used for the analysis related to research question one, while a subsample was used for the analysis for research question two. For the whole sample, the teacher group consisted of 59 teachers in Hesse and 16 teachers in Lower Saxony. Teachers were, on average, 41 (*SD* = 9.19) years old. Two of them were male, the remaining 73 were female. They reported an average of 13.26 (*SD* = 7.58, range = 4–39) years of general teaching experience and 5.68 (*SD* = 5.56, range = 0–27) years of experience teaching third-grade classes. The teachers nominated

some of their students ($N$ = 517 in total) to participate in individual testing sessions. Teachers were allowed to select up to eight children. They were asked to prioritize children with reading difficulties and then add children to be representative of the class, based on their own criteria. These students were on average 8.32 ($SD$ = 0.56) years old. We used the data of these nominated students from all classes for the first set of analyses, with the aim of identifying latent profiles of readers.

For the analysis related to the second research question, which dealt with differential effects of instructional foci, we used a subsample of 52 teachers in Hesse who completed additional online questionnaires on their instructional foci. Teachers in that sample were on average 38 ($SD$ = 8.19) years old, with 11.91 ($SD$ = 6.41) years of teaching experience, 4.60 ($SD$ = 4.34) of which in third-grade classes. The respective student sample comprised 217 students and was on average 8.34 ($SD$ = 0.56) years old.

**Assessment Materials**

*Teachers' Instructional Foci*

Throughout the school year, teachers were presented an online questionnaire every three weeks in which they were asked about their teaching practices during that time. They were asked to fill out this questionnaire eight times during the school year, and the average participation rate was 4.10 times (range 1 to 8). Teachers' self-reported practices in reading instruction showed a moderate to high stability ($ICC(1)$ = 0.21-0.51) across assessments, indicating that a) they did not vary their instructional focus much and b) the aggregated measure can be seen as a reliable estimate of the instructional landscape over the whole school year. Fleiss (1986) proposes that an $ICC(1) > .15$ serves as a reasonable benchmark for these kinds of assessments. To examine the instructional foci of interest in this study, we asked them how much emphasis they put on "vocabulary training," "advanced reading abilities (sentence and passage comprehension)," and "reading motivation." The other measures in the questionnaire

pertained to the organization of instruction (e.g., peer-teaching, individualized attention) or were not reflective of regular third grade reading instruction (e.g., a focus on precursor abilities). All of these were answered on a 4-point Likert scale with the anchors *never* and *often*. For the current analyses, we used average values across time for each of the three selected instructional foci. By regularly asking teachers which of these aspects they emphasized in their classes and aggregating these measures over the whole school year, we can create a picture of the instructional landscape in that specific classroom.

*Reading Comprehension*

As our measure of general reading comprehension, we used the pen and paper version of the ELFE II – Leseverständnistest für Erst- bis Siebtklässler (Lenhard & Schneider, 2006). The ELFE II measures reading comprehension at the word, sentence, and text level. For the word comprehension task, children are presented a picture and a group of four written words and asked to select the word that matches the picture. For the sentence task, children are presented with an incomplete sentence and asked to select one of five written words to complete it. For the text task, children are asked to read short passages and then select the statement that best corresponds to the passage out of a choice of four. The items in the test showed a high internal consistency (alpha = 0.96, omega = 0.97). The ELFE was administered as a pen and paper version to the whole class at the same time. For the first cohort, as well as the pretest of the second cohort, the tests were conducted by our trained research assistants; the posttest of the second cohort was conducted by the teachers themselves instead. For statistical analysis, we used the mean score of each student across the different sub-skills measured by the ELFE as a general indicator of reading comprehension.

*Decoding Ability*

As our measure of decoding ability, we used the pseudoword part of the Salzburg Reading and Writing Test SLRT II – Salzburger Lese und Rechtschreib Test (Moll & Landerl, 2010). In this test, children are asked to read a written list of pseudowords aloud while the experimenter keeps track of the amount of correctly read words. We did not estimate the internal consistency of this measure in our sample as this would have required recording the full sessions with the participants, which we did not do. The test, however, generally shows very high reliability estimates, and it is a test commonly used for diagnostics of individual children that require high precision. The manual reports the reliability (measured via parallel tests) as between .90 and .98 (Moll & Landerl, 2010). For statistical analysis, we used the amount of correctly read pseudowords within one minute.

*Grammatical Comprehension*

As our measure of grammatical comprehension, we used the screening in the TSVK – Test zum Satzverstehen von Kindern (Siegmüller et al., 2011). In this test, children are asked to select the one picture, out of a set of three, that corresponds most to a sentence that was read aloud to them. The items in the test showed a satisfactory internal consistency (alpha = 0.66, omega = 0.69). For statistical analysis, we used the amount of correctly selected pictures as a measure of learners' grammatical comprehension, which we used as an indicator variable for their linguistic comprehension.

*Expressive Vocabulary*

As our measure of expressive vocabulary, we used the WWT – Wortschatz- und Wortfindungstest (Glück, 2011). In this test, children are asked to produce 40 nouns, verbs, or adjectives represented by pictures on a computer screen. The test also provides a list of synonyms that would be counted as correct.

The items in the test showed a high internal consistency (alpha = 0.88, omega = 0.91).

For statistical analysis, we used the sum of correctly produced words as a measure of learners' expressive vocabulary, which we used as the second indicator of linguistic comprehension. The tests for vocabulary, syntax comprehension, and decoding were administered by trained research assistants in individual sessions.

**Analytic Approach**

In order to estimate student profiles of reading skills and reading comprehension, we conducted latent profile analyses (Ferguson et al., 2020; Harring & Hodis, 2016; Hickendorff et al., 2018) using the software package Mplus 8.3 (Muthén & Muthén, 2021). Before undertaking these analyses, we *z*-standardized the indicator variables for improved interpretability.

The indicator variables used as the basis of the student profiles were learners' scores on decoding (SLRT), linguistic comprehension (two scores: one each from the TSVK & the WWT), and reading comprehension (ELFE at T1). Based on these four indicator variables, in a stepwise manner we increased the number of profiles from one to seven, after which it was evident that fit indices were getting worse and model convergence was not possible anymore. As is common practice in latent profile analysis, the model with the actual number of profiles interpreted and used for further analyses was then selected based on fit indices and theoretical considerations (Ferguson et al., 2020; Harring & Hodis, 2016; Hickendorff et al., 2018). To this end, we relied in particular on the fit indices BIC, aBIC, and the VLMR-likelihood ratio test (Edelsbrunner & Flaig, 2021; Ferguson et al., 2020; Hickendorff et al., 2018; Lo et al., 2001). For the BIC and aBIC, lower estimates indicate a better relative model fit (for explanations of these fit indices, see Edelsbrunner & Flaig, 2021), and for the VLMR, the

model with the highest number of profiles reaching significance should be selected (Ferguson et al., 2020; Harring & Hodis, 2016; Hickendorff et al., 2018).

Regarding technical specifications, each model was estimated with 400 initial random starts, of which the most promising 100 were used for further estimation. The estimation method was maximum likelihood with expectation-maximization optimization and Huber-White standard errors that are robust against multivariate kurtosis and heteroscedasticity (Freedman, 2006). We took the multilevel-structure of the data into account through a cluster-robust estimation of the standard errors in all modeling steps (Muthen & Muthen, 2017). We did not, however, employ full multilevel modeling because the relatively stable reading-related abilities at the beginning of a school year should not be strongly influenced by the class of the students (Hoover & Tunmer, 2018). For convergence criteria during estimation, we used the Mplus defaults, and we accepted a model as converged when the best likelihood was achieved multiple times. Apart from the means, we also estimated the variances freely within each profile (Edelsbrunner & Flaig, 2021; Ferguson et al., 2020; Hickendorff et al., 2018).

To examine the second research question, concerning the differential effectiveness of instructional foci, we related profile membership to the reading comprehension of the students in the subsample at the end of the school year and to teachers' foci in reading instruction. To this end, we decided on the BCH-method (for details on this method and its implementation, see Asparouhov & Muthen, 2014). This approach has an advantage in that it allows for a modeling of differential effectiveness for students in different profiles while correcting for measurement error, in a way similar to a structural equation model (Vermunt, 2010). We regressed the participants' reading comprehension at posttest on the instructional foci of their teachers. In a last addition, to check for the interaction between students' reading profiles and teachers' instructional foci, we defined derived parameters. These parameters indicated differences between the different reading profiles in the regression weights of reading

comprehension on teachers' instructional foci. These parameters allowed us to check for differences between profiles, as well as to identify the main effects of each instructional focus across all profiles.

For statistical inference, we present and interpret 90% confidence intervals for all focal model parameters. The present analyses have a rather exploratory than confirmatory approach, which undermines the reliability of significance testing. Instead, we focus on confidence intervals and interpret them as follows. If a confidence interval excludes 0, we interpret this as evidence pointing towards a hypothesis that should be further investigated in future research. If a confidence interval includes 0, we cautiously interpret this as lack of evidence for an effect of interest. The results with 95% confidence intervals are provided in the analysis output. All data, syntaxes and output files are available under https://osf.io/a97gv/?view_only=73f3249ac61f4b618415d3116e1b164f.

## Results

### Descriptive Statistics

Students in our sample achieved slightly below-average scores across all reading abilities at pretest, when compared with a norm sample (*t*-values ranging from 39.97 for vocabulary to 46.61 for reading comprehension). This is probably due to the teacher-selected student sample: Teachers were asked to prioritize children with reading difficulties and then add children to be representative for the class. This leads to a slight over-sampling of struggling readers. The negative correlations between indicators of decoding and linguistic comprehension (see Table 1) have also been reported for samples low on reading comprehension (Hoover & Gough, 1990).

The means of the aggregated teacher self-reports on their instructional foci were consistently in the upper half of the 4-point Likert scale (3.0 for vocabulary training and 3.23

for advanced reading abilities and reading motivation) indicating a general tendency to report the presence, rather than the absence, of specific instructional foci. This is consistent with findings concerning biases resulting from social or educational desirability of certain types of instruction (Kopcha & Sullivan, 2007).

**Table 1**

*Intercorrelations of the Indicator Variables*

| | Syntax Comprehension | Reading Comprehension (T1) | Decoding |
|---|---|---|---|
| Vocabulary | .61 | .11 | -.23 |
| Syntax | - | .17 | -.10 |
| Reading | | - | .62 |

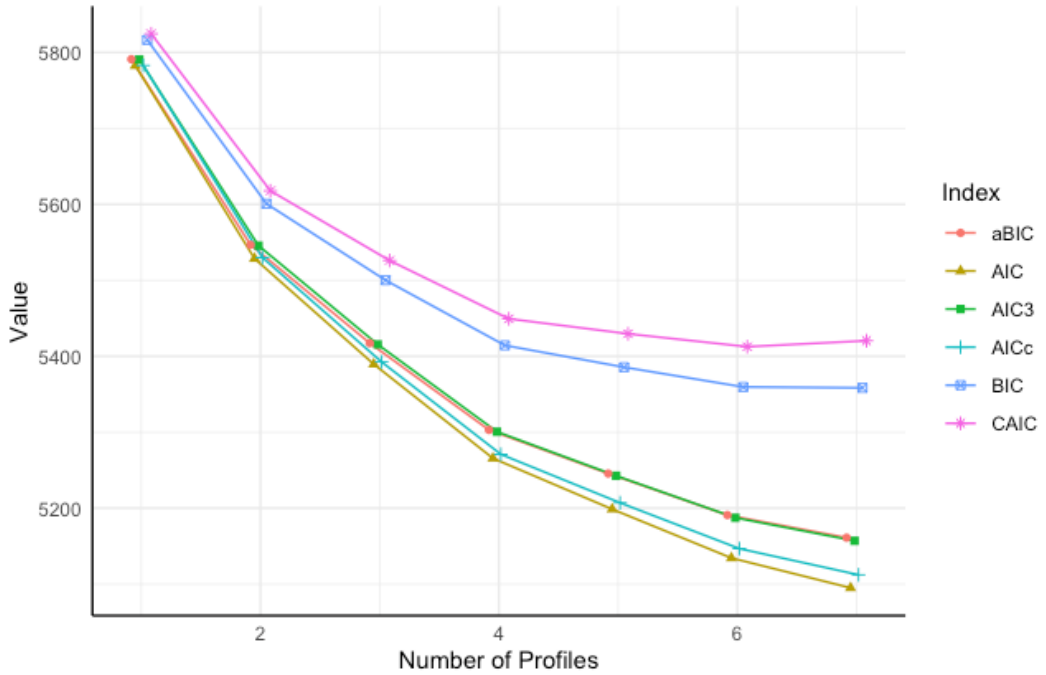
**Student Profiles of Learning Prerequisites**

To investigate the first research question, the first step was to decide how many separate profiles were present in the data. Figure 2 depicts the different fit indices for the models with different numbers of profiles. As shown in Figure 2, the BIC had its lowest value with six profiles and the aBIC with seven profiles. The VLMR-test, however, indicated significant improvement in model fit only with up to four profiles ($p$ = .02) but not, for example, with five profiles ($p$ = .07). In addition, the BIC and aBIC showed a visible decrease in strength of improvement from the four- to the five-profile models. Given these indications and the straightforward interpretability of the four-profile solution (see below), we decided to proceed

with the four-profile model. Results with the five-profile model can be found in the supplementary materials.

**Figure 1**

*Fit indices for different numbers of profiles*



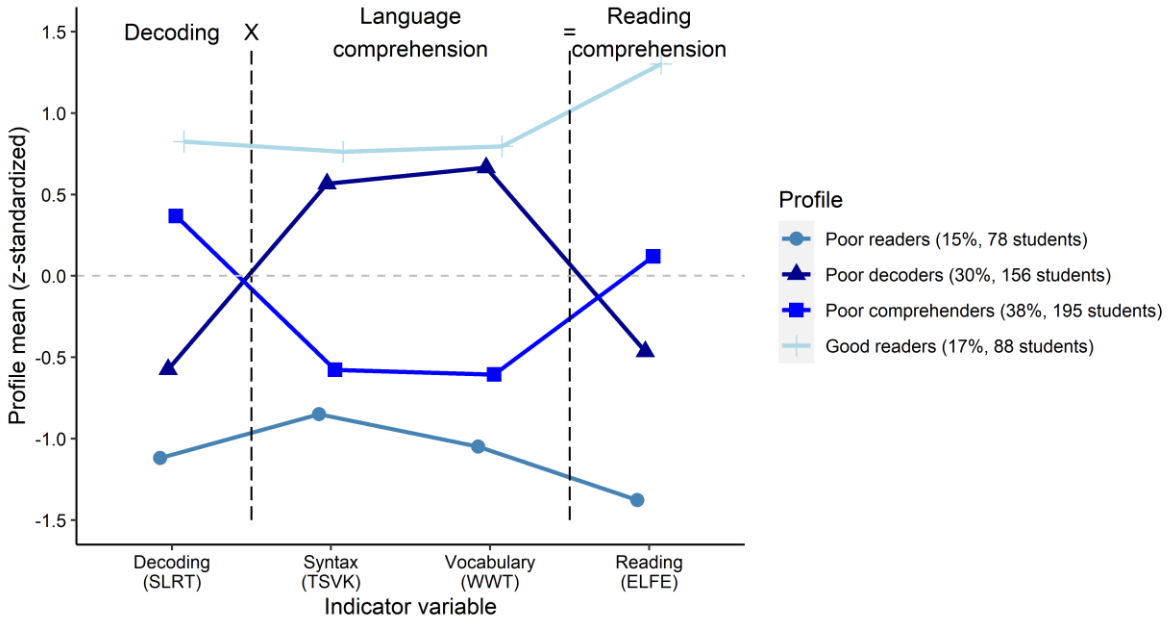

*Note.* the elbows in the curves indicate declines in information gained from the addition of more profiles.

The four identified profiles can be labeled in accordance with the SVR (see Figure 2). The poor readers ($N$ = 78) are mainly characterized by their low performance in the reading comprehension task, but they also score well below the mean on all the other indicator variables. The good readers ($N$ = 88) have extraordinarily high performance in reading comprehension, complemented by scores well above the mean on all the other indicator variables. The poor comprehenders ($N$ = 195) have slightly above average reading comprehension skills, strong decoding skills, weak syntax comprehension, and below average expressive vocabulary. The poor decoders ($N$ = 156) have slightly below average reading comprehension skills as well as decoding skills, balanced by strong performance in the linguistic comprehension indicators.

In sum, by employing latent profile analysis, we were able to identify informative profiles across multiple relevant learner characteristics. The strong correspondence of these profiles with the SVR speaks to their validity and provides some indications concerning potential interaction with instruction, which we investigated next.

**Figure 2**

*Profiles of Readers According to the Simple View of Reading*



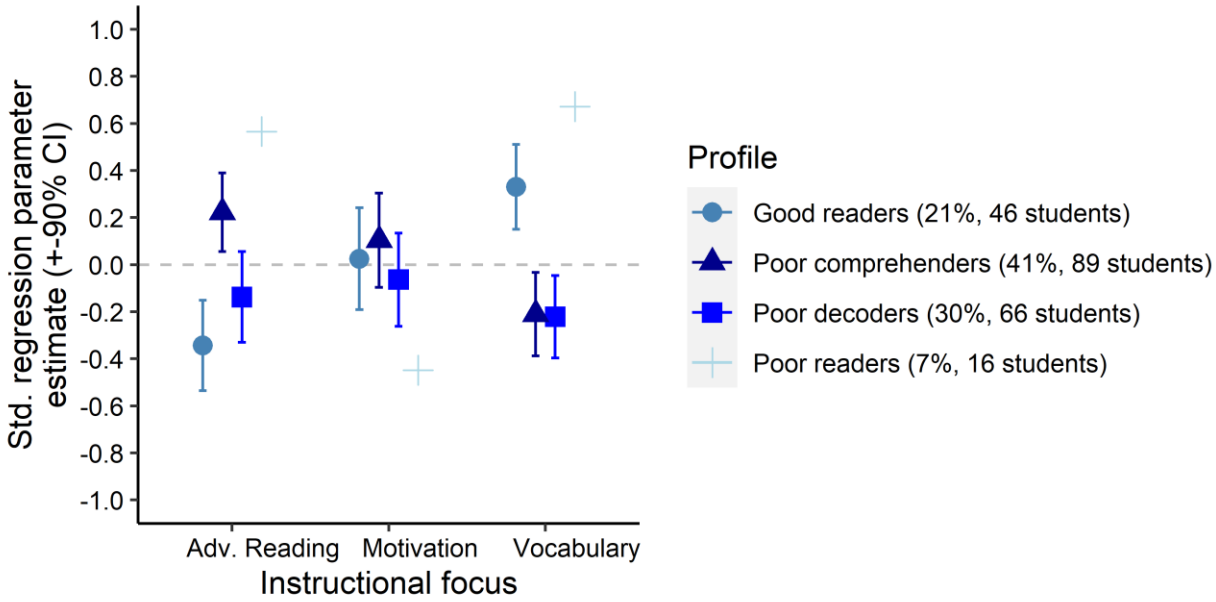
**Differential Effectiveness of Instructional Foci for Students in Different Profiles**

For the second research question, concerning the differential effectiveness of instructional foci for students in the different profiles, we used data from the subsample of third graders whose teachers ($N$ = 52) had participated in the surveys assessing their instructional foci. Of the $N$ = 217 students eligible for this analysis, 89 (41.01%) belonged with the highest probability to the poor comprehenders profile, 66 (30.42%) to the poor decoders, 46 (21.20%) to the good readers, and 16 (7.37%) to the poor readers. Please note that, since only a few students belonged to the poor readers profile, we were not able to investigate differential effectiveness for those students due to too little statistical information.

We first extracted the information on students' profiles from the latent profile analysis conducted for the first research question. We then used the bias correcting BCH-method (Asparouhov & Muthén, 2014) to set up regression models predicting reading comprehension at the end of the school year within each profile from teachers' instructional foci. When looking at the profile-specific regression weights (Figure 3), we can see that—consistent with the predictions made by the SVR—poor decoders do not improve when vocabulary is the focus. Poor comprehenders also do not improve with a focus on vocabulary, but they benefit from a focus on advanced reading skills. Good readers benefit from a focus on vocabulary and suffer when the focus is on advanced reading skills. Since the group of the poor readers which also had self-report data from their teachers only consisted of approximately 16 students, we are unable to make any reliable statement about them.

**Figure 3**

*Parameter Plot of Regression Weights Describing Effects of Instructional Foci on Reading Comprehension for Students in Different Profiles*




*Note***.** Confidence intervals for poor readers are not shown because these would exceed the plot range due to small sample size. Respective parameters are provided in Table A1.

Direct evidence for differential effectiveness is present if the different instructional foci show differential effects for learners in the different profiles. To examine this statistically, we set up contrasts that represented group differences in the effects of the instructional foci for learners in the different profiles. The first information that we examined before inspecting differential effectiveness were the simple main effects of teachers' foci on students' reading comprehension at the end of the school year. These effects, estimated in a simple multiple regression, were all close to 0 (Table 2).

While none of the instructional foci showed any main effect for all participants, the specific contrasts for the different instructional foci (see Table 2) show a strong differential effectiveness of vocabulary training between good readers and poor decoders, as well as between good readers and poor comprehenders. They also indicate a strong differential effectiveness of training advanced reading abilities between good readers and poor

comprehenders, as well as between poor decoders and poor comprehenders. For the fostering of reading motivation, no meaningful differential effectiveness could be found.

**Table 2**

*Main Effects and Specific Contrasts on Vocabulary and Advanced Reading Abilities, Including 90% Confidence Intervals*

| Contrast | Beta | SE | Lower 5% | Upper 5% |
|---|---|---|---|---|
| Vocabulary | | | | |
| Main Effect | -0.034 | 0.095 | -0.190 | 0.122 |
| PD vs. PC | 0.012 | 0.166 | -0.261 | 0.284 |
| GR vs. PC | 0.54 | 0.164 | 0.272 | 0.810 |
| GR vs. PD | 0.552 | 0.167 | 0.278 | 0.827 |
| Advanced | | | | |
| Main Effect | 0.018 | 0.116 | -0.200 | 0.191 |
| PD vs. PC | 0.361 | 0.134 | 0.140 | 0.581 |
| GR vs. PC | -0.567 | 0.150 | -0.813 | -0.320 |
| GR vs. PD | 0.206 | 0.163 | -0.063 | 0.457 |
| Motivation | | | | |
| Main Effect | -0.005 | 0.119 | -0.172 | 0.209 |
| PD vs. PC | 0.168 | 0.170 | -0.112 | 0.448 |
| GR vs. PC | -0.079 | 0.166 | -0.353 | 0.194 |
| GR vs. PD | -0.089 | 0.971 | -0.346 | 0.168 |

Note. PD = Poor Decoders; PC = Poor Comprehenders; GR = Good Readers; Poor Readers were excluded due to their limited sample size.

## Discussion

In the present study, we tackled the ongoing issue of finding an appropriate approach to modeling multivariate learning prerequisites and their interactions with instructional parameters. We found that a person-centered approach, specifically a latent profile analysis, identified profiles of reading skills that were in accordance with the simple view of reading and that different instructional foci proved differentially effective for readers with different profiles. This differential effectiveness emerged despite a lack of main (average) effects of the different instructional foci across all students, indicating that none of the foci is, by itself, preferable to the others. Only when learner characteristics are taken into account can an informed selection be made. To the best of our knowledge, this is one of the first studies successfully demonstrating interactions between multivariate learning prerequisites and instructional interventions (but see Hofer et al., 2018). This discussion will focus first on the general implications for educational research and then on implications for reading instruction.

### The Benefits of Person-Centered Analysis

How to model multivariate learner characteristics and their interaction with instruction has remained a more or less unresolved problem in educational psychology (Cronbach, 1975). In this study, we make a strong case for utilizing person-centered analysis to group learners in accordance with their pattern of means across multiple relevant characteristics as well as to investigate the differential effectiveness of specific instructional parameters for these groups. Instead of asking for which *levels* of a certain characteristic a specific treatment is most effective, researchers should ask for which *learners*—with their specific constellation of learning prerequisites—a treatment is most effective. Person-centered analysis such as latent

profile or class analysis allows researchers to pose and answer this question. This approach has the added benefit of also identifying nonlinear relations between aptitudes and treatment effectiveness that would be lost, or at least be difficult to track and interpret, in traditional variable-centered analyses (Bauer & Shanahan, 2007). Utilizing the 3-step BCH method (Asparouhov & Muthén, 2014) further allows for integration of the class-specific regression terms without either biasing the profile estimation or embezzling measurement error. In sum, person-centered analysis circumvents several problems associated with modeling multivariate learner characteristics and their interaction with instruction and thus can be seen as a promising approach for future studies in these domains. With its multivariate nature, our approach goes beyond the state-of-the-art methods for the analysis of differential effectiveness/aptitude-treatment interactions, which, to the best of our knowledge, still focus on univariate aptitudes (Preacher & Sterba, 2019).

**Implications for Reading Instruction**

Besides the main aim of demonstrating how fruitful the application of person-centered analysis can be for investigating differential effectiveness, our results also provide some support for the simple view of reading (Hoover & Gough, 1990). The four identified profiles suggest that the SVR is valid for describing reading performance on a level that is useful for classroom instruction. This is not a completely new finding: Torppa et al. (2007) used latent profile analysis to identify five subgroups of reading performance that correspond to the four profiles predicted by the SVR, with an additional profile of average readers.

Wolff (2010) identified eight latent profiles across ten reading-related abilities. Of these eight profiles, three proved to be especially stable—good readers, poor decoders (dyslexics), and poor comprehenders. Foorman et al. (2017), on the other hand, conducted several latent profile

analyses to identify reading profiles in different age groups and found that, while profiles in elementary grades show heterogenous deficits, profiles in higher age groups mostly showed a high, medium, low pattern of parallel profiles. This implies that the identified patterns are subject to various developmental trajectories and thereby not necessarily stable over longer timeframes.

In addition to the specific profiles identified, we observed that the deviation from the mean for the good and poor readers is especially pronounced for reading comprehension. This is in line with the presumed multiplicative relation between decoding and linguistic comprehension in the simple view of reading (Hoover & Gough, 1990). Thus, when the two skills of decoding and linguistic comprehension are both not yet developed enough or both well-developed, they have an even stronger effect on the resulting reading comprehension level than when only one of the two prerequisite skills is high or low.

While this was an exploratory study, the class-specific regression weights of specific instructional foci still provide some implications for reading instruction:

1) The negative effect of vocabulary training on the reading comprehension of poor comprehenders indicates that their below-average expressive vocabulary can be seen as a result or symptom of their comprehension deficit (mediated by frequency of reading), rather than a cause (Duff et al., 2015; Suk, 2017). This again highlights the strength of the multivariate approach. Simply looking at the specific deficits of children classified as poor comprehenders would make vocabulary and syntax training the straightforward choice (Galuschka et al., 2014). By instead teaching comprehension skills to complement their already strong decoding abilities, teachers can enable those children to improve their vocabulary on their own, while also increasing their reading comprehension level (Verhoeven et al., 2011; Share, 1999).

2) In contrast, for children who already have strong decoding and comprehension abilities, the most valuable focus seems to be on vocabulary extension, even though they

already possess a good expressive vocabulary. This is plausible since it is possible that the best way to improve a completely automated reading process is to add even more words to the mental lexicon.

3) The observation that fostering reading motivation did not have any significant effect on the development of reading comprehension, regardless of profile, can potentially be explained by mediating mechanisms. If the positive effects of reading motivation on reading comprehension are, for example, mediated by an increased frequency of reading outside the school context, they may take longer to manifest than the time frame of this study was able to capture (Guthrie et al., 1999; Retelsdorf et al., 2011).

It is important to note that, at a higher level, our results indicate a need for individualized or differentiated reading instruction. A uniform instructional focus for an entire class of students is certain to be a wasted opportunity for some students. The strong differential effectiveness of specific instructional approaches depending on measurable multivariate aptitude profiles implies a need for stronger individualization or at least differentiation of instruction—this is in line with previous research on individualized reading instruction (Connor et al., 2009; Connor et al., 2007). Basing instructional adaptations on multivariate aptitude profiles, rather than specific univariate deficits, potentially enables even more effective individualization that also builds on students' individual strengths.

**Limitations**

One limitation of the present study is its primarily exploratory nature. This implies that both the identified profiles and the observed interaction with instructional foci need to be replicated before they can be used to inform specific instructional approaches. Regarding the profile analysis, it should be mentioned that the selected indicator variables are not completely exhaustive indicators of linguistic comprehension. However, both vocabulary knowledge

(Tunmer & Chapman, 2012) and syntax comprehension (Tilstra et al., 2009) have been identified as important aspects of linguistic comprehension.

The teacher self-reports might be biased by educationally desirable response tendencies, leading to an overestimation of the amount of focus they put on specific instructional practices. This is indicated by consistently positive correlations between the different instructional foci. If teachers managed to report their instructional practices in a reliable manner, it would result in lower intercorrelations between the amount of focus they put on each of the individual instructional practices. For example, if teachers who teach more advanced reading during a given period tend to teach less vocabulary because there is only enough time for one of them, then this should result in a negligible or even negative correlation between these two instructional foci. The fact that we found positive correlations between all instructional foci indicates that on average, teachers might tend to engage in consistently positively biased response behavior across all instructional foci. Prior research indicates that, although teacher self-reports about teaching practices are related to actual classroom observations (Mayer, 1999), such self-reports might still be biased by social desirability aspects (Wubbels et al., 1992). However, for our focal interpretations, this might not pose a significant issue because we applied a multiple regression-approach in our analysis of differential effectiveness. This approach should correct for bias by controlling for the shared positive covariance among the predictor variables that indicate the different instructional foci.

In addition to response biases, it is conceivable that the teachers differ in their understanding of what is meant by terms like advanced reading abilities. However, in our assessments we labeled the instructional foci with simple descriptions that were in accordance with the teachers' official instructional guidelines and curricula. Consequently, it should be rather unlikely that the positive correlations of the instructional foci are largely caused by linguistic ambiguities. To further explore and control for potential response biases, future

studies with larger samples of teachers could specify a multilevel mixture in the model (see Flunger et al., 2021; Vermunt, 2008). This would allow for better differentiation of general effects of a specific focus versus teachers' specific implementations.

**Future Directions**

Building upon our findings, an important next step would be to test the more general applicability of our person-centered approach in areas beyond reading. This would show whether profiles of multivariate aptitudes attained via person-centered analysis possess the general potential to reveal more about the differential effectiveness of treatments beyond variable-centered approaches. If our approach fails in other areas where differential effectiveness might be expected, then the present findings would still be valuable but would have to be seen as a peculiarity of reading instruction. As a related next step, it might be informative to broaden the range of learner characteristics that make up the multivariate learner model, including affective/motivational dispositions as well as personality traits (Ackerman, 2003). Examples of multivariate constructs that might lend themselves to building aptitudes and modeling differential effectiveness in combination with the present approach include different facets of working memory (Oberauer et al., 2003) and executive functions (Miyake et al., 2000), self-regulation (Grunschel et al., 2013) and classroom engagement (Reinhold et al., 2020), affective/motivational variables such as different kinds of goal orientations (Wolters, 2004) and need satisfaction (Ratelle & Duchesne, 2017), and epistemic beliefs (Schiefer et al., 2022). As a general consideration, we recommend considering the conceptualization and modeling of aptitudes that encompass different constructs. For example, learners' prior knowledge and motivational aspects might interact with each other in determining the effectiveness of educational interventions, potentially in further interaction with specific cognitive skills. In addition to broadening the scope of aptitudes in these regards, similar

approaches could be taken for outcome variables. The recent study by Grimm et al. (2021) demonstrates that, in the investigation of differential effectiveness, latent profile analysis also lends itself well to the modeling of multivariate learning outcomes. In a similar manner to modeling students' aptitudes, latent profile analysis could be used to model patterns in outcomes and encompass additional outcome variables besides learning gains, as well as variables associated with different constructs.

Another promising extension of the current approach would be the repeated assessment of indicator variables. This would allow for a more dynamic conceptualization of aptitude profiles and their interaction with specific kinds of instruction. Reinhold et al. (2020) demonstrated, for example, how process data can be used to build profiles of students with different patterns of engagement. Such an approach would also enable an investigation of whether teachers adapt their instruction over the school year to changing learner prerequisites. This could be further extended by relaying information about the multivariate aptitude profiles of their students to teachers (either in a dynamic way via formative assessment procedures or even just single measurement point data) and observing if teachers actually adapt their instruction based on that information and if these adaptations translate to improved learning gains in their students. Over one school year, the assessed multivariate learner prerequisites may change during and in interaction with the learning process. We only assessed them once at the beginning of the school year, leading to a potential mismatch when learners make rapid gains in one of these areas in the first few weeks or months of instruction. A more dynamic measurement approach would allow for better differentiation of these effects, as well as a better understanding of the temporal dynamics behind them (Tetzlaff et al., 2021).

**Conclusions**

In this study, we were able to show that profiles of multivariate aptitudes can be used to explain the differential effectiveness of treatments above and beyond univariate

conceptualizations, at least in the domain of reading. The person-centered approach circumvents the exorbitant power requirements and interpretational complexity involved in analyzing higher-order interactions in variable-centered multiple regression models. The differential effectiveness of instructional parameters that do not show a significant main effect across all learners suggests that those parameters need to be selectively adapted to specific learners. Taken together, these findings strengthen the claim that the simple view of reading model can serve as a basis for informing reading instruction. Our analytic approach appears promising for identifying differential effectiveness, potentially providing a way to overcome the long-standing methodological bottleneck in this area across a variety of educational domains.

## References

Ackerman, P. L. (2003). Aptitude complexes and trait complexes. *Educational Psychologist*, *38*(2), 85–93. https://doi.org/10.1207/S15326985EP3802_3

Alt, M., Fox, A., Levy, R., Hogan, T. P., Cowan, N., & Gray, S. (2021). Phonological working memory and central executive function differ in children with typical development and dyslexia. *Dyslexia*. https://doi.org/10.1002/DYS.1699

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary secondary model. *Mplus web notes*, *21*(2), 1-22.

Bauer, D., In, M. S.-M. contextual effects, & 2007, U. (n.d.). Modeling complex interactions: Person-centered and variable-centered approaches.

Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Bracht, G. H. (1970). Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, *40*(5), 627–645. https://doi.org/10.2307/1169460

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies. - PsycNET*. https://psycnet.apa.org/record/1993-97611-000

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5-51.

Connor, C. M. D., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science*, *315*(5811), 464–465. https://doi.org/10.1126/science.1134513

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., Underwood, P., & Morrison, F. (2009). Individualizing student instruction precisely: Effects of child x instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100.

Constas, M. A., & Sternberg, R. J. (2013). Translating theory and research into educational practice: Developments in content domains, large-scale reform, and intellectual capacity. In *Translating Theory and Research Into Educational Practice: Developments in Content Domains, Large-Scale Reform, and Intellectual Capacity*. Taylor and Francis. https://doi.org/10.4324/9780203726556

Corno, L. (2008). On Teaching Adaptively. *Educational Psychologist*, *43*(3), 161–173.

https://doi.org/10.1080/00461520802178466

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116–127. https://doi.org/10.1037/h0076829

Decristan, J., Fauth, B., Kunter, M., Büttner, G., & Klieme, E. (2017). The interplay between class heterogeneity and teaching quality in primary school. *International Journal of Educational Research*, *86*, 109-121.

Deno, S. L. (1990). Individual Differences and Individual Difference. *The Journal of Special Education*, *24*(2), 160–173. https://doi.org/10.1177/002246699002400205

Dockterman, D. (2018). Insights from 200+ years of personalized learning. *Npj Science of Learning*, *3*(1), 1–6. https://doi.org/10.1038/s41539-018-0033-x

Duff, D., Bruce Tomblin, J., & Catts, H. (2015). The Influence of Reading on Vocabulary Growth: A Case for a Matthew Effect. *Journal of Speech, Language, and Hearing Research*, *58*(3), 853–864. https://doi.org/10.1044/2015_JSLHR-L-13-0310

Edelsbrunner, P., Flaig, M., & Schneider, M. (2021). A Simulation Study on Latent Transition Analysis for Examining Profiles and Trajectories in Education: Recommendations for Fit Statistics.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The Impact of Vocabulary Instruction on Passage-Level Comprehension of School-Age Children: A Meta-Analysis. *Http://Dx.Doi.Org/10.1080/19345740802539200*, *2*(1), 1–44. https://doi.org/10.1080/19345740802539200

Ferguson, S. L., G. Moore, E. W., & Hull, D. M. (2020). Finding latent groups in observed data: A primer on latent profile analysis in Mplus for applied researchers. *International Journal of Behavioral Development*, *44*(5), 458-468.

Flunger, B., Trautwein, U., Nagengast, B., Lüdtke, O., Niggli, A., & Schnyder, I. (2021). Using multilevel mixture models in educational research: An illustration with homework research. *The Journal of Experimental Education*, *89*(1), 209-236.

Foorman, B. R., Petscher, Y., Stanley, C., & Truckenmiller, A. (2017). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of Research on Educational Effectiveness*, *10*(3), 619-645.

Freebody, P., & Anderson, R. C. (1983). Effects on text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Literacy Research*, *15*(3), 19–39.

https://doi.org/10.1080/10862968309547487

Fuchs, L. S., Schumacher, R. F., Sterba, S. K., Long, J., Namkung, J., Malone, A., Hamlett, C. L., Jordan, N. C., Gersten, R., Siegler, R. S., & Changas, P. (2014). Does working memory moderate the effects of fraction intervention? An aptitude-treatment interaction. *Journal of Educational Psychology*. https://doi.org/10.1037/a0034341

Galuschka, K., Ise, E., Krick, K., & Schulte-Körne, G. (2014). Effectiveness of Treatment Approaches for Children and Adolescents with Reading Disabilities: A Meta-Analysis of Randomized Controlled Trials. *PLOS ONE*, *9*(2), e89900. https://doi.org/10.1371/JOURNAL.PONE.0089900

Glück, C. (2011). Wortschatz-und Wortfindungstest für sechs-bis zehnjährige (WWT 6-10). *Aufl München: Elsevier*.

Grimm. H., Edelsbrunner P. A., & Moeller, K. (2021). Accommodating Heterogeneity: The Interaction of Instructional Scaffolding with Student Preconditions in the Learning of Hypothesis-Based Reasoning. PsyArXiv Preprint, available from https://psyarxiv.com/sn9c3/

Guthrie, J. T., Wigfield, A., Metsala, J. L., & Cox, K. E. (1999). Motivational and Cognitive Predictors of Text Comprehension and Reading Amount. *Scientific Studies of Reading*, *3*(3), 231–256. https://doi.org/10.1207/s1532799xssr0303_3

Grunschel, C., Patrzek, J., & Fries, S. (2013). Exploring different types of academic delayers: A latent profile analysis. *Learning and Individual Differences*, *23*, 225-233.

Hallquist, M. N., & Wiley, J. F. (2018). *MplusAutomation* : An R Package for Facilitating Large-Scale Latent Variable Analyses in M *plus*. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Harring, J. R., & Hodis, F. A. (2016). Mixture modeling: Applications in educational psychology. *Educational Psychologist*, *51*(3-4), 354-367.

Hessian Ministry of Education (2021). *Rahmenplan Grundschule*. https://grundschule.bildung.hessen.de/rahmenplan/Rahmenplan.pdf

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences*, *66*, 4–15. https://doi.org/10.1016/J.LINDIF.2017.11.001

Hjetland, H. N., Brinchmann, E. I., Scherer, R., Hulme, C., & Melby-Lervåg, M. (2020). Preschool pathways to reading comprehension: A systematic meta-analytic

review. *Educational Research Review*, *30*, 100323.

Hofer, S. I., Schumacher, R., Rubin, H., & Stern, E. (2018). Enhancing physics learning with cognitively activating instruction: A quasi-experimental classroom intervention study. *Journal of Educational Psychology*, *110*(8), 1175–1191. https://doi.org/10.1037/EDU0000266

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

Hoover, W. A., & Tunmer, W. E. (2018). *Features of the Simple View of Reading (SVR)*. https://doi.org/10.1177/0741932518773154

Hunt, D. E. (1975). Person-environment interaction: A challenge found wanting before it was tried. *Review of Educational Research*, *45*(2), 209–230. https://doi.org/10.2307/1170054

Jiang, D., Kalyuga, S., & Sweller, J. (2018). The Curious Case of Improving Foreign Language Listening Skills by Reading Rather than Listening: An Expertise Reversal Effect. *Educational Psychology Review*, *30*(3), 1139-1165 (27 Seiten). http://dx.doi.org/10.1007/s10648-017-9427-1

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kazak, A. E. (2018). Journal article reporting standards.

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading Comprehension: Core Components and Processes. *Https://Doi.Org/10.1177/2372732215624707*, *3*(1), 62–69. https://doi.org/10.1177/2372732215624707

Kopcha, T. J., & Sullivan, H. (2007). Self-presentation bias in surveys of teachers' educational technology practices. *Educational Technology Research and Development*, *55*(6), 627–646. https://doi.org/10.1007/s11423-006-9011-8

Kuşdemir, Y., & Bulut, P. (2018). The Relationship between Elementary School Students' Reading Comprehension and Reading Motivation. *Journal of Education and Training Studies*, *6*(12), 97–110. https://doi.org/10.11114/jets.v6i12.3595

Lenhard, W., Erst-bis, W. S.-E. L. für, & 2006, U. (n.d.). ELFE 1-6. *Hf.Uni-Koeln.De*. Retrieved December 21, 2021, from https://www.hf.uni-koeln.de/data/lernwerkstatt/File/Material des Monats/2014-12 Material des Monats ELFE.pdf

Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the Number of Components in a Normal Mixture. *Biometrika, 88*(3), 767-778. Retrieved August 25, 2021, from http://www.jstor.org/stable/2673445

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*(1), 29–45. https://doi.org/10.3102/01623737021001029

Menghini, D., Finzi, A., Carlesimo, G. A., & Vicari, S. (2011). Working memory impairment in children with developmental dyslexia: Is it just a phonological deficity? *Developmental Neuropsychology*, *36*(2), 199–213. https://doi.org/10.1080/87565641.2010.549868

Mezynski, K. (1983). Issues Concerning the Acquisition of Knowledge: Effects of Vocabulary Training on Reading Comprehension: *Http://Dx.Doi.Org/10.3102/00346543053002253*, *53*(2), 253–279. https://doi.org/10.3102/00346543053002253

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, *41*(1), 49-100.

Muthén, B., & Muthén, L. (2017). *Mplus* (pp. 507-518). Chapman and Hall/CRC.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Https://Doi.Org/10.1080/10705510701575396*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, *31*(2), 167–193.

Preacher, K. J., & Sterba, S. K. (2019). Aptitude-by-treatment interactions in research on educational interventions. *Exceptional Children*, *85*(2), 248-264.

Ratelle, C. F., & Duchesne, S. (2014). Trajectories of psychological need satisfaction from early to late adolescence as a predictor of adjustment in school. *Contemporary Educational Psychology*, *39*(4), 388-400.

Reinhold, F., Strohmaier, A., Hoch, S., Reiss, K., Böheim, R., & Seidel, T. (2020). *Process data from electronic textbooks indicate students' classroom engagement*. https://doi.org/10.1016/j.lindif.2020.101934

Retelsdorf, J., Köller, O., & Möller, J. (2011). On the effects of motivation on reading performance growth in secondary school. *Learning and Instruction*, *21*(4), 550–559. https://doi.org/10.1016/J.LEARNINSTRUC.2010.11.001

Schiefele, U., Stutz, F., & Schaffner, E. (2016). Longitudinal relations between reading motivation and reading comprehension in the early elementary grades. *Learning and Individual Differences*, *51*, 49–58. https://doi.org/10.1016/J.LINDIF.2016.08.031

Schiefer, J., Edelsbrunner, P. A., Bernholt, A., Kampa, N., & Nehring, A. (2022). Profiles of Epistemic Beliefs in Science: An Integration of Evidence from Multiple Studies. *PsyArXiv*.

Segers, E., & Verhoeven, L. (2003). Effects of vocabulary training by computer in kindergarten. *Journal of Computer Assisted Learning*, *19*(4), 557–566. https://doi.org/10.1046/J.0266-4909.2003.00058.X

Seufert, T., Schütze, M., & Brünken, R. (2009). Memory characteristics and modality in multimedia learning: An aptitude-treatment-interaction study (PSYNDEXshort). *Learning and Instruction*, *19*(1), 28–42. https://doi.org/10.1016/j.learninstruc.2008.01.002

Share, D. L. (1999). Phonological recoding and orthographic learning: A direct test of the self-teaching hypothesis. *Journal of Experimental Child Psychology*, *72*(2), 95-129. https://doi.org/10.1006/jecp.1998.2481

Siegmüller, J., & Kauschke, V. M. S., & Bittner, D.(2011). *Test zum Satzverstehen von Kindern. Eine profilorientierte Diagnostik der Syntax*.

Singer, M. H., & Crouse, J. (1981). The Relationship of Context-Use Skills to Reading: A Case for an Alternative Experimental Logic. *Child Development*, *52*(4), 1326. https://doi.org/10.2307/1129525

Smith-Spark, J. H., & Fisk, J. E. (2007). Working memory functioning in developmental dyslexia. *Http://Dx.Doi.Org/10.1080/09658210601043384*, *15*(1), 34–56. https://doi.org/10.1080/09658210601043384

Snow, R. E., Farr, M. J., United States. Office of Naval Research., & Navy Personnel Research and Development Center (U.S.). (1987). *Aptitude Complexes*. 11–34. https://doi.org/10.4324/9781003163244-2

Subban, P. (2006). Differentiated instruction: a research basis. International Education Journal, 7(7), 935–947

Suk, N. (2017). The Effects of Extensive Reading on Reading Comprehension, Reading Rate, and Vocabulary Acquisition. *Reading Research Quarterly*, *52*(1), 73–89. https://doi.org/10.1002/RRQ.152

Team, R. C. (2021). R: A Language and Environment for Statistical Computing, v. 4.0. 2 (R Foundation for Statistical Computing, Vienna, Austria, 2020). *Google Scholar There is no corresponding record for this reference*.

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing Personalized Education: A Dynamic Framework. *Educational Psychology Review*, *33*(3), 863–882. https://doi.org/10.1007/S10648-020-09570-W/FIGURES/3

Tilstra, J., McMaster, K., Van Den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: components of the simple view of reading across grade levels. *Journal of Research in Reading*, *32*(4), 383–401. https://doi.org/10.1111/J.1467-9817.2009.01401.X

Torppa, M., Tolvanen, A., Poikkeus, A. M., Eklund, K., Lerkkanen, M. K., Leskinen, E., & Lyytinen, H. (2007). Reading development subtypes and their early characteristics. *Annals of dyslexia*, *57*(1), 3-32.

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

vanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. In *Educational Psychologist* (Vol. 46, Issue 4, pp. 197–221). https://doi.org/10.1080/00461520.2011.611369

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary Growth and Reading Development across the Elementary School Years. *Https://Doi.Org/10.1080/10888438.2011.536125*, *15*(1), 8–25. https://doi.org/10.1080/10888438.2011.536125

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical methods in medical research*, *17*(1), 33-51.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450-469.

Wolff, U. (2010). Subgrouping of readers based on performance measures: A latent profile analysis. *Reading and Writing*, *23*(2), 209–238. https://doi.org/10.1007/S11145-008-9160-8/FIGURES/7

Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of educational psychology*, *96*(2), 236.

Wubbels, T., Brekelmans, M., & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education*, *8*(1), 47–58.

https://doi.org/10.1016/0742-051X(92)90039-6

Ziegler, E., Edelsbrunner, P. A., & Stern, E. (2020). The benefit of combining teacher-direction with contrasted presentation of algebra principles. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-020-00468-3

Appendix:

**Table A1**

*Differential Effects of Specific Instructional Foci for the Different Learner Profiles*

| | Estimate | Standard Error | Lower 5% | Upper 5% |
|---|---|---|---|---|
| Poor Decoders | | | | |
| Vocabulary | -0.222 | 0.106 | -0.396 | -0.047 |
| Motivation | -0.064 | 0.120 | -0.262 | 0.134 |
| Adv. Reading | -0.138 | 0.117 | -0.330 | 0.055 |
| Poor Comprehenders | | | | |
| Vocabulary | -0.210 | 0.108 | -0.388 | -0.033 |
| Motivation | 0.104 | 0.122 | -0.096 | 0.304 |
| Adv. Reading | 0.223 | 0.102 | 0.055 | 0.390 |
| Poor Readers | | | | |
| Vocabulary | 0.672 | 1.152 | -1.223 | 2.568 |
| Motivation | -0.449 | 0.988 | -2.074 | 1.176 |
| Adv. Reading | 0.566 | 1.054 | -1.168 | 2.300 |
| Good Readers | | | | |
| Vocabulary | 0.331 | 0.109 | 0.150 | 0.511 |
| Motivation | 0.025 | 0.132 | -0.191 | 0.241 |
| Adv. Reading | -0.344 | 0.117 | -0.536 | -0.151 |

***Paper 3***

**Tetzlaff, L.,** Hartmann, U., Dumont, H., & Brod, G.: *Assessing Individualized Instruction in the Classroom: Comparing Teacher, Student and Observer Perspectives.* Manuscript revised and resubmitted at *Learning & Instruction*.

# Assessing individualized instruction in the classroom: Comparing teacher, student, and observer perspectives

Leonard Tetzlaff[a], Ulrike Hartmann[a,b], Hanna Dumont[a,c] & Garvin Brod[a,d]

[a]DIPF | Leibniz Institute for Research and Information in Education

Rostocker Straße 6, 60323 Frankfurt am Main, Germany

[b]Bergische Universität Wuppertal

Gaußstraße 20, 42119 Wuppertal, Germany

[c]Universität Potsdam

Am Neuen Palais 10, 14469 Potsdam, Germany

[d]Johann Wolfgang Goethe-Universität Frankfurt

Theodor-W.-Adorno-Platz 1, 60323 Frankfurt am Main, Germany

Correspondence concerning this article should be addressed to Leonard Tetzlaff,

DIPF | Leibniz Institute for Research and Information in Education,

Rostocker Straße 6, 60323 Frankfurt, email: tetzlaff@dipf.de

**Abstract**

In this article, we address the measurement of individualized instruction in the context of regular classroom instruction. Our study assessed individualized instruction in German third grade reading lessons by combining self-report data from 621 students and their teachers ($N$ = 57) with live observations. We then investigated the reliability of these different approaches to measuring individualization as well as the agreement between them. All three approaches yielded reliable indicators of individualization, but not all of them corresponded with each other. We found considerable agreement between students and observers, but neither agreed with teachers' self-reports. Upon closer examination, we found that students' ratings only correlated with teacher ratings that were provided close to the timepoint of interest. This correlation increased when teacher measures were corrected for response tendencies. We conclude with some recommendations for future studies that aim to measure individualization in the classroom.

Keywords: Individualization, Personalization, Differentiation, Adaptive Teaching, Individualized Education, Instructional Quality, Learning Environments, Live Observations, Classroom Research

## 1. Introduction

One of the most pressing issues in educational research and practice has been the individualization of classroom instruction. For more than 200 years, educators, as well as scientists and policymakers, have been calling for a shift away from one-size-fits-all whole class instruction towards a more individualized and personalized approach (Dockterman, 2018). Especially in times of growing heterogeneity (Markic & Abels, 2014), individualized instruction is seen as a prerequisite for each student to receive instruction situated within his or her zone of proximal development (Vygotsky, 1930-1944/1978).

This need contrasts starkly with the dearth of empirical studies investigating individualized instruction in a regular classroom context. While there are plenty of studies dealing with the concept of individualized instruction, only a few of them focus on regular classroom instruction (e.g. Förster et al., 2018; Hachfeld & Lazarides, 2020; Suprayogi et al., 2017))—a recent systematic review (Bernacki et al., 2021) found that 80% of investigated studies examined some form of technology-based learning. Even in formative assessment studies—a prominent approach to classroom-based individualized instruction— the specific instructional adaptations are usually left up to the practitioners, providing almost no information on how teachers individualize and adapt their instruction to individual students' needs (Förster & Souvignier, 2014; Jung et al., 2018; Kingston & Nash, 2011). Despite these gaps in the current knowledge base, there is emerging evidence that the introduction of interventions that facilitate individualization is generally associated with positive student outcomes (Connor et al., 2009, 2018; Jung et al., 2018; Stecker et al., 2005; Waxman et al., 1985).

In order to improve our understanding of individualization and its effects on learning in the context of regular classroom instruction, the amount and type of individualized practices need to be reliably measured. We believe that the discrepancy between the ubiquitous demand

and the lack of concrete operationalizations stems at least in part from the difficulties associated with assessing individualized education in a classroom context. This article aims to address this issue by assessing indicators of individualized education from three different perspectives—students, teachers, and observers—to investigate the reliability of each instrument as well as the agreement among them.

### 1.1 What is Individualized Instruction?

Even though there is wide agreement in the literature that the one-size-fits-all whole class instruction should move to a more individualized instructional approach, this idea is reflected in various different terms and concepts. “Individualized instruction” (e.g. Connor et al., 2018), “individualization” (e.g. Hachfeld & Lazarides, 2020), “personalized learning” (e.g. Bernacki et al., 2021; Dockterman, 2018), “personalization” (e.g. Daruwala et al., 2021), “personalized education” (e.g. Tetzlaff et al., 2021), “differentiated instruction” (e.g. Bondie et al., 2019; van Geel et al., 2019), “differentiation” (e.g. Deunk et al., 2018; Prast et al., 2018), “adaptive teaching” (e.g. Corno, 2008; Hardy et al., 2019; Vaughn et al., 2021), “instructional adaptations” (e.g. Parsons et al., 2018) are among the most common found in the literature. When taking a closer look at the definitions for the different terms provided, it is almost impossible to arrive at conceptual clarity and clearly differentiate them from each other as the same term is sometimes used for different concepts and different terms are sometimes used for the same concept. The different terms seem to be rather the result of researchers coming from different research traditions and disciplinary backgrounds. For example, whereas the term “personalized learning” is mostly used by researchers focusing on educational technologies and is thus a rather new concept (Bernacki et al., 2021; Daruwala et al., 2021; Major et al., 2021; Roberts-Mahoney et al., 2016). “Individualized instruction” and “differentiated instruction” have been around to describe teaching practices in classrooms for a long time (e.g.

Miller, 1976; Slavin & Karweit, 1985; Stradling & Saunders, 2006), whereas differentiated instruction typically describes the adaptation of instruction to small homogenous groups of learners within a classroom (Bondie et al., 2019).

Despite there not being a consensus regarding which term to use, there is however, a wide consensus in the literature regarding the underlying goal of individualizing, personalizing, differentiating or adapting instruction: to align instruction with the specific characteristics of individual learners in order to better meet the needs of each. For the purpose of this article, we use individualized instruction as an umbrella term to describe such an instructional approach. While individualized instruction may be implemented in the classroom in different ways (Suprayogi et al., 2017), in the present study, we focus on three key aspects: 1) providing different tasks for different students, 2) providing individualized attention (in the form of specific instruction or feedback) to specific students, and 3) providing different amounts of time for different students working on the same task, which can include both extending a time frame for slower students as well as providing faster students with additional material. We believe the first aspect, that is, individualized task assignment, to be at the heart of individualized instruction. Note, that whenever instruction is adapted to smaller groups of students and not individual students, we use the term differentiated instruction.

## 1.2 Measurement of Classroom Processes

Classroom instruction is a complex process that is constituted by an interplay of multiple processes. There are several possible approaches for measuring these processes—which all have their specific advantages and disadvantages—including external observers, student self-reports, and teacher self-reports.

### *1.2.1 Observers*

One possible approach is to observe the entire process in situ assuming that a) the outside observer is able to access the relevant information and b) the specific timeframe of

observation is a good representation of the process in general. Examples of studies employing classroom observations are plentiful, especially in the context of measuring the basic dimensions of teaching quality—classroom management, cognitive activation, and supportive climate (e.g., Bell et al., 2019; Nava et al., 2019; Praetorius et al., 2018), and results indicate that these dimensions can be reliably measured by observing a single lesson (Praetorius et al., 2014).

Classroom observations are especially informative when the construct of interest manifests in directly observable behavior (e.g. individualized task assignment) and can thus be assessed by low-inference ratings (McConnell & Bowers, 1979).

### *1.2.2 Students*

Students have a direct recipient perspective on most instructional processes, but their ratings might be influenced by teacher popularity (Aleamoni, 1999; Greenwald, 1997). Fauth et al. (2018) were able to show, however, that teacher popularity also significantly correlated with instructional quality as measured by outside observers, indicating that those correlations potentially stem from a true relation of these constructs, rather than a bias in the student ratings. Moreover, younger children, in particular, might have a hard time identifying and differentiating between pedagogical constructs. Previous research indicates, however, that student ratings are a good source of information concerning both the suitability of tasks and instructional approaches as well as their difficulty (Fauth et al., 2014; Kunter & Baumert, 2006). Asking all students in a class to provide ratings has the additional advantage of having multiple observers of the same construct, which—given satisfactory agreement—increases the reliability of the aggregated measure (see Lüdtke et al., 2009). Student ratings of individualized instruction are potentially less reliable than those of other classroom processes though, as the amount of individualized instruction single students receive might vary based on the need

perceived by the teachers and students might not be able to correctly identify whether their classmates receive individualized instruction.

### *1.2.3 Teachers*

While teachers also have a direct view on instructional processes, their perspectives differ from those of students in a few key aspects. Most importantly, their ratings might be influenced by self-presentation bias (e.g., Kopcha & Sullivan, 2007)—especially for highly desirable aspects of teaching such as individualized instruction. It is also plausible that they sometimes assess their intentions/ideals instead of their actual behavior (Wubbels et al., 1992). But their expert knowledge on instructional theory also allows them to note aspects that students are unable to classify. Previous research indicates that teacher ratings are a good source of information for the details of instructional processes as well as for a differentiation between related concepts (e.g., task selection and task presentation) that might be hard to distinguish for students (Kunter & Baumert, 2006).

To conclude, all three perspectives—classroom observation, student ratings, and teacher ratings—can offer unique and valuable insights. They might emphasize different aspects of the respective constructs and be influenced by different factors. In the next paragraph, we will discuss previous research on whether these different perspectives converge or diverge when assessing the same construct.

## 1.3 Convergence and Divergence of the Different Perspectives on Classroom Processes

A literature review by Den Brok et al. (2006) found that correlations between teacher and student perspectives are, in general, low to moderate. Additionally, for behavior that is positively related to student outcomes, teachers report, on average, higher ratings than their students. Kunter & Baumert (2006), reporting similar results, concluded that the low to moderate correlations between student and teacher ratings do not necessarily stem from a low reliability of the single measures but rather reflect systematic differences in the constructs—

even when they are assessed with parallel items. This indicates that students and teachers do have some shared perception of pedagogical constructs, but also include some aspects that are unique to their respective roles.

With regard to the agreement between external observers and ratings provided by teachers or students, Fauth et al. (2014) have shown that basic dimensions of teaching quality can be reliably measured both by observation as well as by surveying teachers and students. They also found considerable overlap between the three assessment approaches, indicating that they all measure the same construct, albeit from slightly different perspectives. Scherzinger & Wettstein (2019) found a moderate correlation between observers and students when rating classroom management and virtually no correlation between observers and teachers. Clunies-Ross et al. (2008), on the other hand, found considerable agreement between teachers and observers when assessing the usage of specific classroom management strategies, indicating a possible advantage of framing self-report questions as specifically as possible.

The majority of the above-mentioned studies investigated the basic dimensions of teaching quality, as they are of high importance and clearly defined by specific indicators, making them comparatively easy to measure (Praetorius et al., 2014). Concerning individualized instruction, Fraser (1981) established the ICEQ, an individualization questionnaire with parallel scales for students and teachers and found that teacher and student perspectives correlated moderately (between $r$ =.39 and $r$ =.68 for the different subscales), but also that teachers generally rated their instruction as more individualized than their students did. We are not aware of any studies that use classroom observations to quantitatively measure individualized instruction, let alone compare them to other perspectives.

In summary, while there is a lack of research on individualized instruction specifically, research on other aspects of teaching quality indicates that there is a systematic overlap

between all three perspectives across different constructs. This overlap is usually largest between students and observers and smallest between teachers and observers. The differences between the perspectives are systematic and vary across constructs. Agreement between students and observers depends on the complexity of the construct, and teachers seem to rate their behavior more favorably than students or observers, at least when they are asked about their general behavior.

### 1.4 Challenges in Self-report Measures

A possible explanation for the low to moderate agreement between students and teachers lies in the way their self-reports are usually framed. In self-report studies, a distinction can be made between the believing, remembering, and experiencing self (Conner & Barrett, 2012). The experiencing self is usually assessed via ambulatory assessment—repeated measurements that are collected during a specific activity—and depicts the behavior or experience in the moment the question is asked. The remembering self is usually assessed via retrospective reports—measurements collected at some point after a specific activity—and depicts the memory of the behavior or experience in question. It is thus prone to well-known memory biases such as peak-end effects (Kahneman et al., 1993). The believing self is usually assessed via trait measures—measurements of stable characteristics without references to a specific situation—and depicts participants' beliefs about their general tendency to act or experience things in a certain way. Especially for constructs where it may be assumed that participants have strong beliefs about themselves, assessing behavior via trait measures tends to give results closer to what they believe their behavior is or what they would like it to be (e.g., Houtveen & Oei, 2007; Robinson & Clore, 2002).

In the majority of the above-mentioned studies on agreement, the teacher perspective is assessed via trait measures. It is conceivable that those tap mainly into the believing self, which could lead to an overestimation of teaching behaviors that are positively related to

student outcomes. The student perspective, on the other hand, is usually assessed via retrospective reports and thus taps into the remembering self, given that students are rating explicit past behavior of their teachers and not some enduring trait. We therefore assumed that the agreement between teacher and student raters would be higher when the teacher perspective is also assessed via retrospective reports instead of self-reports involving trait measures.

### 1.5 Research Questions and Hypotheses

This study incorporates multiple approaches to measuring individualization in a classroom context, in order to assess different perspectives and their reliability. Besides the instruments themselves, we were also interested in comparing the different perspectives to explore to what extent and under what conditions they agree with each other. To achieve these goals, we used classroom observations to assess different aspects of individualized instruction -task assignment, attention and time allocation- and then compared the measures of individualized task assignment to self-report measures from the student and teacher perspective as assessed by established instruments. We chose to focus on task assignment as it is clearly referenced in all of the used instruments. We further believe that individualized task assignment is a useful tool for addressing learner variability independent of specific pedagogical traditions.

We posed the following specific research questions: 1) whether student reports, teacher reports, and live observations could reliably assess the occurrence of individualization; 2) whether the assessments of individualized task assignment from these different perspectives correlate with each other; and 3) whether the kind of self-report measure used for the teacher perspective—trait or retrospective—influences its agreement with the other perspectives.

Based on previous research that investigated agreement between assessment methods of other aspects of teaching quality (Brekelmans et al., 2011; Donker et al., 2021; Fauth et al., 2014; Scherzinger & Wettstein, 2019), we expected the highest overlap to occur between

students and observers, less overlap between teachers and students, and only low to moderate overlap between teachers and observers. We also expected retrospective teacher reports to correlate more highly with the other perspectives than teacher trait measures.

## 2. Method

The current study was carried out as part of a larger research project running from 2018–2020. Data were collected in 35 German public elementary schools in two cohorts, one in the school year 18/19, the other in the school year 19/20. Schools were recruited in cooperation with the ministry of education of Hesse, without any special requirements (e.g., location, demographics). Each cohort was presented with a pretest at the beginning of the school year and a posttest before the summer break. Teachers and their students participated in the study on a voluntary basis. Participants did not receive any compensation for their participation. Of the 73 teachers initially recruited, 57 completed the posttest, 33 of whom also took part in our observations (17 in the first cohort, 16 in the second). The 57 teachers were on average 40.14 (SD = 9.37) years old and had an average of 13.32 (SD = 7.78) years of teaching experience. Teachers were asked to provide explicit consent, separately for the questionnaires and for the observations. Not all teachers who provided consent for the questionnaires also agreed to the observations, leading to a reduced sample size for the comparison between the two. The study was approved by the ethics committee of DIPF. For participating students, we requested active consent from the parents and hence excluded all students for whom this consent was not obtained. The student ratings were obtained from between 7 and 25 third grade students per class ($M$ = 10.90), leading to a total of 621 students from 57 classes. Those students were on average 8.32 (SD=0.56) years old and showed average reading comprehension skills, when compared to a norm sample of the same grade level (T = 48.32, SD = 10.70).

## 2.1 Measures

We collected data on individualized instruction in third grade German reading lessons from three different perspectives: observers, students, and teachers. All of the measures we used only deal with the sight-structure of individualized instruction. This means that we assessed directly observable instructional actions without addressing the thought processes that potentially lead to, or were induced by those actions.

### *2.1.1 Observational Measures*

We observed one third grade reading lesson per teacher in the middle of the school year for both cohorts (February/March 2019 and February 2020) with a standardized observation protocol. Lessons were selected based on the schedules of the teachers (e.g. not directly before a test) and the teachers were instructed to conduct a "typical" lesson. They received no additional information on the goal of the study and did not know anything about the focus of the observations.

For every aspect, observers rated whether the behavior occurred at least once during the lesson, or not. The three observed aspects of individualized instruction were:

1) the individualization of tasks, which was assessed at two levels:

    a) the assignment of different tasks to different groups of students: this code was applied whenever at least one subgroup of the class was working on a different task

    b) the individualized selection of tasks for specific students: this code was applied in addition to 1a whenever an individual student was working on a task that no one else was working on

2) the individualization of provided time, which was captured via two different indicators:

    a) the provision of additional time for students that were slower than the class average and were not able to finish the group task within the allocated time

b) the provision of additional tasks for students who were faster than the class average and finished the group task early

3) the individualization of attention, which was captured via two different indicators:

a) the allocation of attention to a specific subgroup of students (e.g., a teacher talks to a group of students while the others are working on a written task)

b) the allocation of attention to an individual student: we coded this when there was at least a short conversation between teacher and student(s) that had some relation to the subject matter (in contrast to organizational conversations such as a student asking to go to the restroom).

The six raters consisted of student assistants who were trained by Author 1 and 2 for about 18 hours (over six sessions) that took place over the course of two weeks. The training consisted of watching videos of teaching situations in German elementary school classrooms (taken from the VERA study (10.7477/20:1:1)). Raters rated the videos independently using the standardized form. Differences in ratings were discussed afterwards**.**

Each lesson was observed in situ by two independent raters who both filled out the entire form. Immediately after the lesson, they discussed any differences in their ratings and filled out a third, unified form. All further analyses were performed with this agreed-upon form. The two original forms were used to calculate interrater reliability. The allocation of the raters to the lessons was conducted based on the availability of the raters, i.e., the lessons were rated by changing pairs of raters and not by fixed teams.

### *2.1.2 Teacher Reports*

We used two different measurement approaches to capture the teachers' perspective—a trait assessment at the end of the school year and a short online questionnaire that was repeated every three weeks. These can be seen as corresponding to the believing and remembering self, respectively (Conner & Barrett, 2012).

At the end of the school year, we used a translation of the Differentiated Instruction scale from the DSAQ (Differentiated Self-Assessment Questionnaire; Prast et al., 2015) to obtain summative information on teachers' individualization practices during the school year. The differentiated instruction scale comprises seven items that detail possible instructional adaptations e.g.: "I regularly provide high-achieving students with additional instruction or guidance at their level." All of these items are answered on 4-point Likert scales ranging from *I don't agree at all* to *I agree completely*. These questions all target regular or general behaviour and can thus be conceptualized as trait measures. While the DSAQ was originally designed for mathematics lessons, the selected items are all worded in a general way that is applicable to lessons independent of subject.

The online questionnaire was made available eight times throughout the school year. Teachers were asked to provide details concerning their teaching practice during the previous three weeks. This assessment included information concerning the distribution of learning time to specific training of vocabulary, basic reading skills, advanced reading skills, reading motivation, and reading precursor abilities for the whole class. We also gave the teachers the option to indicate how much time was spent with different students working on different tasks. This was assessed with a single item: "How often have you given different tasks to different children during the last 3 weeks?" and was used as an indicator for within-class differentiation as experienced by the teacher. All of these questions were answered on a 4-point Likert scale with answers ranging from *never* to *often*. As the questions target specific past behavior, they can be understood as retrospective measures. The online questionnaire was presented on the REDCap platform (Harris et al., 2009). Teachers received a link via email every three weeks. If they did not fill out the questionnaire one week after receiving the link, a reminder email was

sent. 42 teachers participated in the repeated online questionnaires and, on average, filled out 4.10 of the 8 questionnaires (range from 1 to 8).

### *2.1.3 Student Reports.*

To assess the student perspective on teachers' differentiated instruction, we used a short self-report questionnaire (Dumont, 2016). This questionnaire was presented at the end of the school year. Students made a retrospective judgement of the amount of differentiated instructional practice they had received during the year. The 9 items were completed during regular classroom hours under the supervision of our test administrators. Example items included: "In our class, everyone is always working on the same tasks" and "In our class, everyone gets the same tasks as homework." Answers were given on a 4-point Likert scale with the anchors *definitely correct* and *not at all correct*.

## 3. Analyses

All analyses were computed with R (Version 4.0.2) and Mplus (Version 8).

For the observer ratings, we calculated the interrater reliability between the two original forms by using Cohen's kappa (Cohen, 1968). For the teacher trait measurements, we used sum-scores to aggregate the items from the DSAQ instruction scale and computed Cronbach's alpha for the internal consistency. We decided against a latent variable model because we believe the measurement model to be formative rather than reflective. In a formative model, different indicators additively form the construct together, without necessarily being related to each other. For an in-depth discussion on formative models and how to assess their reliability, see (Coltman et al., 2008; Diamantopoulos et al., 2008).

For the repeated online questionnaires, we calculated ICC(1) and ICC(2) to determine the amount of variance that can be attributed to the teacher/stability over time. The ICC(1) specifies the amount of variance that can be explained by group membership (in this case:

teacher), while the ICC(2) is a measure for the reliability of the mean rating, taking group size into account (Bliese, 2000).

We removed one item from the student questionnaire ("all children have the same subject at the same time") because all children were taught the same subject at the same time in our sample. We used sum-scores to aggregate the data from each single student and afterwards aggregated the scores from all students in a specific teacher's class using mean scores. We used Cronbach's alpha to determine the internal consistency of the scale and we again calculated ICC(1) and ICC(2) to determine the reliability of the class aggregate. To investigate the agreement between the different perspectives, we used the indicator 1a)—provision of different tasks to different students—on the observation sheet because it corresponds best to the construct as it is operationalized in the student and teacher self-report scales. This is potentially due to task assignment being a directly observable action with clear indicators – both for teachers and for students.

## 4. Results

### 4.1 Reliability of the Different Measures

Before comparing the different perspectives, we first tested whether they were assessed in a reliable way.

#### *4.1.1 Observer Ratings*

As all observed constructs were operationalized via low-inference ratings (see McConnell & Bowers, 1979), we were able to attain a high level of objectivity and reliability. We used Cohen's kappa (Cohen, 1968) to measure the interrater reliability of our observations. As a Cohen's kappa above .6 is generally considered adequate agreement (McHugh, 2012), our attained values from .66 to .90 (see Table 1) suggest that our observational measures were

reliable. We then calculated the means and variances of each variable across all teachers. These results can be found in Table 1.

**Table 1**

*Means, variances, and reliabilities of observational measurements of individualized instruction*

| | Mean | SD | Range | Reliability (Cohen's kappa) |
|---|---|---|---|---|
| **Individualized Tasks 1** | .26 | .36 | 0-1 | .79 |
| **Individualized Tasks 2** | .10 | .24 | 0-1 | .89 |
| **Attention - Group** | .48 | .31 | 0-1 | .87 |
| **Attention - Individual** | .60 | .31 | 0-1 | .90 |
| **Time – wait for slow** | .32 | .33 | 0-1 | .89 |
| **Time – engage fast** | .12 | .20 | 0-1 | .66 |

### *4.1.2 Teacher*

The items in the trait teacher questionnaire showed a similarly moderate internal consistency in our sample ($\alpha = 0.68$) as in the original study ($\alpha = .72$) (E.J. Prast et al., 2015). We believe that individualized instruction should best be modelled as a formative construct, which means that a high internal consistency is neither required nor to be expected. Concerning the reliability of the aggregated repeated measurements, we again looked at the ICC to determine the relation of variance within and between teachers. The ICC(1) of 0.40 and ICC(2) of 0.73 imply that these scores measure constructs at the teacher level and can be reliably aggregated over multiple timepoints. The means and variances of these items across all teachers and averaged over all timepoints can be found in Table 2.

**Table 2**

*Means, variances, and reliabilities of self-report measures for individualized instruction at the classroom level*

| | mean | SD | Range | Reliability |
|---|---|---|---|---|
| **Teacher (repeated retrospective)** | 3.28 | 0.62 | 1-4 | ICC(1) = 0.40<br>ICC(2) = 0.73 |
| **Teacher (trait)** | 3.35 | 0.35 | 1-4 | α = 0.68 |
| **Student** | 2.22 | 0.43 | 1-4 | α = 0.72<br>ICC(1) = 0.42<br>ICC(2) = 0.88 |

### *4.1.3 Student*

The items in the student questionnaire showed satisfactory internal consistency (α = 0.72), indicating that they indeed measured the same construct. The students in the same class also showed a high relative agreement with ICC(1) = 0.42 and ICC(2) = 0.88. Thus, the ICC(1) shows that these scores actually measure constructs at the teacher level, as opposed to individual student experience, while the ICC(2) shows that aggregating those scores results in reliable measures. According to Fleiss (1986), an ICC(1) of >.15 and an ICC(2) of >.75 indicate that the data can be reliably aggregated. Both of these values are necessary prerequisites for the correlational analyses with the other variables, which are also assessed at the teacher level. Descriptive results of these aggregated measures can be found in Table 2.

## 4.2 Correcting for Response Tendencies

As this study was situated within a larger research project, the repeated teacher questionnaire also contained questions concerning which aspect of reading instruction teachers emphasized in a given timeframe. While their content is not relevant for this study, the answering patterns of the teachers exhibited a peculiarity that provided some insight into potential response tendencies (see section 4.2).

Upon examining the correlations between items of the repeated teacher questionnaire, we noticed that the different instructional parameters were consistently positively correlated (see Table 3). As the items measure the amount of time allocated to different aspects of whole-class reading instruction, however, negative correlations between items should be expected. In a limited timeframe, more time allocated to a specific aspect of teaching should indicate less time allocated to the others. We thus believe that these correlations stem from a general tendency to answer the questions in a more favorable way rather than a genuine relatedness of the underlying constructs. If this is the case, any variance that is shared by all of them should be uniquely attributable to this common response tendency. Under this assumption, we specified a reflective measurement model in which all Likert-scale items from the questionnaire are loading on a common factor. This resulted in an acceptable model fit (CFI = .95; RMSEA = .073) with factor loadings from .49 (for individualized tasks) to .90 (for training of precursor abilities). We understand the residuals in this model to approximate the true score of a teacher on that specific variable, untainted by the common response tendency (but still containing measurement error). We used those residuals in addition to the original scores (raw) on that item for the correlational analysis. Figure 1 depicts the structural model underlying these measurements.

**Table 3**

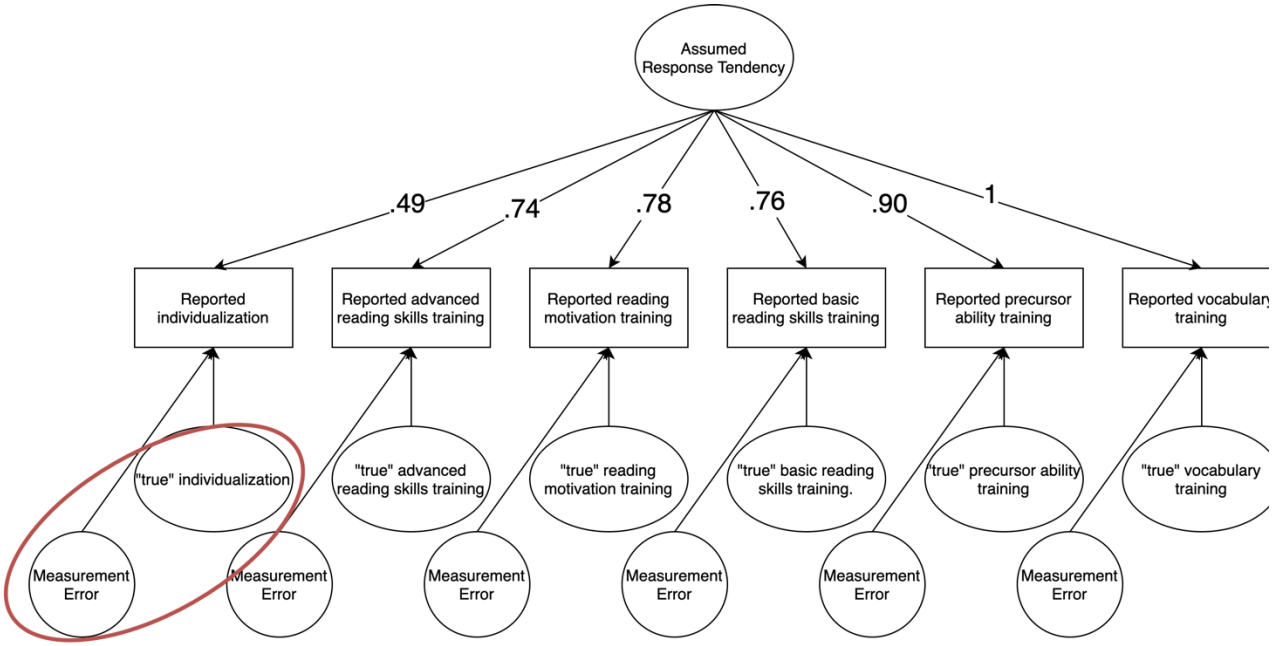
*Intercorrelations in the repeated teacher questionnaire*

| | Vocabulary | Advanced Reading Skills | Basic Reading Skills | Precursor Abilities | Reading Motivation | Individualized Tasks |
|---|---|---|---|---|---|---|
| **Vocabulary** | 1 | **.38** | **.31** | **.48** | **.55** | **.30** |
| **Advanced Reading Skills** | | 1 | **.48** | **.30** | **.51** | **.38** |
| **Basic Reading Skills** | | | 1 | **.44** | **.40** | .15 |
| **Precursor Abilities** | | | | 1 | **.45** | **.28** |
| **Reading Motivation** | | | | | 1 | .19 |
| **Individualized Tasks** | | | | | | 1 |

*Note.* Correlations with $p < .05$ are depicted in bold.

**Figure 1**

*Measurement model for the assumed response tendency*




*Note.* The manifest variables from the retrospective teacher questionnaire are all loading on the assumed response tendency factor. The red ellipse depicts the residual we use for the correlational analyses.

### 4.3 Agreement Between Perspectives

We computed the Pearson product-moment correlations between the indicators of individualized instruction for the different perspectives (student, teacher, and observer) and tested them using t-tests. As can be seen in Table 4, there is substantial agreement ($r$ = .43, $p$ = .01) between students and observers but no agreement between teachers and observers. This pattern is slightly more complex for the agreement between teachers and students. While the trait teacher questionnaire was uncorrelated with students' ratings, the repeated teacher self-reports did correlate with the student perspective ($r$ =.38 $p$ = .01), especially when controlled for response tendencies ($r$ = .49, $p$ < .01). This pattern suggests that the type of self-report measure used is an important factor that influences agreement. Finally, reports from the teacher trait questionnaire were significantly correlated with the uncontrolled repeated measurements

($r$ = .39, $p$ = .01), providing further support for the assumption that those trait measures mainly capture an idealized self-image of the teachers.

**Table 4**

*Correlations – individualized tasks*

| | **Observer** | **Teacher - Trait** | **Teacher – retrospective (raw)** | **Teacher – retrospective (corrected)** | **Student** |
|---|---|---|---|---|---|
| **Observer** | 1 | .06 | -.07 | .16 | **.43** |
| **Teacher - Trait** | | 1 | **.39** | .32 | .03 |
| **Teacher – retrospective (raw)** | | | 1 | **.95** | **.38** |
| **Teacher – retrospective (corrected)** | | | | 1 | **.49** |
| **Student** | | | | | 1 |

*Note.* Correlations with $p < .05$ are depicted in bold.

## 5. Discussion

Our study investigated the reliability of different instruments for assessing individualized instruction in a classroom context. We used these instruments to capture the

teacher, student, and observer perspectives, which allowed us to determine the extent of their agreement with each other. We further explored possible factors that influence this agreement, in particular the type of self-report measure used and associated biases. To the best of our knowledge, this is the first study that assessed individualized task assignment from all three perspectives. Considering these perspectives simultaneously seems paramount for an accurate assessment of within-classroom individualization. In order to investigate the prevalence and associated outcomes of individualization in the classroom, it is important to ensure that individualization reflects the scientific operationalization of the construct as well as the actual classroom processes as perceived by students and teachers. We showed that individualized task selection in the classroom context can be reliably assessed from three different perspectives—teacher self-reports, student self-reports, and live observations (research question 1). The live observations additionally assessed individualized allocation of attention and time with a high interrater reliability. The three different perspectives do not strongly correlate with each other, however. For the facet of individualized task selection, student ratings were moderately correlated with both teacher self-reports and observer ratings, but the latter two did not correlate significantly (research question 2). A similar pattern of results has already been shown for other constructs, mainly classroom management (Den Brok et al., 2006; Fauth et al., 2014; Scherzinger & Wettstein, 2019). While this is the first study investigating agreement between all three perspectives for the construct of individualized task selection, the pattern of correlation is the same as for the basic dimensions of teaching quality. This indicates that the systematic differences between the perspectives might not be inherent to a specific construct, but rather reflect a general tendency of the specific perspective on classroom processes. Based on this observation we believe it to be likely that the other identified aspects of individualized instruction in a classroom context – individualized time and attention – also behave in a similar way, but future research will be needed to confirm this.

Besides the investigated construct, one of the main differences between our study and previous work concerning agreement between measures from students, teachers, and observers is that we did not use completely parallel scales for assessing the different perspectives but rather approached each perspective with an instrument appropriate for the target group. The fact that we found the same pattern presents strong evidence, in our view, that the shared variance is genuinely caused by the construct under investigation and not just by a common measurement method. These findings not only provide a foundation for future research questions concerning individualized instruction, but also shed some additional light on the different factors that influence teacher self-reports – namely the framing of the question and the timepoint of asking.

**5.1 Response Tendencies**

Another interesting finding is the discrepancy between the different types of self-report assessments. Teacher reports were more highly correlated with the other perspectives when assessed regularly and with questions targeted at specific past behavior rather than a summative questionnaire about their general behavior (research question 3). These findings are in line with the theory of different selves evoked by different kinds of self-reports (Conner & Barrett, 2012). According to this account, respondents that are asked to describe their behavior *in general* (as is the case in the DSAQ) tend to answer more in line with their *believing self*, i.e., based on how they see themselves as a person, rather than their concrete behavior. When people are asked about past behavior, they tend to answer more in line with their *remembering self*, i.e., based on their memory of said behavior—including typical memory biases such as peak-end-effect or self-presentation bias (Redelmeier & Kahneman, 1996; Zhang et al., 2018).

This self-presentation bias is also a plausible explanation for the positive correlations between the different items in the retrospective questionnaire. When asked about the presence

of behavior that is assumed to be positively connotated, teachers might tend to answer more in the affirmative than they would otherwise. The fact that the retrospective self-report correlated more highly with the other perspectives when corrected for this presumed bias further increases the plausibility of that hypothesis. Correcting the retrospective responses for the presumed self-presentation bias further increased the discrepancy between retrospective and trait self-reports—although it is reasonable to assume that the correlation of the trait measures to the other perspectives could also have been improved by some form of bias correction. The fact that only the uncorrected retrospective measures were significantly correlated with the trait measures provides some additional evidence for the hypothesis that those measures are biased by idealized self-perception.

## 5.2 Evaluation of Instruments

### *5.2.1 Classroom Observation*

By using low-inference ratings of directly observable behavior, we were able to attain very high reliability scores. These come at a price, though: by focusing on processes that are apparent on the surface of the lesson, we were only able to register the presence or absence of specific behavior without any information on the quality or adequacy. This focus on directly observable behavior offers a high amount of face validity and should increase the agreement with the student and teacher ratings (which also target directly observable behavior). It does, however, sacrifice a possible advantage provided by observers: the amount of professional knowledge they can use to make sense of observed situations. Future studies could use trained observers (probably through video and not live situations) to further differentiate between

effective and ineffective individualization attempts in a way that student and teacher ratings are likely unable to accomplish.

### *5.2.2 Teacher Questionnaires*

While the DSAQ trait questionnaire provides reliable measures of teachers' self-reported individualization, these estimates are most likely influenced by self-serving biases and thus tend to capture teachers' self-concept regarding individualized instruction. In order to better assess actual individualization, we recommend using items that target specific past behavior, presenting them as close as possible to the relevant point in time, and/or using some form of bias correction. Using an average of repeated assessments during the school year is recommended when a high level of stability of the behavior in question is to be expected, further increasing the reliability of the aggregate measure.

### *5.2.3 Student Questionnaire*

The brevity of the student questionnaire—9 items that were answered once at the end of the school year—and its ease of use make it a highly parsimonious instrument that can be integrated seamlessly in all kinds of study designs. The high ICC showed that student ratings reliably measure something unique to their teacher, while the fact that there are many independent observers of the same construct further increases the reliability of the aggregated measure. The simple statements in the questionnaire offer a high amount of face validity while the significant correlations with the observer and teacher perspective also indicate at least some convergent validity.

## 5.3 Limitations

The generalizability of these findings is limited by the characteristics of our specific sample. As in most classroom observation studies, the cooperating teachers are a subset of the teachers we asked to participate. It is reasonable to assume that this is not a random subset but

rather influenced by possible self-selection biases among the teachers. Another limitation is the fact that we only investigated third grade students from German schools, limiting the strength of generalizations for assessing individualized instruction in different age groups or cultures. Most importantly, we only investigated reading lessons; it is possible that the opportunities for individualized instruction, as well as the way this instruction is best implemented, vary between subjects and/or age-groups, with possible implications for the level of agreement between the perspectives as well. The fact that similar patterns have been found for different constructs, independent of the subject or the age-group, provides at least tentative evidence that our findings are generalizable. Future research is clearly needed, however, that replicates these findings in different subjects and age-groups.

Another limitation of the study lies in the brevity of the used instruments. Because it was part of a larger study and space in the assessment instruments was strongly limited, we could only probe one facet of individualization – individualized task selection – from all three perspectives, and even that facet was only assessed by a single item in the repeated teacher questionnaire. Having the same facets in all the instruments as well as several items per facet would have been advantageous for increasing reliability and could provide additional insights into the differences between the facets and how they are perceived by the different perspectives.

A final limitation of the current study is the fact that we only assessed indicators on the surface level of the lesson. While we were, for example, able to measure *if* a specific student was assigned a different task from the rest of the class, we were not able to assess whether the individualized task was actually a good fit for the student's current learning situation. The latter can also be defined as adaptive teaching, where the adaptations taken by the teacher are meant to enable and enhance self-regulated learning on the student side and thus engender a long-term process of academic aptitude development (Corno, 2008).

### 5.4 Future Directions

Going forward, student ratings seem to be a very efficient and valid way of assessing individualized instruction on a surface level. They showed the strongest correlation to both observer and teacher ratings, despite being assessed only once via a quick questionnaire. The fact that these ratings are composed of many individual student observers aggregated at class level also greatly improves their reliability. This does not mean that the other perspectives are without merit, however. Depending on the specific research questions, strong arguments can be made for using teacher reports or observational measures. When using teacher self-reports, we suggest that study designers should consider, in advance, possible biases and how to circumvent them. One recommendation would be to use items that are targeted at specific past behavior rather than general tendencies. These should be assessed as temporally close to the actual behavior as possible. When utilizing in situ observations, we recommend using low-inference indicators that are accessible on the surface of the lesson to guarantee satisfactory reliability and objectivity of the measures.

The ability to precisely operationalize and measure different types of individualized instruction is essential not only for assessing their merits, but also for identifying ways to foster individualized approaches in day-to-day classroom instruction. Only with a solid base of empirical research can the ubiquitous demand for more individualized education in regular classroom practice be met. By providing several instruments that capture individualized instruction from different perspectives, this article lays a foundation for future empirical research addressing this need.

An interesting example of such future research would be a more precise analysis of predictors on the teacher level–both of the amount of present individualization and of the alignment with other perspectives. Knowing the necessary prerequisites (both on a cognitive and an experiential level) for successful individualization would allow interventions to better

foster it. Investigating the determinants of a strong alignment between teachers and their respective students would shed further light on the specific factors influencing those perspectives.

## 5.5 Conclusion

This study shows that individualized task assignment can be reliably assessed from the teacher, student, and observer perspectives. The pattern of correspondence between these perspectives is similar to the one shown for other classroom processes. Students and observers show a moderate agreement in their ratings, students and teachers show a lesser – but still significant – agreement, and no agreement could be found between teachers and observers. We also show how some biases can be explained by the theory of multiple selves, which posits that such response tendencies can be partially alleviated by framing the questions in a way that targets the remembering rather than the believing self. Our results suggest that this can be achieved by asking about specific past behavior, as close as possible to the timepoint of interest. These findings pave the way for future research investigating the prevalence, prerequisites, and outcomes of individualized classroom instruction.

## References


Aleamoni, L. M. (1999). Student Rating Myths Versus Research Facts from 1924 to 1998. *Journal of Personnel Evaluation in Education 1999 13:2*, *13*(2), 153–166. https://doi.org/10.1023/A:1008168421283

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

Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). *A Systematic Review of Research on Personalized Learning : Personalized by Whom , to What , How , and for What Purpose ( s )?* Educational Psychology Review.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). https://psycnet.apa.org/record/2000-16936-008

Bondie, R. S., Dahnke, C., & Zusho, A. (2019). How Does Changing "One-Size-Fits-All" to Differentiated Instruction Affect Teaching?: *Https://Doi.Org/10.3102/0091732X18821130*, *43*(1), 336–362. https://doi.org/10.3102/0091732X18821130

Brekelmans, M., Mainhard, T., den Brok, P. D., & Wubbels, T. (2011). Teacher control and affiliation: Do students and teachers agree? *Journal of Classroom Interaction*, *46*(1), 18–27.

Clunies-Ross, P., Little, E., & Kienhuis, M. (2008). Self-reported and actual use of proactive and reactive classroom management strategies and their relationship with teacher stress and student behaviour. *Educational Psychology*, *28*(6), 693–710. https://doi.org/10.1080/01443410802206700

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement ☆. *Journal of Business Research*, *61*, 1250–1262. https://doi.org/10.1016/j.jbusres.2008.01.013

Conner, T. S., & Barrett, L. F. (2012). Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine. In *Psychosomatic Medicine* (Vol. 74, Issue 4, pp. 327–337). NIH Public Access. https://doi.org/10.1097/PSY.0b013e3182546f18

Connor, C. M., Mazzocco, M. M. M., Kurz, T., Crowe, E. C., Tighe, E. L., Wood, T. S., & Morrison, F. J. (2018). Using assessment to individualize early mathematics instruction. *Journal of School Psychology*, *66*, 97–113. https://doi.org/10.1016/j.jsp.2017.04.005

Connor, C. M., Piasta, S. B., Glasney, S., Schatschneider, C., Fishman, B. J., Underwood, P. S., & Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child-by-instruction interactions on students' literacy. *Child Development*, *80*(1), 77–100.

Corno, L. (2008). On teaching adaptively. *Educational Psychologist*, *43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Daruwala, I., Bretas, S., & Ready, D. D. (2021). When Logics Collide: Implementing Technology-Enabled Personalization in the Age of Accountability. *Educational Researcher*, *50*(3), 157–164. https://doi.org/10.3102/0013189X20960674

Den Brok, P., Bergen, T., & Brekelmans, M. (2006). Convergence and divergence between students' and teachers' perceptions of instructional behaviour in Dutch secondary education. *Contemporary Approaches to Research on Learning Environments: Worldviews*, 125–160. https://doi.org/10.1142/9789812774651_0006

Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation Practices:A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review*, *24*, 31–54. https://doi.org/10.1016/j.edurev.2018.02.002

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*(12), 1203–1218. https://doi.org/10.1016/j.jbusres.2008.01.009

Dockterman, D. (2018). Insights from 200+ years of personalized learning. *Npj Science of Learning*, *3*(1), 1–6. https://doi.org/10.1038/s41539-018-0033-x

Donker, M. H., van Vemde, L., Hessen, D. J., van Gog, T., & Mainhard, T. (2021). Observational, student, and teacher perspectives on interpersonal teacher behavior: Shared and unique associations with teacher and student emotions. *Learning and Instruction*, *73*, 101414. https://doi.org/10.1016/j.learninstruc.2020.101414

Dumont, H. (2016). Die empirische Untersuchung von individueller Förderung als Perspektive für die Unterrichtsqualitätsforschung. In *Bedingungen und Effekte guten Unterrichts* (pp. 107–116).

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2018). Exploring teacher popularity: associations with teacher characteristics and student outcomes in primary school. *Social Psychology of Education*, *21*(5), 1225–1249. https://doi.org/10.1007/s11218-018-9462-x

Fleiss, J. L. (1986). Reliability of measurement. In: The design and analysis of clinical experiments. New York, Wiley, pp 1–32

Förster, N., Kawohl, E., & Souvignier, E. (2018). Short- and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension. *Learning and Instruction*, *56*(March), 98–109. https://doi.org/10.1016/j.learninstruc.2018.04.009

Förster, N., & Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learning and Instruction*, *32*, 91–100. https://doi.org/10.1016/j.learninstruc.2014.02.002

Fraser, B. J. (1981). *Validity and use of individualized classroom environment questionnaire*.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *The American Psychologist*, *52*(11), 1182–1186. https://doi.org/10.1037//0003-066X.52.11.1182

Hachfeld, A., & Lazarides, R. (2020). The relation between teacher self-reported individualization and student-perceived teaching quality in linguistically heterogeneous classes: an exploratory study. *European Journal of Psychology of Education*, 1–21. https://doi.org/10.1007/s10212-020-00501-5

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online 11 (2019) 2, S. 169-191*, *11*(2), 169–191. https://doi.org/10.25656/01:18004

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Houtveen, J. H., & Oei, N. Y. L. (2007). Recall bias in reporting medically unexplained

symptoms comes from semantic memory. *Journal of Psychosomatic Research*, *62*(3), 277–282. https://doi.org/10.1016/j.jpsychores.2006.11.006

Jung, P.-G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of Data-Based Individualization for Students with Intensive Learning Needs: A Meta-Analysis. *Learning Disabilities Research & Practice*, *33*(3), 144–155. https://doi.org/10.1111/ldrp.12172

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When More Pain Is Preferred to Less: Adding a Better End. *Psychological Science*, *4*(6), 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Kopcha, T. J., & Sullivan, H. (2007). Self-presentation bias in surveys of teachers' educational technology practices. *Educational Technology Research and Development*, *55*(6), 627–646. https://doi.org/10.1007/s11423-006-9011-8

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

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Major, L., Francis, G. A., & Tsapali, M. (2021). The effectiveness of technology-supported personalised learning in low- and middle-income countries: A meta-analysis. *British Journal of Educational Technology*, *52*(5), 1935–1964. https://doi.org/10.1111/bjet.13116

Markic, S., & Abels, S. (2014). Heterogeneity and diversity: A growing challenge or enrichment for science education in German schools? *Eurasia Journal of Mathematics, Science and Technology Education*, *10*(4), 271–283. https://doi.org/10.12973/eurasia.2014.1082a

McConnell, J. W., & Bowers, N. D. (1979, April). *A Comparison of High-Inference and Low-Inference Measures of Teacher Behaviors as Predictors of Pupil Attitudes and Achievements.*

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3),

276–282. https://doi.org/10.11613/bm.2012.031

Miller, R. L. (1976). Individualized Instruction in Mathematics: A Review of Research. *Mathematics Teacher*.

Nava, I., Park, J., Dockterman, D., Kawasaki, J., Schweig, J., Quartz, K. H., & Martinez, J. F. (2019). Measuring Teaching Quality of Secondary Mathematics and Science Residents: A Classroom Observation Framework. *Journal of Teacher Education*, *70*(2), 139–154. https://doi.org/10.1177/0022487118755699

Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., Pierczynski, M., & Allen, M. (2018). Teachers' Instructional Adaptations: A Research Synthesis. *Review of Educational Research*, *88*(2), 205–242. https://doi.org/10.3102/0034654317743198

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12. https://doi.org/10.1016/J.LEARNINSTRUC.2013.12.002

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM - Mathematics Education*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Prast, E. J., van de Weijer, E., Kroesbergen, E. H., & van Luit, J. E. H. (2015). Readiness-based differentiation in primary school mathematics : Expert recommendations and teacher self-assessment. *Frontline Learning Research*, *3*(2), 90–116. https://doi.org/https://doi.org/10.14786/flr.v3i2.163

Prast, E. J., Van de Weijer-Bergsma, E., Kroesbergen, E. H., & Van Luit, J. E. H. (2018). Differentiated instruction in primary mathematics: Effects of teacher professional development on student achievement. *Learning and Instruction*, *54*, 22–34. https://doi.org/10.1016/J.LEARNINSTRUC.2018.01.009

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *66*(1), 3–8. https://doi.org/10.1016/0304-3959(96)02994-6

Roberts-Mahoney, H., Means, A. J., & Garrison, M. J. (2016). Netflixing human capital development: personalized learning technology and the corporatization of K-12 education. *Journal of Education Policy*, *31*(4), 405–420. https://doi.org/10.1080/02680939.2015.1132774

Robinson, M. D., & Clore, G. (2002). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes Feelings as Information View project Dynamic Negativity Effects in Emotional Responding: Onsets, Peaks, and Influences from Repetition View project. *Article in Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.83.1.198

Scherzinger, M., & Wettstein, A. (2019). Classroom disruptions, the teacher–student relationship and classroom management from the perspective of teachers, students and external observers: a multimethod approach. *Learning Environments Research*, *22*(1), 101–116. https://doi.org/10.1007/s10984-018-9269-x

Slavin, R. E., & Karweit, N. L. (1985). Effects of Whole Class, Ability Grouped, and Individualized Instruction on Mathematics Achievement. In *American Educational Research Journal Fall* (Vol. 22, Issue 3). http://journals.sagepub.com/doi/pdf/10.3102/00028312022003351

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

Stradling, B., & Saunders, L. (2006). Differentiation in practice: responding to the needs of all pupils. *Http://Dx.Doi.Org/10.1080/0013188930350202*, *35*(2), 127–137. https://doi.org/10.1080/0013188930350202

Suprayogi, M. N., Valcke, M., & Godwin, R. (2017). Teachers and their implementation of differentiated instruction in the classroom. *Teaching and Teacher Education*, *67*, 291–301. https://doi.org/10.1016/j.tate.2017.06.020

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing Personalized Education: A Dynamic Framework. *Educational Psychology Review*, *33*(3), 863–882. https://doi.org/10.1007/S10648-020-09570-W/FIGURES/3

van Geel, M., Keuning, T., Frèrejean, J., Dolmans, D., van Merriënboer, J., & Visscher, A. J. (2019). Capturing the complexity of differentiated instruction. *School Effectiveness and School Improvement*, *30*(1), 51–67. https://doi.org/10.1080/09243453.2018.1539013

Vaughn, M., Parsons, S. A., & Gallagher, M. A. (2021). Challenging Scripted Curricula With Adaptive Teaching. *Educational Researcher*, 1–11. https://doi.org/10.3102/0013189X211065752

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner & E. Souberman., Eds.) (A. R. Luria, M. Lopez-Morillas & M. Cole [with J. V. Wertsch], Trans.) Cambridge, Mass.: Harvard

University Press. (Original manuscripts [ca. 1930-1934])

Waxman, H. C., Wang, M. C., Anderson, K. A., Herbert, J., Waxman, C., Wang, M. C., & Anderson, K. A. (1985). *Adaptive Education and Student Outcomes : A Quantitative Synthesis*. *78*(4), 228–236.

Wubbels, T., Brekelmans, M., & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education*, *8*(1), 47–58. https://doi.org/10.1016/0742-051X(92)90039-6

Zhang, Y., Pan, Z., Li, K., & Guo, Y. (2018). Self-Serving Bias in Memories: Selectively Forgetting the Connection between Negative Information and the Self. *Experimental Psychology*, *65*(4), 236–244. https://doi.org/10.1027/1618-3169/a000409

***Paper 4***

Schmitterer A., **Tetzlaff, L.,** Hasselhorn, M., & Brod, G: *Who benefits from Computerized Learning Progress Assessment in Reading Education? Evidence from a Two-Cohort Longitudinal Study.* Manuscript submitted for Publication in Journal of Computer Assisted Learning.

# Who benefits from Computerized Learning Progress Assessment in Reading Education? Evidence from a Two-Cohort Longitudinal Study

Schmitterer, A.M.A.[1,2], Tetzlaff, L. D.[1], Hasselhorn, M.[1,3], & Brod, G.[1,3]

[1] DIPF | Leibniz Institute for Research and Information in Education

[2]University of Paderborn

[3] Goethe-University Frankfurt am Main

## Statements and Declarations

Funding: This research was funded by the German Federal Ministry of Education and Research, as part of the LONDI research program. GB was supported by a Jacobs Foundation Research Fellowship.

This research was approved by the Ethics committees of our research institution (DIPF) and the Hessian Ministry of Education prior to data collection. Participants or participants' care-givers consented to participation after being informed of their right to drop out of the study at any time without negative consequences. The authors have no conflict of interest.

## Open Science

Data availability: https://osf.io/62fdy/?view_only=d7f90e7b6ab040e79748900b8963c12e

Code availability: https://osf.io/62fdy/?view_only=d7f90e7b6ab040e79748900b8963c12e

## Abstract

**Background:** Learning Progress Assessments (LPA) have been developed to help teachers individualize their curriculum. The use of LPA is facilitated by an increasing number of computerized LPA tools. However, little is known about student factors that influence the effectiveness of computerized LPA.

**Objectives:** In this study, we explored whether a computerized LPA that focused on reading comprehension was differentially effective depending on students' initial reading comprehension abilities. Moreover, effects of the LPA implementation on other, related literacy skills (i.e., decoding, spelling) were explored.

**Methods:** The development of reading and spelling skills of 668 third graders was assessed in 41 LPA and 36 control classes in a pre- and posttest design. We analyzed effects of the LPA on reading comprehension, decoding, and spelling skills, and whether this effect was qualified by students' initial achievement level.

**Results:** The LPA treatment proved beneficial for improving reading comprehension but not for improving decoding or spelling. Children with low levels of reading comprehension at the beginning of the school year benefitted particularly from LPA.

**Takeaways:** Our results suggest that teachers used the data offered by the Computerized LPA to identify children with reading comprehension difficulties and to adapt their teaching to their specific deficits. This leads to an accelerated increase in reading comprehension ability that is specific to this group.


Keywords: learning progress assessment, formative assessment, reading education, effectiveness

## Open Practices Statement

The materials, the data, and the script that was used to analyze the data are available via the Open Science Framework and can be accessed at

https://osf.io/62fdy/?view_only=d7f90e7b6ab040e79748900b8963c12e

# Who benefits from Computerized Learning Progress Assessment in Reading Education?

## Evidence from a Two-Cohort Longitudinal Study

Reading acquisition is an individual journey. Children progress at different rates and from the very beginning of reading education, studies find a lot of variability in students' ability to read (e.g., Mullis et al., 2017). Large-scale studies in Germany suggest that even with formal education, about 20% of students leave primary school without advanced reading comprehension skills (Hußmann et al., 2017 [IGLU-Studie]; Stanat et al., 2017 [IQB Bildungstrend]). One approach to dampen this trend has been to individualize reading instruction. Individualized instruction has become a dominant educational approach with respected philosophies increasing all over Europe during the last decade (Lai & Schildkamp, 2013; Peters et al., 2021). One tool that has been developed to support teachers in individualized instruction is Learning Progress Assessments (LPA; Förster & Souvignier, 2011; Fuchs & Fuchs, 1986; Walter, 2014).

Originally developed for special education classrooms in the U.S. (Deno, 1985; Fuchs & Fuchs, 1986), LPAs have since been developed for several areas of academic achievement (e.g., decoding, reading speed, reading comprehension, spelling, mathematics; Förster & Souvignier, 2011; Jung et al., 2018; Lembke et al., 2012; Schütze, Souvignier & Hasselhorn, 2018). The use of LPAs has expanded to general education thanks to the availability of computerized LPA tools that facilitate their use with larger groups of students (e.g., Espin et al., 2012; Förster et al., 2018). The key difference to standardized summative tests is that LPAs consist of repeated parallel tests that closely match the curriculum and thereby allow for a more fine-grained monitoring of students' progress (e.g., weekly or every couple of weeks). LPAs are, therefore, suited to detect short-term changes (Deno, 1985; Fuchs, Deno, & Mirkin, 1984;

Lembke et al., 2012; Walter, 2014) and can help teachers to react swiftly to negative changes in students' learning trajectory.

Regarding efficacy, a recent meta-analysis reported low to medium sized beneficial effects of an implementation of LPAs (e.g., Jung et al., 2018) in comparison to control classes. However, not all meta-analyses revealed the same conclusion. Kingston and Nash (2011) found overall small effect sizes of LPA implementations but identified factors that influenced the LPA's efficacy. For example, they found that computer-based implementations or implementations of LPA that include courses for professional development had the largest effects in comparison to other types of implementations. Thus, they noticed that the type of LPA implementation and how teachers use the data provided by an LPA might affect the efficacy of the tool. Therefore, they pointed out that instead of focusing broadly only on the efficacy of the usage of LPA, more studies should focus on *factors that influence* the efficacy of LPA.

In line with that suggestion, some recent studies focused on the teacher level to identify factors that can influence the efficacy of an LPA implementation. For example, Zeuch and colleagues (2017) showed that teachers' ability to read and interpret learning progress graphs provided by LPA tools differed as a function of their subject knowledge. Schildkamp and colleagues (2020) went on to identify several teacher characteristics (e. g., prior knowledge) that influence the efficacy of LPA implementations. By comparison, factors on the level of students have been studied fewer but might also contribute to explaining the efficacy of LPA implementations.

For example, Peters and colleagues (2021) studied whether data from several LPA efficacy studies show benefits for low-achieving students in reading fluency or reading comprehension (up to the 25$^{th}$ percentile). Overall, results indicated that LPA implementations

were not effective for low achieving students. The result is interesting and might indicate that teachers might not use the data provided by LPA to support low achievers – but might also have been obscured by the absolute cut-off criterion for low achievement. Some studies (e.g., Schmitterer & Brod, 2021) indicate that teachers orientate on their class mean rather than on an absolute or standardized criterion of students' reading competencies when estimating who needs reading support. Focusing on children within the lowest 25% in a standardized test might cut some variability of children that teachers perceive as low achieving in comparison to their class mean but might not be low achievers in a standardized test. Therefore, in this study we aimed to check whether the general initial reading level of students had an impact on the efficacy of the LPA implementation.

Results by Peters and colleagues (2021) further revealed class level effects. For example, data from second and fourth grade indicated no beneficial effects of LPA treatment but data from third grade indicated low to medium-sized benefits of an LPA implementation for low-achieving readers. As mentioned above, LPAs are developed to match the school curriculum. The overlap of the LPA's task construct with the curriculum and the skills assessed might be another factor contributing to LPA implementation efficacy. For example, if the focus of the curriculum is reading comprehension and the LPA task focuses on reading comprehension but we measure decoding, one might hardly find beneficial effects of the LPA implementation. If teachers did not receive information about their student's learning progress in decoding abilities, they might not focus on students' respected reading difficulties in their adaptive teaching approaches. In the case of the third grade LPA implementation reported by Peters and colleagues (2021), the construct of the LPA task matched the curriculum and the skills measured at the beginning and at the end of the school year. This matching, to our knowledge, has not been focused on in previous LPA efficacy studies.

Studying translating effects of LPA efficacy to not directly assessed reading skills might also enhance our understanding of the effects of LPA. There are several things teachers could do if the data provided by the LPA indicates, say, slow reading comprehension progress. They could focus on training the specific skills assessed in the LPA (i.e., reading comprehension) or they could focus on underlying or related skills (e.g., decoding, spelling) or on other factors altogether (e.g., reading motivation). Little is known so far about what teachers are doing with the data provided by the LPA, which is likely the case because it would require in-depth interviews and classroom observations. Indirect evidence can be drawn from comparing the effects of the LPA on the assessed reading skill versus related skills that were not directly assessed. If teachers used the data offered by the LPA to provide individualized training on a specific skill, students should improve more on the specific skill rather than on related skills (e. g., Ise et al, 2012; Sala et al., 2019). If this was not the case, effects on related skills should be similarly strong.

In this study, we present data from a two-cohort longitudinal study with third graders. In our analyses, we first aimed to replicate the finding that children's reading comprehension abilities improve more in classes using a computerized LPA of reading comprehension in comparison to control classes without LPA (e.g., Jung et al., 2018; Kingston & Nash, 2011; Peters et al., 2021). Second, we strived to test whether children's initial achievement level is related to the efficacy of applying LPA. We expected that children with lower initial achievement levels would be more likely to benefit than children with high initial achievement. Finally, we aimed to explore whether the usage of LPA specifically improves students' reading comprehension as assessed in the LPA, or also transfer to other reading or reading-related skills (i.e., decoding and spelling). In light of previous research, we expected treatment effects

specific to reading comprehension - the skill measured with the LPA that matched the curriculum.

## Methods

The current study was part of a larger research project running from 2018 to 2020. The study followed a quasi-experimental design with a treatment group having access to a computerized version of an LPA and a control group without access to the LPA. Recruitment was conducted at the school level. Teachers could choose freely whether they wanted to participate in the study. Data were collected in two cohorts, one in the school year 2018/2019 in the state of Hesse in Germany, the other in the school year 2019/2020 in the states of Hesse and Lower Saxony. Each cohort completed a pretest at the beginning of the school year and a posttest before the summer break. Due to the school lockdowns and distance measures in Germany during the Covid-19 pandemic, the posttests in the school year 2019/2020 were administered not by trained research assistants, but by the respective teachers. Teachers received an elaborate handbook beforehand and were instructed to keep as closely to the handbook as possible. Teachers and children's caregivers gave their consent to the participation before data collection. Children did not receive any incentives for their participation. Teachers and parents received summary statistics of their own classes and details regarding the changes of their children's reading and spelling skills at the end of the school year.

### Participants

#### *Teachers*

We recruited 61 teachers from primary schools in Hesse, and 16 teachers from schools in Lower Saxony for participation in our study. Of the 61 Hessian teachers, 46 used the LPA program "*quop*" in their reading lessons (treatment group). When carrying out the study,

"*quop*" was available free of charge for schools in Hesse as part of a state-wide program and schools were encouraged to use it. If a school decided to take part in the program, most teachers in a school used it. Control group classes were sampled from schools who did not take part in the program. Because both participation in "*quop*" and in our study was voluntary, assignment of classes to conditions was not random. The teachers in the treatment group were on average 40.34 (*SD* = 9.95) years old. The teachers in the control group were on average 42.0 (*SD* = 8.55) years old. Three of the teachers were male (two of them in the treatment group), the remaining 74 were female. Teachers in the treatment group reported an average of 13.11 (*SD* = 7.95, range = 4–30) years of general teaching experience and 5.65 (*SD* = 5.82, range = 0 – 27) years of experience teaching third grade classes. Teachers in the control group reported an average of 13.56 (*SD* = 6.99, range = 4 – 30) years of teaching experience and 5.73 (*SD* = 5.18, range = 1 – 19) years of experience teaching third grade classes.

***Students***

From the more than 700 third graders in the participating classes, we excluded 52 (7 %) children with a nonverbal IQ below 70 from analysis, to make sure that all children were able to comprehend all instructions and because we assumed that most of these children participated in special education programs that would affect the interpretation of our data. The remaining 668 children had full datasets of measures administered during group sessions at both pre- and posttest. These 668 children were nested in 77 (31 control) classes. They were on average 8;8 years; months old (*SD* = 5.59 months) and 57 % of them were female. Information about socio-economic backgrounds (SES) were retrieved from 549 of 668 parents and operationalized with the HISEI (Highest International Socio-Economic Index of Occupational Status; Ganzeboom et al.,1992 and Ganzeboom, 2010). The SES was highly variable in both groups (treatment: mean = 54.27, *SD* = 17.07, range = 15 – 89; control: mean = 51.15, SD = 17.45, range = 16 –

89). Children in the treatment group scored significantly higher than those in the control group on measures of nonverbal intelligence ($p < .001$) and reading comprehension ($p = .03$) at pretest (see Table 1). Means and standard deviations of group differences are reported in Table 1. The children in the treatment group also showed a higher proportion of native German speakers (74 %) and a lower amount of bilingual children (16 %) than the control group (46 % native 27 % bilingual; $p < .001$). To control for these pre-existing group differences, we added them (separately) as control variables to our final model. Those additional models can be found in the Appendix (Tables A2 and A3).

**Table 1**

*Mean and Standard Deviations of Background Variables for the Student Sample*

| | Mean (*SD*) Control | Mean (*SD*) Treatment | Range Control | Range Treatment |
|---|---|---|---|---|
| Reading Comprehension[a] | 60.16 (*20.34*) | 63.46 (*20.42*) | 22 – 118 | 15 – 113 |
| Nonverbal Intelligence[b] | 32.32 (*6.17*) | 33.99 (*5.55*) | 20 – 44 | 20 – 43 |

*Note.* [a]raw sum scores from ELFE-II; [b]raw sum scores from the CFT; see materials section for descriptions of the respective tests; $N = 668$.

In addition to group sessions, we also administered some reading and language tasks (i.e., productive tasks involving speaking) in individual sessions. For these sessions teachers nominated up to seven students from their classroom that either had issues with literacy or were representative to the classes reading level and diversity. This was the case, because the greater framework of the research project had one focus on underlying abilities of reading in children with reading difficulties. Regarding age, gender and socio-economic background there were no relevant differences between the samples from group and individual sessions. However, by

comparison the group session, children in individual sessions, expectedly, showed lower achievement levels in reading and more children spoke German as a second language. The sample differences are reported in detail in this study (Schmitterer & Brod, 2021). Regarding the present study, data from individual sessions were only used for the analyses of decoding abilities in the small sample to analyze whether beneficial effects of the LPA implementation for reading comprehension also translated to underlying reading abilities. To this end, we report relevant variables and ran additional analyses to ensure comparability that are referred to in the results section and reported in the Appendix.

**LPA**

The computer-based LPA "*quop*" (Souvignier & Förster, 2011; Souvignier et al., 2021) used in the treatment classes is a web-based LPA program developed at the Universität Münster. For third grade students, *quop* measures reading comprehension as well as reading fluency with a maze task every three weeks and is providing teachers with graphs of the development of their students in these domains (for a detailed description of the tasks used in *quop,* see Souvignier et al., 2021). Teachers receive graphs displaying the progress of individual students' average test scores (accuracy and speed) compared to the class mean. Thus, teachers are provided with information about which children stay below or above the class mean across the school year. Furthermore, they have some information that is connected to the specific task selection of the LPA.

To monitor teachers' use of the LPA, we asked them to complete an online questionnaire every three weeks in which they reported the number of times they looked at the LPA data for their students. In a three-week timeframe, teachers reported to look at the data of 6.17 children on average and to look at each child's data 1.69 times. Teachers also reported to find the data "somewhat useful" (3.01 on a 1-4 Likert scale). This indicates that teachers were

using the information provided by *quop*. We also asked them which information they found especially useful, with 88% selecting the individual progress graphs, 52% selecting the graph of the class mean and 50% the comparison of the mean and the individual. Multiple selections were possible.

## Materials

### *Reading Comprehension*

For assessing the level of reading comprehension, we used the ELFE II (Lenhard et al., 2017). ELFE II is a speeded test, measuring reading comprehension on the word-, sentence- and text-level. Since we were interested in general reading comprehension proficiency, we used sum scores over all subtest as a general reading comprehension score. The internal consistency (alpha = 0.97) indicates a high reliability of the sum scores. For a detailed description of the ELFE and its specific items, see Lenhard et al., (2017).

### *Decoding*

Children's decoding ability was assessed by the pseudoword reading task of the SLRT II (Moll & Landerl, 2010). In this task, children were presented a list of pseudowords and asked to read as many of the pseudowords out loud as possible in one minute. The variable of interest was the number of pseudowords that were correctly read aloud in one minute.

### *Spelling*

Spelling ability was measured by the spelling test of the SLRT II (Moll & Landerl, 2010). In this test, children were presented 48 written sentences that each missed one word. Each sentence was read aloud, including the missing word. Then the missing word was repeated, and children had to write down the missing word into the blank spaces. The variable of interest in this case was the number of correctly spelled words.

### *Nonverbal Intelligence*

As our measure of nonverbal intelligence, we used the Culture Fair Intelligence Test (Weiß & Osterland, 2013). The CFT is a nonverbal intelligence test and was administered by trained test supervisors in a pen and paper format to the whole class at the same time. For a detailed description of the CFT and the specific tasks used in it see Weiß & Osterland (2013).

**Analyses**

We ran multilevel models to account for the nested structure of students attending classes of teachers who did or did not use LPA. Thus, we accounted for the variability in reading proficiency between classes (Level 2) and between students (Level 1). We fitted the models using the lme4 package (Bates et al., 2015) in R. The dependent variable was children's reading comprehension score at the end of third grade. For the first model, we added reading comprehension at the beginning of the school year and group (LPA vs. control) and their interaction as fixed effects to the model described above. Moreover, classes were added as a random effect factor to account for variability between classes. For all analyses, we quantified the severity of multicollinearity using variance inflation factors. Variance inflation factors were all below 8, which has been estimated to be within an acceptable range (O'Brian, 2007; Marcoulides & Raykov, 2019).

Since our participants attended school in different regions of Germany and the second wave of our sample was affected by Covid-19 lockdowns, we controlled for region and wave with additional fixed effects. Finally, in all models, we checked for variance inflation ({vif} function in *car* package; Fox & Weisberg, 2019). Furthermore, we calculated Pseudo $R^2$ ({r.squaredLR} function in the *MuMin* package; Bartón, 2019) to show the explained variability of the testing models based on Level 1 effects in comparison to null models (only

random effects). Explained variances were high and Pseudo-$R^2$s are reported in the notes to the tables of the respective model.

**Table 2**

*Results of the Multilevel Model focusing on the Effect of initial Achievement on Reading Comprehension Abilities at the End of Third Grade as a Function of Attendance of an LPA Class*

| **Fixed Effects** | **Estimate** | ***SE*** | ***t* value** |
|---|---|---|---|
| Intercept | 24.06 | 2.53 | 9.49*** |
| Reading Comprehension at T1 | 0.93 | 0.03 | 28.16*** |
| Treatment Group (LPA/ No LPA) | 7.36 | 3.23 | 2.27 * |
| Year of Assessment (Corona/ No Corona) | -0.29 | 1.58 | -0.18 |
| Region (Hesse/ Lower Saxony) | 2.41 | 2.48 | 0.97 |
| Reading Comprehension at T1 * TGroup Interaction | -0.10 | 0.04 | -2.53* |

Note. *** $p < .001$; ** $p < .01$; * $p < .05$; For this analysis data from 668 students with full data sets were used; Pseudo-$R^2 = 0.74$; $N = 668$.

## Results

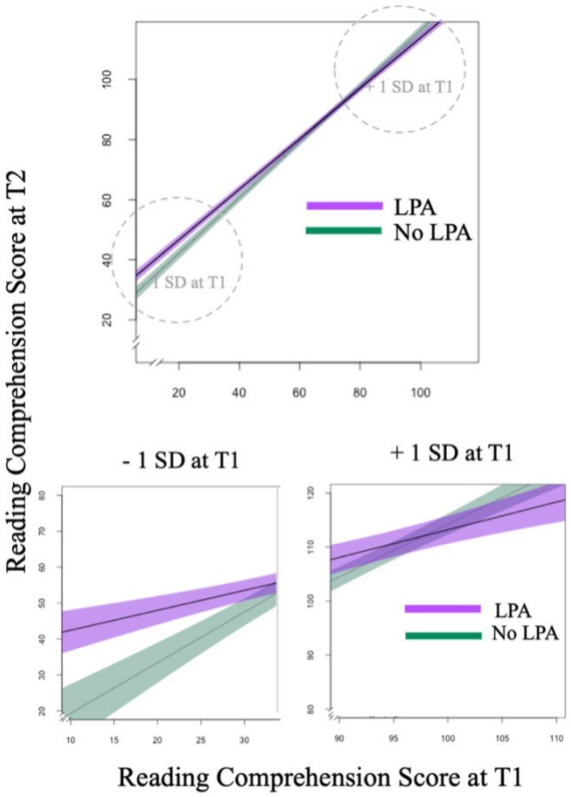
### Computerized LPA Particularly Supports Poor Readers

Our results (see Table 2) show that reading comprehension at the beginning of the school year was a strong predictor for reading comprehension at the end of the school year. This finding indicates a high stability of individual differences in reading comprehension

proficiency, already during grade 3. Nevertheless, children in classes that used LPA had higher gains in reading comprehension at the end of third grade than their peers in classes without usage of LPA. This finding held when year of assessment or region were controlled for. These effects were qualified by a significant interaction effect between group and children's reading comprehension at pretest. This interaction suggests that the differential stability of reading comprehension differed between classes with and without LPA. Figure 1 indicates that this interaction effect is mainly driven by students with low reading comprehension scores at the beginning of the school year. When these children were in the LPA group, they achieved higher levels of reading comprehension at the end of the school year than their peers in classes without LPA. In contrast, children with high initial scores improved in a similar way regardless of their classes' use of LPA. Thus, results suggest that children with lower achievement levels were more likely to profit from the application of LPA in their classroom.

**Figure 1**

*Interaction effect of the between-group Differences of the Effect of Achievement level at T1 on the Reading Comprehension Score at T2*




*Note*. Polygons around the lines represent 1 *SE* deviation from the mean (i.e., lines); Lines represent the effect of achievement level at T1 on achievement at T2 in a reading comprehension test (i.e., ELFE); Graphs at the bottom zoom into effects for children with low or high achievement at T1.

**Do the Effects of LPA Translate to other Literacy Skills?**

Moreover, we also ran two modified versions of the first model to check for effects of the LPA treatment on further reading and reading-related skills (see Table 3). In one model, we focused on decoding and, thus, used decoding at posttest as the dependent variable. Decoding at pretest and its interaction with the group variable (treatment vs. no treatment) were entered as fixed effect in the model. In the second model, we did the same for spelling. Since decoding proficiency was assessed in individual sessions, the sample size in this model was smaller. We, therefore, ran an additional model to replicate the first analysis to ensure that results regarding reading comprehension would be comparable in the large and small sample.

Results indicated that, indeed, effects in both sample sizes were comparable. An overview of differences in initial reading ability of the large and small samples are provided in the Appendix (Tables A1, A2). In both cases, no transfer effects of the LPA treatment were found (i.e., no main or interaction effect involving the group variable).

**Table 3**

*Effects of Initial Achievement in Decoding and Spelling Development at the End of Third Grade as a Function of Attendance of an LPA Class*

| **Decoding (*N* = 230)** | | | |
|---|---|---|---|
| **Fixed Effects** | **Estimate** | ***SE*** | ***t* value** |
| Intercept | 16.21 | 1.74 | 9.34*** |
| Decoding at T1 | 0.95 | 0.03 | 28.6*** |
| Treatment Group (LPA/ No LPA) | -3.27 | 3.18 | -1.03 |
| Region (Hesse/ Lower Saxony) | -1.23 | 1.33 | -0.93 |
| Decoding at T1 * TGroup Interaction | -0.01 | 0.08 | 0.16 |
| **Spelling (*N* = 660)** | | | |
| Intercept | 14.73 | 0.98 | 14.93*** |
| Spelling at T1 | 0.72 | 0.04 | 24.51*** |
| Treatment Group (LPA/ No LPA) | -0.38 | 1.27 | -0.30 |
| Year of Assessment (Corona/ No Corona) | 0.40 | 0.58 | 0.67 |
| Region (Hesse/ Lower Saxony) | 0.39 | 0.94 | 0.42 |
| Spelling at T1 * TGroup Interaction | 0.05 | 0.04 | 1.24 |

Note. *** $p < .001$; ** $p < .01$; * $p < .05$; The number of data points per analysis represent all children for whom Decoding or Spelling data at T1 and T2 were complete (see Table A2 and Method section for details); Pseudo-$R^2_{Decoding} = 0.99$; Pseudo-$R^2_{Spelling} = 0.71$.

### Is the LPA Conditions' Efficacy explained by Other Covariates?

We discovered several differences in background variables between treatment groups. Therefore, we ran a number of analyses with additional covariates that controlled for social or intellectual differences between groups. To achieve this, we added measures of language background and nonverbal intelligence of the children as additional covariates, separately. Controlling for language background attenuated the strength of the main effect but did not explain additional variance in reading comprehension gains. The interaction effect (RC X Treatment group) also remained untouched. Controlling for nonverbal intelligence did not affect the results. These models are reported in the Appendix (Tables A3 and A4).

## Discussion

Computerized LPAs carry the promise of supporting teachers in individualizing their teaching (e.g., Jung et al., 2018). In this study, we aimed to replicate the efficacy of computerized LPA for promoting reading comprehension in elementary reading education (Förster et al., 2018; Förster et al., 2015; Hebbecker & Souvignier, 2018; Jung et al., 2018), and we aimed to expand this research in two ways. First, we asked if the response to the LPA depends on children's initial achievement level. Second, we asked if the effects of LPA usage were specific to reading comprehension or also transferred to further reading or reading-related skills. With our study, we respond to previous research that has recommended to not focus solely on the efficacy of the implementation of LPA but also on factors that influence the efficacy of the use of LPAs (e.g., Kingston & Nash, 2011).

### General Efficacy of Computerized LPA

Of the 77 classes that were used in the analysis, 31 were control classes and 46 were classes in which teachers made use of a computerized LPA to follow their students' reading comprehension development throughout the school year. Our results show a positive main

effect of treatment, indicating that children in classes where the LPA was implemented showed overall more gains in their reading comprehension than their peers in the control group. This effect was also found when differences between the treatment groups were controlled (i.e., nonverbal intelligence, language background; see Tables A3 and A4 in the Appendix). These results are in line with meta-analyses that show beneficial effects for the use of LPA in teaching reading (Jung et al., 2018).

**Differential Effectiveness Based On Students' Initial Reading Comprehension Abilities**

In our models, an interaction effect of the treatment group with the initial reading level indicated that the treatment effect was moderated by the level of reading comprehension at the beginning of the school year. Considering the positive main effect for treatment group, the negative interaction effect indicated that specifically children with low initial levels in reading comprehension benefited from the implementation of the LPA in their class (see Figure 1). These findings suggest that either teachers used the information provided by the LPA particularly for identifying and supporting children with reading difficulties in an individualized manner, or that they generally adapted their teaching towards more individualized instruction and low achievers particularly benefited from this adaptation. Either way, this finding suggests that the initial achievement level of children is one factor that is of relevance for the efficacy of the implementation of LPA. However, we would need more precise data on how teachers used data to support their students on an individual level to draw more specific conclusions.

Our results regarding achievement level add to recent findings of Peters and colleagues (2021), who found some beneficial effects for the LPA implementation in low-achieving third grader. These findings were connected to special teacher training programs, which was not the case in our sample. Furthermore, our data was collected in a different region. We also used a

continuous scale of achievement rather than a standardized cut-off criterion, because previous studies suggest that teachers orientate on the mean skill level in their class rather than on an absolute criterion (Schmitterer & Brod, 2021). Nevertheless, ours and their studies both found beneficial of computerized LPA in third grade for the support of reading comprehension development – specifically for children that were initially low achievers in this domain of reading comprehension. More longitudinal studies, however, would be necessary to see whether these beneficial effects can stabilize the reading acquisition process long-term and whether - if other pre- and posttest measures are used – these beneficial results can also be found in other grades.

**Efficacy of LPA: Broad or Specific?**

Our results show that the positive effects of an implementation of LPA in reading comprehension measures were not found for other variables measuring literacy and literacy-related skills. Neither the positive treatment group effect nor the interaction effect of treatment group and initial achievement skill was present in models that had decoding or spelling at the end of the school year as outcome measures. This was the case even though decoding, spelling and reading comprehension were moderately correlated ($r = 0.5 – 0.75$). In addition, decoding - as an underlying skill of reading comprehension (Hjetland et al., 2020; Van Viersen et al., 2018; Wang et al., 2019) - was not affected by the LPA implementation. Thus, our results indicate that beneficial treatment effects were specific to the skill that was measured with the LPA and trained in third grade according to the curriculum. This specificity of effects indicates that teachers indeed used the specific data on reading comprehension provided by the computerized LPA to individualize their reading instruction. However, based on our study we cannot say what this individualization looked like. In-depth interviews with the teachers along with classroom observation studies are needed to clarify how teachers used the data provided by the LPA to support children's reading comprehension skills.

## Limitations

Our study had several limitations. Most importantly, assignment of classes to the intervention and control group was not random since teachers could decide freely whether they wanted to participate in the study or in using the LPA. Our data indicate that the two groups differed at pretest. This difference could be connected to the fact that some control classes were from a different region within Germany. In addition, data collection took place in two waves, and the second wave was affected by the Covid-19 pandemic, including three weeks of complete school closures. For the same reason, we had to postpone some of the post-tests during the second year of data collection and had most of the post-tests conducted by teachers. We dealt with this situation by controlling for various variables that we could measure (e.g., year of assessment, region of assessment, cognitive abilities) and we did not find effects indicating any significant interference with the results by these measures. Furthermore, due to the nested data structure we accounted for between-class effects. These effects would be expected to outweigh systematic effects of region or cohort. However, of course, we cannot ensure that our actions to control group differences were sufficient.

## Future Directions

We found that the LPA for reading comprehension was particularly beneficial for students with low initial levels of reading comprehension. Future studies could collect more data on how teachers use the information of the LPA to support these students on an individual level, too. Another interesting question to study would be in what way computerized LPAs can be implemented in a classroom most effectively to achieve the overarching academic goal of teaching all children to read well. Here, the rise of computerized LPA allows for the assessment itself to be tailored to individual students. Following up on our finding that the positive effects of LPA implementation are specific to the skills that were measured by the LPA, an important

question for future research is whether the LPA should be tailored to individual needs of students (e.g., LPA for decoding if children have difficulties in that domain), continue to be tailored to the curriculum (e.g., reading comprehension in third grade) or whether both approaches should be combined.

**Summary**

In summary, this study corroborated earlier studies that found that computerized LPA helps to foster third grade students' reading comprehension. Going beyond these previous studies, we found that children with low levels of reading comprehension at the beginning of the school year profited most from computerized LPA. These effects were specific to reading comprehension and did not transfer to cognate literacy skills such as decoding or spelling. Taken together, these findings suggest that teachers used the specific data provided by the LPA to identify children with reading difficulties and to adapt their teaching to their specific deficits, which leads to an accelerated increase in reading comprehension ability that is specific to this group. Future studies on the efficacy of LPA implementation are needed that combine these data with information about how exactly teachers used the data provided by the LPA to support children's reading comprehension skills. The identification of factors influencing the success of computerized LPAs is of high relevance for boosting individualized instruction in regular classrooms. Knowledge of these factors could be incorporated in teacher trainings in order to improve their data-based decision making.

## References


Bartoń, K. (2019). MuMIn: Multi-Model Inference. *R package version 1.43.15.* https://cran.r project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.

Bishop, D. V. M., & Snowling, M. J. (2004). Developmental dyslexia and specific language impairment: Same or different? *Psychological Bulletin, 130*, 858–886. https://doi.org/10.1037/0033-2909.130.6.858

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5-51. https://doi.org/10.1177/1529100618772271

Compton, D.L., Olinghouse, N.G., Elleman, A., Vining, J., Appleton, A.C., Vail, J. & Summers, M. (2005). Putting Transfer Back on Trial: Modeling Individual Differences in the Transfer of Decoding-Skill Gains to Pther Aspects of Reading Acquistion. *Journal of Educational Psychology 97*, 55-59. https://doi.org/10.1037/0022-0663.97.1.55

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219 –232. https://doi.org/10.1177/00144 0298505200303

Espin, C., McMaster, K. L., Wayman, M. M., & Rose, S. (Eds.). (2012). *A measure of success: The influence of curriculum-based measurement on education*. Minneapolis, MN: University of Minnesota Press.

Fox, J. and Weisberg, S. (2019). An {R} Companion to Applied Regression, Third Edition Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, *41*, 163-171. https://doi.org/10.3758/BRM.41.1.163

Förster, N., Kawohl, E., & Souvignier, E. (2018). Short-and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and

reading comprehension. *Learning and Instruction*, *56*, 98-109. https://doi.org/10.1016/j.learninstruc.2018.04.009

Förster, N., & Souvignier, E. (2011). Curriculum-based measurement: Developing a computer-based assessment instrument for monitoring student reading progress on multiple indicators. *Learning Disabilities: A Contemporary Journal, 9*, 21–44.

Fox, J. and Weisberg, S. (2019). *An {R} companion to applied regression (Third Edition)*. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fuchs, L.S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement, using a reading maze task. *Exceptional Children*, *58*(5), 436-450. https://doi.org/10.1177/001440299205800507

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*(3), 199-208. https://doi.org/10.1177/001440298605300301

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, *21*(2), 449-460. https://doi.org/10.3102/00028312021002449

Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1-56. http://dx.doi.org/10.1016/0049-089X%2892%2990017-B

Ganzeboom, H. B. G. (2010, May). *A new International Socio-Economic Index (ISEI) of occupational status for the International Standard Classification of Occupation 2008 (ISCO-08) constructed with data from the ISSP 2002-2007.* Paper presented at the Annual Conference of International Social Survey Programme, Lisbon.

Gelderblom, G., Schildkamp, K., Pieters, J., & Ehren, M. (2016). Data-based decision making for instructional improvement in primary education. *International Journal of Educational Research*, *80*, 1-14. https://doi.org/10.1016/j.ijer.2016.07.004

Hebbecker, K., & Souvignier, E. (2018). Formatives Assessment im Leseunterricht der Grundschule – Implementation und Wirksamkeit eines modularen, materialgestützten Konzepts. Zeitschrift für Erziehungswissenschaft, 21, 735-765. https://doi.org/10.1007/s11618-018-0834-y

Hjetland, H. N., Lervåg, A., Lyster, S. A. H., Hagtvet, B. E., Hulme, C., & Melby-Lervåg, M. (2019). Pathways to reading comprehension: A longitudinal study from 4 to 9 years of

age. *Journal of Educational Psychology, 111*, 751–763. https://doi.org/10.1037/edu0000321

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

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E. M., ... & Valtin, R. (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster; New York: Waxmann.

Ise, E., Engel, R. R., & Schulte-Körne, G. (2012). Was hilft bei der Lese-Rechtschreibstörung? *Kindheit und Entwicklung*, *21*, 122-136. https://doi.org/10.1026/0942-5403/a000077

Jenkins, J. R. & Fuchs, L. S. (2012). Curriculum-based measurement: The paradigm, history, and legacy. In C., A. EspinK., L. McMasterS., RoseM., M. Wayman (Eds.), *A measure of success. The influence of curriculum-based measurement on education* (pp. 7–23). Minneapolis, MN: University of Minnesota Press.

Jung, P. G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of data-based individualization for students with intensive learning needs: A meta-analysis. *Learning Disabilities Research & Practice*, *33*(3), 144-155. https://10.1016/j.lindif.2011.11.017

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational measurement: Issues and practice*, *30*(4), 28-37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Lai, M. K., & Schildkamp, K. (2013). Data-based decision making: An overview. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based Decision Making in Education* (pp. 9-21). Dordrecht: Springer.

Lembke, E., McMaster, K. L., & Stecker, P. M. (2012). Technological applications of Curriculum-Based Measurement in elementary settings. In C. A. Espin, K.L. McMaster, S.Rose, M.M. Wayman (Eds.), *A Measure of Success: The Influence of Curriculum-based Measurement on Education* (*pp*. 125-138). Minneapolis, MN: University of Minnesota Press.

Lenhard, W., Lenhard, A., & Schneider, W. (2017). *ELFE II – Ein Leseverständnistest für Erst- bis Siebtklässler*. Göttingen: Hogrefe.

Leppänen, U., Aunola, K., Niemi, P., & Nurmi, J. E. (2008). Letter knowledge predicts Grade 4 reading fluency and reading comprehension. *Learning and Instruction*, *18*(6), 548-564. https://doi.org/10.1016/j.learninstruc.2007.11.004

Lervåg, A., Hulme, C., & Melby-lerv, M. (2018). Unpicking the developmental relationship between oral language skills and reading comprehension: It's simple, but complex. *Child Development, 89*, 1821–1838. http://doi.org/10.1111/cdev.12861

Marcoulides, K. M., & Raykov, T. (2019). Evaluation of variance inflation factors in regression models using latent variable modeling methods. *Educational and Psychological Measurement*, *79*(5), 874-882. https://doi.org/10.1177/0013164418817803

Moll, K., & Landerl, K. (2009). Double dissociation between reading and spelling deficits. *Scientific Studies of Reading*, *13*(5), 359-382. https://doi.org/10.1080/10888430903162878

Moll, K., & Landerl, K. (2010). *SLRT-II: Lese-und Rechtschreibtest; Weiterentwicklung des Salzburger Lese-und Rechtschreibtests (SLRT)*. Zürich: Huber.

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., ... & Landerl, K. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learning and Instruction*, *29*, 65-77. https://doi.org/10.1016/j.learninstruc.2013.09.003

McArthur, G., Badcock, N., Castles, A., & Robidoux, S. (2021). Tracking the relations between children's reading and emotional health across time: Evidence from four large longitudinal studies. *Reading Research Quarterly*. https://doi.org/10.1002/rrq.426

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. Boston College, TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/pirls2016/international-results/

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673-690. https://doi.org/10.1007/s11135-006-9018-6

Peters, M. T., Förster, N., Hebbecker, K., Forthmann, B., & Souvignier, E. (2021). Effects of data-based decision-making on low-performing readers in general education classrooms: Cumulative evidence from six intervention studies. *Journal of Learning Disabilities*, https://doi.org/10.1177/00222194211011580

Sala, G., Aksayli, N. D., Tatlidil, K. S., Tatsumi, T., Gondo, Y., Gobet, F., ... & Verkoeijen, P. (2019). Near and far transfer in cognitive training: A second-order meta-analysis. *Collabra: Psychology*, *5*(1). https://doi.org/10.1016/j.tics.2018.10.004

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, *103*, 101602. https://doi.org/10.1016/j.ijer.2020.101602

Schmitterer, A. M., & Brod, G. (2021). Which data do elementary school teachers use to determine reading difficulties in their students?. *Journal of Learning Disabilities*, 0022219420981990. https://doi.org/10.1177/0022219420981990

Schütze, B., Souvignier, E., & Hasselhorn, M. (2018). Stichwort - formatives Assessment. *Zeitschrift für Erziehungswissenschaft*, *21*, 697–715. doi:10.1007/s11618-018-0838-7

Share, D.L. (1999). Phonological recoding and orthographic learning: A direct test of the self-teaching hypothesis. *Journal of Experimental Child Psychology*, *72*(2), 95-129. https://doi.org/10.1006/jecp.1998.2481

Souvignier, E., Förster, N., Hebbecker, K., & Schütze, B. (2021). quop: An effective web-based approach to monitor student learning progress in reading and mathematics in entire classrooms. In S. Jornitz & A. Wilmers (Eds.), *International perspectives on school settings, education policy and digital strategies: A transatlantic discourse in education research* (pp. 283–298). Budrich

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S., & Haag, N. (2017). *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich*. Münster; New York: Waxmann.

Strathmann, A. M., & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42,* 111-122. https://doi.org/10.1026/0049-8637/a000011

van Viersen, S., de Bree, E. H., Zee, M., Maassen, B., van der Leij, A., & de Jong, P. F. (2018). Pathways into literacy: The role of early oral language abilities and family risk for dyslexia. *Psychological Science*, *29*(3), 418-428. https://doi.org/10.1177/0956797617736886

Walter, J. (2014). Lernfortschrittsdiagnostik Lesen (LDL) und Verlaufsdiagnostik sinnerfassenden Lesens (VSL): Zwei Verfahren als Instrumente einer formativ orientierten Lesediagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 166 –201). Göttingen: Hogrefe.

Wang, Z., Sabatini, J., O'Reilly, T., & Weeks, J. (2019). Decoding and reading comprehension: A test of the decoding threshold hypothesis. *Journal of Educational Psychology*, *111*(3), 387 –401. https://doi.org/10.1037/edu0000302

Weiß, R. H., & Osterland, J. (2013). *Grundintelligenztest Skala 1-Revision: CFT 1-R*. Göttingen: Hogrefe.

Zeuch, N., Förster, N., & Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learning Disabilities Research & Practice*, *32*(1), 61-70. https://doi.org/10.1111/ldrp.12126

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*(1), 3-29. doi: 0.1037/0033-2909.131.1.3

# Appendix

**Table A1**

*Results of the Multilevel Model focusing on the Effect of initial Achievement on Reading Abilities at the End of Third Grade as a Function of Attendance of an LPA Class in the Small Sample*

| **Fixed Effects** | **Estimate** | ***SE*** | ***t* value** |
|---|---|---|---|
| Intercept | 21.67 | 2.98 | 7.29*** |
| Reading Comprehension at T1 | 0.98 | 0.04 | 21.97*** |
| Treatment Group (LPA/ No LPA) | 9.04 | 4.22 | 2.14 * |
| Year of Assessment (Corona/ No Corona) | 1.78 | 2.17 | 0.82 |
| Region (Hesse/ Lower Saxony) | 0.19 | 3.16 | 0.06 |
| Reading Comprehension at T1 * TGroup Interaction | -0.17 | 0.07 | -2.64* |

Note. $^{***}$ $p < .001$; $^{**}$ $p < .01$; $^{*}$ $p < .05$; For this analysis data from 344 students with full data sets in group and individual sessions were used.

**Table A2**

*Overview of relevant Sample Characteristics in the Large and Small Samples at T1*

| | Large Sample ($N$ = 668) | | Small Sample1 ($N$ = 344) | | Small Sample2 ($N$ = 230) | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Reading Comprehension | 62.16 | 20.44 | 52.05 | 16.9 | 52.47 | 16.92 |
| Decoding | -- | -- | 28.45 | 9.89 | 28.81 | 10.39 |

*Note.* Small Sample 1 represents all children with full data sets in individual and group sessions at T1 in both waves; Small Sample 2 represents all children with full data sets in individual and group sessions at T1 for whom we could also measure decoding proficiency during the second wave at T2 during Covid-19 distancing measures.

**Table A3**

*Results of the Multilevel Model focusing on the Effect of initial Achievement on Reading Abilities at the End of Third Grade as a Function of Attendance of an LPA Class, controlling for language background*

| **Fixed Effects** | **Estimate** | ***SE*** | ***t* value** |
|---|---|---|---|
| Intercept | 24.64 | 2.60 | 9.51*** |
| Reading Comprehension at T1 | 0.92 | 0.03 | 27.95*** |
| Treatment Group (LPA/ No LPA) | 6.08 | 3.24 | 1.88 |
| Year of Assessment (Corona/ No Corona) | -0.29 | 1.58 | -0.19 |
| Region (Hesse/ Lower Saxony) | 2.62 | 2.48 | 1.06 |
| Bilingual | -0.18 | 1.01 | -0.18 |
| German as second language | -1.21 | 1.17 | -1.04 |
| Reading Comprehension at T1 * TGroup Interaction | -0.09 | 0.04 | -2.11* |

Note. $^{***}$ $p < .001$; $^{**}$ $p < .01$; $^{*}$ $p < .05$; For this analysis data from 667 students were used. For the factor language background contrast coding was used with the reference level German as first language; Pseudo-$R^2 = 0.74$.

**Table A4**

*Results of the Multilevel Model focusing on the Effect of initial Achievement on Reading Abilities at the End of Third Grade as a Function of Attendance of an LPA Class, controlling for nonverbal intelligence.*

| **Fixed Effects** | **Estimate** | ***SE*** | ***t* value** |
|---|---|---|---|
| Intercept | 22.07 | 3.20 | 6.90*** |
| Reading Comprehension at T1 | 0.92 | 0.04 | 27.51*** |
| Treatment Group (LPA/ No LPA) | 6.40 | 4.22 | 1.99* |
| Year of Assessment (Corona/ No Corona) | -0.42 | 1.58 | -0.27 |
| Region (Hesse/ Lower Saxony) | 2.74 | 3.16 | 1.09 |
| Nonverbal Intelligence | 0.07 | 0.07 | 1.00 |
| Reading Comprehension at T1 * TGroup Interaction | -0.09 | 0.04 | -2.18* |

Note. *** $p < .001$; ** $p < .01$; * $p < .05$; For this analysis data from 667 students were used; Pseudo-$R^2 = 0.74$.