Hermann Astleitner (Ed.)

Intervention Research in Educational Practice

Alternative Theoretical Frameworks and Application Problems

WAXMANN

Hermann Astleitner (Ed.)

Intervention Research in Educational Practice

Alternative Theoretical Frameworks and Application Problems

Waxmann 2020
Münster • New York

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the
Deutsche Nationalbibliografie; detailed bibliographic data

Print-ISBN 978-3-8309-4197-2
Ebook-ISBN 978-3-8309-9197-7
https://doi.org/10.31244/9783830991977

Münster, Germany

www.waxmann.com
info@waxmann.com

Cover Design: Anne Breitenbach, Münster
Typesetting: MTS. Satz & Layout, Münster
Print: CPI Books GmbH, Leck

Contents

Contributors

Hermann Astleitner, A.Univ.-Prof. Mag. Dr.
https://www.uni-salzburg.at/erz/hermann.astleitner
Department of Educational Science
University of Salzburg
Erzabt-Klotz-Str. 1, 5020 Salzburg, Austria

Josef Eisner, Senior Lecturer Mag. Dr.
https://www.uni-salzburg.at/index.php?id=211343
Ulrike Greiner, Univ.-Prof. MMag. DDr.
https://www.uni-salzburg.at/index.php?id=57561
Franz Hofmann, Ao.Univ.-Prof. MMag. Dr.
https://www.uni-salzburg.at/index.php?id=211348
Michaela Katstaller, Mag. Dr.
https://www.uni-salzburg.at/index.php?id=60825
School of Education
University of Salzburg
Erzabt-Klotz-Str. 1, 5020 Salzburg, Austria

Nomy Dickman, Dr.
https://medicine.biu.ac.il/en/node/866
The Azrieli Faculty of Medicine
Bar-Ilan University
Ramat Gan 5290002, Israel

Gabriela Gniewosz, Dr.
https://diagnostik.sbg.ac.at/team/gniewosz/
Department of Psychology
University of Salzburg
Hellbrunner Straße 34, 5020 Salzburg, Austria

Stephanie Musow, MA
https://www.phsg.ch/de/team/ma-stephanie-musow
St. Gallen University of Teacher Education (PHSG)
Notkerstrasse 27
9000 St. Gallen, Switzerland

Alexander Naumann, Dr.
https://www.dipf.de/de/institut/personen/naumann-alexander
DIPF, Leibniz Institute for Research and Information in Education
Rostocker Straße 6, 60323 Frankfurt/Main, Germany

Foreword

The greatest difficulties lie where we are not looking for them.
Johann Wolfgang von Goethe (1821/2000, p. 484)

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.
John Wilder Tukey (1962, p. 13–14)

This book aims to address special problems in "educational interventions" which are "measures that attempt to solve problems in the field of education" embedded within training, instructing, coaching, or counseling (Astleitner, 2010, p. 48). Such interventions are "purposeful actions" which "operate at the individual, family, organizational (e.g., school), neighborhood, regional, national, or other level. Interventions may be comprised of a single action or a cluster of actions" (Fraser, Richman, Galinsky, & Day, 2009, p. 5).

The selected problems in this book concern different areas of educational intervention research and are related to theoretical foundations as well as research designs and measurements. We focus on problems which are often overlooked in recent intervention research or excluded from scientific discourse although problem-related solutions could, above all, reduce the theory-practice gap or build links between basic research, applied research, and educational practice (e.g., Astleitner, 2003; Herber, 1998; Patry, 2001; Zumbach & Mandl, 2008). The given book should therefore be a supplement to existing standard approaches on educational intervention research (e.g., McBride, 2016; Phyne, Robinson, & Levin, 2005; Riley-Tillman & Burns, 2009; Stylianides & Childs, 2019).

Our problems and related studies were embedded in practical settings in the fields of higher education, teacher education, and teaching-learning-research.

A first major problem on theories in educational intervention research lies in distinguishing and handling successfully different types of theories from basic research, applied research, and educational practice. In particular, traditional theories suffer from their limited capacities in connecting different stages of human development with adaptive support strategies, in optimizing context sensitivity during implementation, and in handling the dark side of educational experiences. We will show, in chapter 1, how development-support, implementation-, and dark-side-theories could stimulate educational intervention research.

A second problem focuses on ineffective interventions in the field of teaching. Interventionistic researchers and practitioners need to know what kind of ineffective interventions and why ineffectiveness are given. Here, different types of teach-

ing malfunctions are identified. The contribution in chapter 2 also aims at formulating a macro- and a micro-theory about why teaching malfunctions occur.

Also, an important third problem in exploring effects patterns of educational interventions is given when a control group is missing because of practical, ethical, or other constraints. Here, we will demonstrate within an empirical study, in chapter 3, how a control group can be simulated and therefore have an impact on the conclusiveness of findings from quasi-experimental settings.

A fourth problem is about inconclusive or negative research findings. Even having a sophisticated theory- and research-based foundation within an educational intervention study will not guarantee that findings are significant and confirming given hypotheses. Sometimes, educational interventions improve some goal areas, but worsen others at the same time. In chapter 4, we will show how to provide theoretical and statistical analyses as well as a focus on alternative approaches in research design in order to handle such problems.

A fifth problem in educational intervention research is about the invalidity of measurements, especially when they are multidimensional in nature. Sometimes, it is necessary in educational intervention research to use measurement instruments without having the possibility to pretest them. Low validity of the measurement instruments might be the consequence. In chapter 5, we will show how statistical procedures on analyzing construct validity like confirmatory factor analysis can be used in order to test and handle these problems.

A sixth problem concerns a situation in which the correlations of pretests from different measurements are different for the posttests. When the correlations are different, then one might conclude that also the reliabilities and validities of measurements are different. When the reliability and validity of pretests are different from those on the posttests, then effect patterns in interventions (as differences between pre- and posttests) might be affected. In chapter 6, we will show and reflect a study in which such problems occurred.

A seventh problem is on valid interpretations and uses of test scores. In chapter 7, there is a focus on the question whether tests are appropriate or inappropriate to evaluate intended intervention goals. Here, the focus is on instructional sensitivity as a property of a test to capture effects of classroom interventions. The complex concept of instructional sensitivity appears to be very challenging for intervention research in educational practice and future research will show whether it can be beneficial.

Finally, an eight problem is on the question whether participants of an intervention are willing to give their best in a test on intervention effects. In order to answer this question, in chapter 8, a comprehensive overview of the state-of-the-art on test-taking motivation is given. Also, practical suggestions are depicted, how problems on test-taking motivation should be handled in educational intervention studies.

All problems are embedded in different types of research works, all of them can be found within the context of educational intervention research. There are concep-

tual analyses, theory building approaches, quasi-experimental intervention studies, or measurement validity studies. In case of the empirical studies, some of the mentioned problems were identified after the study was done. So, some of the problems were not first or starting points, but were found during the research process. Therefore, especially, within most of the empirical studies in this book, the orientation on the mentioned problems remains subdominant. This problem orientation is supplemented by innovative topics in the field of higher education resp. university teaching, teacher education and teaching-learning activities like alternative types of intervention theories, volitional competences, minimal guidance interventions, developmental measurement approaches, and so on.

Of course, these selected problems do not provide a complete list of problems in educational intervention research. We have focused on problems which are highly relevant for research activities in educational practice, but which are, at the same time, not prominently anchored especially in the field of teacher training or teaching-learning activities. It is also clear, that our way of handling such problems depends on our theoretical and methodological focus. This focus has multiple facets. These facets can be made explicit in the context of another important problem in educational intervention research concerning multiple testing.

The Focus of This Book Exemplified by the Problem of Multiple Testing in Educational Intervention Research

In educational intervention research, it is often important to know effects of an intervention on multiple dependent variables. This situation implies that multiple statistical tests are used in one study. However, when conducting multiple tests, the problem of overall Type I error inflation (i.e. of rejecting the null hypothesis when it is true) occurs. Therefore, Alpha-levels should be adjusted on a lower level than the conventional .05 level. For such adjustments, many different methods can be found within the literature, some of them in different advanced versions and with available statistical software (e.g., Bretz, Hothorn, & Westfall, 2011; Shaffer, 1995). The most prominent procedure is an easy-to-use Bonferroni method in which each hypothesis is tested at an adjusted significance level of Alpha/n, whereas n is the number of hypotheses. Ignoring such methods has a strong impact in the scientific community: It could lead to false discoveries and result in wrong decisions on the effectiveness of interventions.

However, there are also critiques of such adjustment methods (Fink, McConnell, & Vollmer, 2014; O'Keefe, 2003): First, many correction methods dependent on the number of tests and therefore on the number of variables in a study. However, it is not always clear how many variables were in fact in the original study plan. Sometimes, researchers only report results of some pre-selected variables and not on all variables in a study. Second, there is some dispute on the question what counts as a distinct hypothesis. Some argue that all or many variables in a study are correlated,

and thus reflect only one single hypothesis. When there is only one hypothesis, then no adjustments of the Alpha-level are necessary. Third, reducing the Alpha-level leads to a problem concerning statistical power resp. Type II error. It reduces the chance of detecting a genuine (nonzero) effect. Forth, there is an inconsistent application of adjustment methods in a way that some researchers are applying such methods and some researchers in the same area of research are not using such procedures. As a consequence, the same results are sometimes statistically significant and sometimes confusingly non-significant.

Despite these problems, there is no doubt that Alpha-level adjustment is necessary especially in basic research, in which researchers have to be very conservative about accepting new evidence from experiments. However, in applied research as in the field of intervention research in educational practice, the situation is different in a way that higher Alpha-levels (e.g., $p > .05$) are accepted under certain circumstances. For example, Lipsey (1990, p. 39) wrote:

> A promising treatment might be investigated to determine if it has beneficial effects in some problem area. In such an applied research the implications of errors of inference may be quite different from those in basic research. To 'discover' that an applied treatment is effective when, in fact, it is not, does indeed mislead practitioners just as the analogous case misleads theoreticians. Practitioners, however, are often in situations where they must act as effectively as they can irrespective of the state of their formal knowledge, and it is unusual for them to use treatments and techniques of plausible but unproven efficacy. Moreover, demonstrably effective treatments for many practical problems are not easy to come by and candidates should not be to easily dismissed.

Educational intervention research is not only situated in the field of applied research, but also has often an exploratory goal focus. Within educational intervention research, this complex situation also leaded to the formulation of sophisticated guidelines for multiple testing. In such guidelines, it was suggested that "multiplicity adjustments are not required for exploratory analyses" (Schochet, 2008, p. 6), however that

> reports should explicitly state that exploratory analyses do not provide rigorous evidence of the intervention's overall effectiveness. Results from post hoc analyses should be reported as providing preliminary information on relationships in the data that could be subject to more rigorous future examination.

Others argued that in such exploratory analyses and related significance testing, *p*-values lose their meaning due to an unknown inflation of the Alpha-level and therefore research should focus on alternative criteria and methods instead of significance testing (e.g., simple graphic techniques) (e.g., Harlow, Mulaik, & Steiger, 1997). Recently, Rubin (2017, p. 272) stated that whether *p*-values lose their meaning depends on how we define the "family" of the error rate and argued that

> it is not necessary to lower the nominal Alpha-level when undertaking single tests of several different hypotheses. Exploratory analyses often involve many tests of this type. Consequently, Alpha-level adjustments are less necessary in exploratory analyses than would be the case if researchers adopted a multiple hypotheses approach to the familywise error rate. Again, it remains the case that the more hypotheses that a researcher tests, the greater probability that they will make a Type I error. However, this increased probability is distributed across the entire collection of hypotheses that are tested rather than localized to any one specific hypothesis. Consequently, it does not threaten the validity of any single test. If (a) researchers are interested in determining whether evidence supports or falsifies specific hypotheses rather than amorphous collections of hypotheses (i.e., universal null hypotheses), and (b) the probability that one hypothesis is true does not influence the probability that another hypothesis is true, then there is no need to adjust Alpha-levels for single tests of multiple hypotheses.

Finally, Trafimow et al. (2018) argued that manipulating the Alpha-level cannot cure significance testing and that the relative importance of Type I and Type II errors might differ across fields of research, studies, and researchers. Factors contributing to such differences include, for example, the clarity of theory, auxiliary assumptions, practical or applied concerns, or experimental rigor.

In this book, we reacted with the following strategies on these discussions in the field of multiple testing and other related problems:

a) We are clear about that our empirical studies represent applied research and have a significant exploratory orientation. In our studies, we tested certain interventions, or measurement approaches as well as related hypotheses for the first time. We will formulate and substantiate specific hypotheses, but most of them are in an early more or less exploratory state of theory development or scientific progress. Therefore, our findings are preliminary and have to be re-tested in future studies before they can be applied in educational practice.

b) As our empirical studies are situated in exploratory and applied settings, we will do no Alpha-level adjustments. We deliver exact *p*-values and therefore the possibility to evaluate their significance in case of simple Alpha-level adjustments. We assume that many of our findings will be non-significant if highly conservative Alpha-level adjustments would be made.

c) In educational practice in general, there are many factors that would decrease statistical power like low treatment integrity, small sample sizes, or floor or ceiling effects (Lipsey, 1990, p. 171). In the empirical studies in this book, there are some of these factors given, especially small sample sizes, due to uncontrollable situations in educational practice. In view of this problem, we do not apply Alpha-level adjustments because they would further decrease an already low power.

d) Of course, we are aware of the fact that multivariate testing would allow to give up Alpha-level adjustments resp. handle them effectively. However, we avoided multivariate testing due to methodological limitations of our studies. Multivariate testing, for example, with repeated measures multivariate analysis of variance would

Tab. 1: Differences between Basic Research, Applied Research, and Practice (based on: Astleitner, 2018, p. 149)

Dimensions	Basic Research	Applied Research	Practice
Theory	Scientific theory	Technological theory, program theory	Subjective, implicit theory
Intervention	Experiment	Quasi-experiment, design-experiment	Case study
Significance	Statistical significance	Practical significance	Success
Validity	Validity of measurement and design	Usefulness	Problem solving
Measurement	Testing	Checking	Estimating
Explanation	Causality	Plausibility	Trial-and-error

need adequate sample size, no univariate or multivariate outliers, multivariate normality, no multicollinearity, and others. On the one hand, such criteria are often not reached in small group studies in educational practice. On the other hand, results of multivariate testing are not significantly different from univariate testing in many cases (e.g., Keselman, Algina, & Kowalchuk, 2001). For example, within this book, in the study from Astleitner, Katstaller, and Greiner, five univariate analysis of variance produced the following 15 *p*-values (for five variables and related group, time, and group x time effects): .996, .815, .024, .303, .164, .073, .074, .079, .474, .336, .029, .796, .840, .042, and .421. A repeated multivariate analysis of variance produced the following *p*-values for the same tests: .594, .984, .039, .167, .156, .036, .085, .089, .446, .087, .016, .560, .755, .013, and .231. Only in 1 of 15 tests, the decisions on the hypotheses would be different: 0.073 (as non-significant) and 0.036 (as significant). Overall, *p*-values in both tests correlated with $r = .913$ ($p < .001$). The differences are mainly due to the fact that in multivariate analysis, all variables are included into the tests, what – due to missing cases – changes means and standard deviations as well as the resulting *F*-Tests.

e) Finally, we have to admit that our studies in this book suffer from typical methodological problems (e.g., small sample sizes) in the field of intervention research in educational practice. From a very strict experimental point of view, these shortcomings suggest to focus on descriptive results of our studies only. Therefore, in our studies, we do not only report and discuss statistical tests, but also descriptive statistics. An experimental psychologist in basic research might consider such descriptive information only, however, an educational researcher in applied and practical settings, would also consider results on hypotheses testing. Overall, we have to stress that there are significant differences between basic and applied research settings and that our studies are situated in applied resp. practical settings (e.g., Astleitner, 2013). Astleitner (2018, p. 149) outlined such differences (see Table 1). For example, in basic research, there is a focus on scientific theories (as classical if-then-statements). In applied research, we also focus on technological or program theories

(as if-do-statements). In practice, practitioners have subjective or implicit theories (as personal opinions).

Case-based Learning

Having such differences in view of other areas of research and facing such problems in educational intervention research, lead to the conclusion that our findings should not be interpreted as effectiveness tests, but rather as possibilities for case-based learning in educational intervention settings. Zumbach, Haider, and Mandl (2008, p. 1) argued that such a perspective allows that learners "acquire knowledge through authentic problems from multiple perspectives, combining both foundations and application". To stimulate comprehensively and sustainably the acquisition of such a perspective should be an important goal of this book. This perspective lies in line with concepts of research like "transformative research" (e.g., Mertens, 2009) or "translational science" in which the major goal is to shorten and optimize the gap between knowledge production in science and application in practical contexts (e.g., Wehling, 2015). Therefore, the book is for researchers, instructors, designers, evaluators, or practitioners in many areas of education, especially on teacher education and teaching-learning activities. The book could be used within innovative scenarios on transdisciplinary educational approaches in which basic researchers, applied researchers, and practitioners work together to solve problems in educational practice (e.g., Ciesielski, Aldrich, Marsit, Hiatt, & Williams, 2017).

Multiple testing and related problems represent significant areas of concern in the field of educational intervention research. This book will show that there are many other problems to be found. However, despite all the difficulties and shortcomings in intervention research in educational practice, this book should be seen as an attempt to connect educational research and practice on a comprehensive and sophisticated basis.

Although we had a strong and farsighted scientific focus in our empirical studies, many of our results were non-significant or even negative. In our opinion, this is nothing unusual in highly complex and dynamic applied and exploratory settings, even if one can hardly find it in published work in professional journals. Results of this book stressed that educational activities, like, for example, teacher education might, at least partially, be ineffective and that this severe problem has to be taken seriously. In this book, we will present different explanations for such problems on a theoretical, design- as well as measurement-based orientation. Of course, all these explanations are preliminary and have to be tested in future research activities. Our book could represent a starting point for such attempts.

Of course, such an attempt needs the support of the scientific community. I would like to thank all the authors in this book for bringing their interest, creative ideas, and scientific skills to their contributions. For the different chapters in the book, the stated authors and co-authors were involved significantly in different stages of the research process (like planning, conducting, data analysis, and writing as

well as reviewing). Reviewing was conducted in a way that all co-authors delivered feedback to the texts and contributions of the other co-authors. The whole research was orientated on guidelines for good scientific practice and data protection from the University of Salzburg (retrieved from https://www.uni-salzburg.at). All participants in our studies were informed about the study goals, participated in the studies voluntarily and were able to cancel their participation at any time. Women were over-represented in most studies, but the generalizability of the results is not given anyway. All data was properly saved and statistical analyses were double checked. We also have evidence, that, as assumed before the interventions, control groups did not suffer from ethically questionable disadvantages in comparison to the intervention groups.

We have to thank all the participants in the studies. Thankfully, Jackie and Jörg Sams supported us in editing some texts in relation to the English language. We also thank for support by the Department of Educational Science and the School-of-Education of the University of Salzburg. We are also grateful for the help of student assistants in data entry and data management. We also want to thank Jörg Zumbach (University of Salzburg) for his critical, but true and constructive comments, especially on the problem of multiple testing. Thanks also to Hans-Jörg Herber and Jean-Luc Patry (University of Salzburg) who had and have a strong influence on my theoretical and methodological thinking. I was inspired by the shared work and encouraged to look at problems from multiple different perspectives what made me open and curious about future challenges in the field of educational intervention research. May that also be true for the readers of this book!

Hermann Astleitner
Salzburg, Spring 2020

References

Astleitner, H. (2003). Praktische Signifikanz. *Journal für Lehrerinnen- und Lehrerbildung*, *3*, 48–53. Retrieved from https://www.studienverlag.at/zeitschriften/journal-fuer-lehrerinnen-und-lehrerbildung/

Astleitner, H. (2010). Methodische Rahmenbedingungen zur Entdeckung der Wirksamkeit von pädagogischen Interventionen [Methodological conditions for identifying the impact of educational interventions]. In T. Hascher & B. Schmitz (Eds.), *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* (pp. 48–62). Weinheim: Juventa.

Astleitner, H. (2013, December). *Ist die Schulforschung naiv?* [Is school research naive?] Paper presented at the research colloquium of the School-of-Education, University Salzburg, Salzburg. Retrieved from https://www.uni-salzburg.at/fileadmin/multimedia/Erziehungswissenschaft/SOE2013_8.pdf

Astleitner, H. (2018). *Spezielle Verfahren sozialwissenschaftlicher Theorieentwicklung* [Special methods of theory building in social research]. Weinheim: Beltz Juventa.

Bretz, F., Hothorn, T., & Westfall, P. (2011). *Multiple comparisons using R*. Boca Raton: Chapman and Hall/CRC.

Ciesielski, T. H., Aldrich, M. C., Marsit, C. J., Hiatt, R. A., & Williams, S. M. (2017). Transdisciplinary approaches enhance the production of translational knowledge. *Translational Research*, *182*, 123–134. doi: https://doi.org/10.1016/j.trsl.2016.11.002

Fink, G., McConnell, M., & Vollmer, S. (2014). Testing for heterogeneous treatment effects in experimental data: False discovery risks and correction procedures. *Journal of Development Effectiveness*, *6*, 44–57. doi: https://doi.org/10.1080/19439342.2013.875054

Fraser, M. W., Richman, J. M., Galinsky, M. J., & Day, S. H. (2009). *Intervention research. Developing social programs*. New York: Oxford University Press.

Goethe, v. J. W. (1821/2000). *Werke* (Band 8). München: Deutscher Taschenbuch Verlag.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah: Erlbaum.

Herber, H. J. (1998). Theorien und Modelle der Pädagogik, Psychologie und pädagogischen Psychologie – Annäherungsmöglichkeiten an ein komplexes Beziehungsproblem [Theories and models of pedagogy, psychology and pedagogical psychology – Approaching a complex relationship problem]. *Salzburger Beiträge zur Erziehungswissenschaft*, *2*, 41–101.

Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, *54*, 1–20. doi: https://doi.org/10.1348/000711001159357

Lipsey, M. R. (1990). *Design sensitivity. Statistical power for experimental research*. Newbury Park, CA: Sage.

McBride, N. (2016). *Intervention research: A practical guide for developing evidence-based school prevention programmes*. Singapore: Springer. doi: https://doi.org/10.1007/978-981-10-1011-8

Mertens, D. M. (2009). *Transformative research and evaluation*. New York: Guilford.

O'Keefe, D. J. (2003). Colloquy: Should familywise alpha be adjusted? Against familywise alpha adjustment. *Human Communication Research*, *29*, 431–447. doi: https://doi.org/10.1111/j.1468-2958.2003.tb00846.x

Patry, J. L. (2001). Situation specificity of behavior: Triple relevance for research and practice in social research. *Salzburger Beiträge zur Erziehungswissenschaft*, *5*, 41–62. Retrieved from https://www.researchgate.net/profile/Jean_Luc_Patry

Phyne, G. D., Robinson, D. H., & Levin, J. (Eds.). (2005). *Empirical methods for evaluating educational interventions*. Burlington, MA: Elsevier Academic Press.

Riley-Tillman, T. C., & Burns, M. K. (2009). *Evaluating educational interventions. Single-case design for measuring responses to intervention*. New York: Guildford.

Rubin, M. (2017). Do p values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology*, *21*, 269–275. doi: https://doi.org/10.1037/gpr0000123

Schochet, P. Z. (2008). *Guidelines for multiple testing in impact evaluations of educational interventions* (Final report). Princeton: Mathematica Policy Research Inc. Retrieved from https://files.eric.ed.gov/fulltext/ED502199.pdf

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561–584. doi: https://doi.org/10.1146/annurev.ps.46.020195.003021

Stylianides, G. J., & Childs, A. (Eds.). (2019). *Classroom-based interventions across subject areas. Research to understand what works in education*. Abingdon, New York: Routledge. doi: https://doi.org/10.4324/9781315170077

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C. J., Beh, E. J., Bilgiç, Y., et al. (2018). Manipulating the Alpha-level cannot cure significance testing. *Frontiers in Psychology*, *9*, 699. doi: https://doi.org/10.3389/fpsyg.2018.00699

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*, 1–67. doi: https://doi.org/10.1214/aoms/1177704711

Wehling, M. (Ed.). (2015). *Principles of translational science in Medicine* (2nd ed.). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-12-800687-0.00033-5

Zumbach, J., & Mandl, H. (Eds.). (2008). *Pädagogische Psychologie in Theorie und Praxis: Ein fallbasiertes Lehrbuch* [Educational Psychology in theory and practice: A case-based textbook]. Göttingen: Hogrefe.

Zumbach, J., Haider, K., & Mandl, H. (2008). Fallbasiertes Lernen: Theoretischer Hintergrund und praktische Anwendung [Case-based learning: Theoretical background and practical application]. In J. Zumbach & H. Mandl, H. (Eds.), *Pädagogische Psychologie in Theorie und Praxis: Ein fallbasiertes Lehrbuch* (pp. 1–11). Göttingen: Hogrefe.

PART 1.
Theoretical Frameworks

1. Alternative Theoretical Frameworks for Educational Interventions[1*)]

Hermann Astleitner

Abstract: Educational intervention research is based on models of educational effectiveness. The goal of this contribution is to analyze the status of educational effectiveness models by focusing on theoretical concepts and criteria in basic and applied social sciences. This evaluation, which is based on a comprehensive exploratory review of literature, leads to the formulation of three alternative theoretical frameworks: Development-support-, implementation-, and dark-side-theories. Development-support-theories link developmental stages or competence levels (e.g., taxonomies of motivation) with adaptive strategies for establishing attainment-based support mechanisms (e.g., motivational tactics). Implementation-theories are about how interventions have to be designed for optimizing sensitivity to different contexts (like low- or high-quality educational scenarios). Dark-side-theories describe and explain the negative, non-transparent, or faulty facets of interventions (like trust and distrust as co-existing social realities for educational improvement). Discussions reflect on how to stimulate and develop intervention activities by using these alternative theoretical frameworks.

Educational interventions have the general goal to be effective in solving a given problem in educational contexts. Educational effectiveness is indispensable and depends on certain well-established criteria like statistical and practical significance, but also on theories or models of effectiveness.

Challenges for Educational Effectiveness Theories

Existing theories on educational effectiveness which are essential in intervention research have to face multiple challenges arising from current developments in the field (e.g., Astleitner, 2010). Such challenges were identified in this contribution by conducting a comprehensive review of literature on educational effectiveness research and an evaluation of existing shortcomings. Such a review was not undertaken to find effect sizes or other indicators of effectiveness (as in traditional meta-analyses), but to get ideas for advancing theory building and research in the field.

1 *) This chapter is based on an unpublished paper presentation by Astleitner (2019).

It represents an exploratory review for stimulating theory building (e.g., Manzano Vázquez, 2018).

The challenge of integrating multiple content areas, backgrounds, and ineffectiveness. In field of research on educational interventions, there are many specific theories on educational effectiveness (e.g., Maag Merki, Emmerich, & Holmeier, 2017), improvement (e.g., Creemers & Kyriakides, 2012), leadership (e.g., Lynch, 2012), or counseling (e.g., Dollarhide & Lemberger-Truelove, 2019). In addition, there are many background theories related to educational theories (e.g., Irby, Brown, Lara-Alecio, & Jackson, 2013), theories of learning (e.g., Olson & Hergenhahn, 2016), instructional design theories (e.g., Reigeluth, Beatty, & Myers, 2017), or theories from related disciplines like educational psychology (e.g., Furlong, Gilman, & Huebner, 2014). A closer look into these approaches revealed that educational effectiveness theories are highly complex as they have to integrate multiple content areas. It was also discovered that they sometimes did not have consistent relationships to background theories. In addition, it can easily be seen that most approaches focused primarily on effectiveness and not on co-existing ineffectiveness (for example: demotivation (as an indicator of ineffectiveness) is not low motivation (as an indicator of effectiveness), it can exist simultaneously and independently from motivation (e.g., Addison & Brundrett, 2008).

The challenge of different types and criteria of theories. Educational interventions concern basic research, but more often applied settings. Therefore, multiple types of theories and related criteria existed in the field. First, there are traditional scientific or objective theories in basic research (as if-then-statements and related to criteria like accuracy, logical consistency, or testability) (Jaccard & Jacoby, 2010). Second, we dispose of prescriptive technological theories in applied research (as if-do-statements and criteria like usability, efficiency, or usefulness) (Swanson & Chermack, 2013). Third, there are practical or program theories in (interventionist) practice (as a model on how an intervention contributes to processes and outputs with criteria like chain of outcomes, or mechanisms for change) (Funnell & Rogers, 2011). Finally, there are implicit, personal, or subjective theories in individuals (as naive assumptions and criteria like coherence, or richness) (e.g., Barger & Linnenbrink-Garcia, 2017). It is obvious that these different types of theories and criteria are relevant for intervention research, but researchers are often not aware of the type of theory they are using or have difficulties in recognizing and handling their incompatibilities.

The challenge of identifying vital variables. Within social research and therefore also within educational intervention research, theories must have a limited explanatory power or scope in order to be testable in an efficient manner. However, such a limitation should not lead to a situation in which vital variables are missing. Vital variables are ones which are related to many other variables and which have at the same time a strong impact. For example, within a capital theory of effectiveness, variables like the motivation (to invest capital), the resistance (to change profits), the management (of capital flow), or failures (due to crisis in capital systems) should

not be missing (see for comparison, for example, the approach from Hargreaves, 2001). Or, within a cultural model for effectiveness, culturally anchored values, cultural sensitivity, conflict management, or cultural change resp. instability are vital, but often missing (see the approach from Rooney, 2018). Such examples show that even recent and comprehensive approaches on effectiveness sometimes fail to identify vital variables. Help for efficiently identifying vital variables could come from considering innovative meta-theoretical frameworks as heuristics like "living systems" (e.g., Pavé, 2012) or "simple and complex systems" (e.g., Proctor & Van Zandt, 2018).

The challenge of coupled mechanisms of change. Often, educational intervention research cannot handle successfully complex problems which are situated on different levels of reality like educational systems, schools, classrooms, and individuals. Interventions then have to be "multi-level" interventions (e.g., Erbacher, Singer, & Poland, 2015) in which problem-solving activities are coordinated and implemented on different levels simultaneously. Such interventions are, for example, also related to statistical multi-level analysis (e.g., Humphrey & LeBreton, 2019). Within such interventions, the outputs of a higher level (e.g., teaching quality in classrooms) often represent the antecedents on a lower level (e.g., individual learning processes) (Scheerens, 2015). From a theoretical perspective, the problem is to find links or coupled mechanisms of change which can conclusively connect different levels. There are, for example, effectiveness models on the school context (e.g., Fleener, 2016), on the school level (e.g., Sammons, 1995), on the classroom level (e.g., Astleitner, 2018a), or on the individual level (e.g., Mayer, 1999). Theoretical concepts which might deliver such links concern, for example, leadership, climate, or culture. However, most of these concepts do not distinguish between different levels. Only few approaches in educational effectiveness or intervention research can handle conclusively different levels. For example, within the path-goal theory, there is leadership behavior, also subordinate behavior, and task characteristics which allow to establish coupled mechanisms of change between different levels (e.g., Phillips & Phillips, 2016).

The challenge of confounding methodological standards with standards on theoretical excellence. In general, there are areas in educational intervention research which could have a stronger theoretical foundation. For example, recently, it was criticized that especially school effectiveness and also educational intervention research suffer from theoretical shortcomings (e.g., Anderson & Shattuck, 2012; Hanberger, 2014). However, there are also existing and well-established theories like the dynamic model from Creemers and Kyriakides (2010) which was tested successfully with promising results (e.g., Reynolds et al., 2014). The problem here is that such and similar approaches were not evaluated based on theoretical standards. Such standards exist, but are not in the focus of intervention research. For example, Maag Merki, Emmerich, and Holmeier (2015) listed core elements of school-effectiveness theories like multi-level-structure, dynamic perspective, linear and non-linear resp. direct and indirect effects, differential effects, longitudinal perspective, and multidi-

mensional output criteria. These authors also suggested strategies to advance educational effectiveness theories, like, for example, to include alternative models, extend methodology to analyze processes and mechanisms, analyze differential processes and instruments, carry out complex multivariate analyses, or combine theory and practice in real school situations. However, most of these core elements and strategies for theories are about social research methods or methodological principles but not on sophisticated types and criteria for theories and theory building procedures like relevance, problem orientation, originality, elegance, or stimulation of research (e.g., Astleitner, 2018b). Of course, first, there should be a theoretical solution to a theoretical problem, and not a methodological one.

Multiple other special challenges for theory building on educational interventions. There are also other specific challenges for theory building on educational effectiveness. First, simple models need to be integrated into more complex models, because complex models are more compatible with complex situations in interventions in educational practice (e.g., Creemers & Kyriakides, 2006). Second, it was criticized that theoretical models and related research were strongly focusing on distant factors in respect to learning. However, a higher distance from individual learning processes reduces the effectiveness of educational interventions (Seidel & Shavelson, 2007). Third, there is little functional creativity in the field in a way that most highly different models of, for example, educational leadership actually include many of the same practices (Leithwood & Sun, 2012). Fourth, overall progress in theory building is missing an encompassing, pro-active perspective together with a consequent search for explanatory mechanisms. There is a clear need to make studies more theory-driven (Scheerens, 2013). Fifth, many educational interventions produced small to medium effects and there is still a significant necessity to find new effective factors within theory building processes (Goldberg et al., 2019).

Need for increasing importance and status of theories. Having all these and similar challenges for theories on educational effectiveness in mind, it is obvious that such theories have potentials for further developmental steps on the importance of theories and their theoretical status (e.g., Embretson & Gorin, 2001; Greenwald, 2012).

The importance of theory concerns a weak up to a strong theoretical focus of a research approach. This focus ranges from (1) considering no theory at all, (2) to have a strong method which is assumed to solve also theoretical problems, (3) the rejection of given theories, (4) the acceptance of given theories, (5) a synthesis of given theories, (6) the development of a new theory, and (7) finally to the assumption that also decisions on methods should be based on theoretical assumptions. The theoretical status is about the degree of evidence. It ranges from (1) naive everyday assumptions, (2) research-based working assumptions, (3) theories in a first version, (4) weakly tested theories, (5) strongly tested theories, (6) (nearly) proven theories, up to (7) calibrated and established theories (e.g., Astleitner, 2018b).

Given the challenges which were identified in a review of literature and considering the importance and status of theoretical approaches in the field of educational effectiveness research, one could have the reasoned assumption that the current

situation in educational intervention research can be described with what DiSessa and Cobb (2004, p. 79) said on the role of theory in design experiments: "Theories ... seem to replace one another, rather than subsume, extend, or complement other theories. Although the state of the art constantly changes, it is often difficult to tell that progress is being made".

Perspectives for Educational Effectiveness Theories

Based on the challenges for theory development in educational effectiveness research, three important research questions with related types of theories can be identified which should stimulate progress in the field of educational intervention research:

1. How can we calibrate support measures? Calibration means improving mechanisms of change, especially in view of the diversity of participants in educational interventions. As an innovative theoretical perspective, so-called "development-support-theories" are proposed. Such theories link developmental levels with adaptive level-based support strategies.
2. How can we improve context sensitivity? This question is based on a general dilemma which lies behind many of the mentioned challenges. In educational intervention research, we must have general theories, but we apply and implement them into specific contexts. General theories do not allow to handle specific practical problems at first sight. In order to reduce this problem, so-called "implementation-theories" are suggested. These theories are about how interventions have to designed in order to increase sensibility to different contexts.
3. What goes wrong and why? This question is especially related to the challenge of ineffectiveness. As an innovative theoretical perspective, it is suggested to develop so-called "dark-side-theories". Such theories are about negative or faulty aspects of educational interventions (see also chapter 2 with a dark-side theory on teaching).

Development-support-theories

Development-support-theories link developmental stages or competence levels with adaptive strategies for establishing attainment-based support mechanisms (see Figure 1). The essential assumption is that within each level (from, for example, A to E), there is a certain percentage of goal attainment. If goal attainment on a level reaches a high percentage, then the probability increases that the next level is reached. For each level, there are level-based support strategies (A- to E-strategies). Such theories combine scientific, objective theories (on developmental stages) with prescriptive strategies on what to do to change the situation. All parts of such a theory should be based on findings from empirical research.

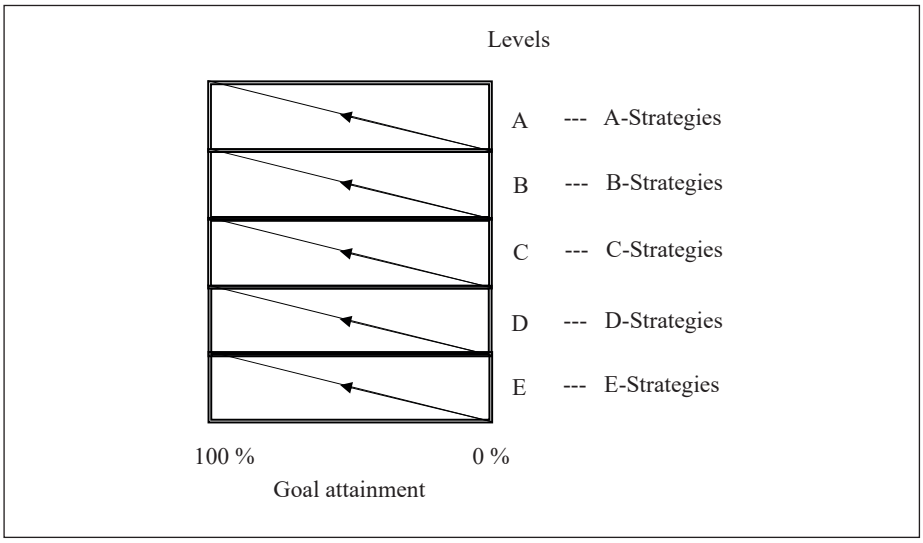

Fig. 1: A hierarchically-organized developmental model.

Within the literature, it is possible to find well-established theories on developmental stages or competence models (for an overview, see Astleitner, 2018a). There are also support strategies for different competence levels like strategies for gifted students (e.g., Wood & Peterson, 2018) or for struggling students (e.g., Jackson & Lambert, 2010). What cannot be found within research on educational effectiveness are models that combine both aspects. Astleitner as one of the few presented such an approach on student engagement (Astleitner, 2018a, see also chapter 5 in this book) and on fostering love as interpersonal competence (e.g., Astleitner & Baumgartner, 2015; see Table 1). For example, the developmental-support theory on fostering love can be described as follows (Astleitner & Baumgartner, 2015, p. 5):

> The developmental model consists of five developmental steps from awareness (i.e., the perception of somebody on the basis of a cognitive construction process), acceptance (i.e., the respect for somebody because of a positive evaluation), care (i.e., the support of somebody's welfare), trust (i.e., the degree a person believes to can count on somebody), and love (i.e., a strong feeling based on intimacy, passion, and commitment; see the triangular theory of love by Sternberg, 1986). For each of these developmental steps, it is assumed that they can be stimulated by certain instructional strategies. Awareness can be affected by knowledge-based interactions (e.g., telling personal histories), acquiring emotional intelligence (e.g., expanding empathic behavior), or allowing positive bias (e.g., seeing the partner in a positive light). Expressing high meaning (e.g., by rewarding), searching for similarities or complementarities (e.g., by finding common goals in life), and promoting tolerance (e.g., by showing the interdependences of problems) should affect acceptance. Care can be realized by expanding others (e.g., putting aside self-interest), achieving compassionate goals (e.g., having a "boy scout"-perspective in life), and

doing perspective taking (e.g., by role-play activities). The instructional strategies for trust concern being positive and open (e.g., by conducting "self-science"), negotiating identities (e.g., by finding solutions without harming others), or keeping balance (e.g., by coordinating personality developmental activities). Finally, realizing togetherness, passionate emotions, and defending might affect love (e.g., by increasing time spent together). It is also important to communicate love (e.g., by giving compliments) and to maintain novelty (e.g., by undertaking new intellectual or physical activities).

Why do we need such development-support-theories in educational intervention research? There is a need for such theories, because researchers in the field of educational interventions have to consider the high diversity in human beings to a larger extent in their theories. There is evidence that educational institutions are becoming even more diversified (e.g., Theoharis & Scanlan, 2015). When people are highly different, then educational interventions have to be designed in a way that this diversity is focused during the design and implementation process. This idea is not new. Within educational intervention research, there are traditional and well-established methodological approaches on "aptitude-treatment-interaction" (e.g., Astleitner, Kriegseisen, & Riffert, 2009) or on "design-based research" (e.g., Mintrop, 2016) in which differences between people are considered when designing different interventions for different participants. However, these approaches are method-driven and not accompanied by suitable complementary theoretical approaches. General theory-driven approaches in educational intervention contexts concern "differentiation" (i.e., the adaptation of instructional methods on different needs of students; e.g., Cash, 2017), "instructional alignment" (i.e., the adaptation of different teaching goals, instructional methods, and evaluation of learning; e.g., Carter, 2007), or "adaptive" and "personalized" learning environments (e.g., Kinshuk, 2016).

How to get from traditional scientific theories to development-support-theories? In general, Astleitner (2018b, p. 135) showed in detail how to build development-support-theories: The first step is defining a final goal of a development. The next step is to formulate a process with intermediate stages, which are related to this final goal. After the process is formulated in stages, influencing factors (support strategies) are determined which are different for different stages. In a final step, the resulting development model is constantly revised and checked for validity. Social science theories and corresponding empirical evidence come into play in all phases of this theory development process.

Another possibility is to search for theoretical models which are hierarchically organized and then link them to other models which contain matching support strategies. An example for this way of theory building can be shown in the field of motivational interventions. Here, on the one hand, for example, Ryan and Deci (2000) delivered a hierarchically organized taxonomy of human motivation as a developmental model ranging from (low) amotivation, external regulation, introjection, identification, integration to (high) intrinsic motivation. On the other hand,

Tab. 1: A Human Developmental Model on Fostering Love (based on: Astleitner & Baumgartner, 2015)

Developmental Steps	Support Strategies
Awareness	(1) Establishing knowledge-based interactions (2) Acquiring emotional intelligence (3) Allowing positive bias
Acceptance	(4) Expressing high meaning (5) Searching for similarities and complementarities (6) Promoting tolerance
Care	(7) Expanding others (8) Achieving compassionate goals (9) Doing perspective taking
Trust	(10) Being positive and open (11) Negotiating identities (12) Keeping balance
Love	(13) Realizing togetherness, passionate emotions, and defending (14) Communicating love (15) Maintaining novelty

Keller (2010) provided with his ARCS-model many support strategies for stimulating motivational parameters like attention, relevance, confidence, and satisfaction. Now, the building of a development-support-theory can be realized when the stages of the developmental model are linked with certain support strategies. For example, it might be theoretically and empirically conclusive to link the support strategy of "extrinsic rewards" (from the ARCS-model) with the developmental level of "external regulation" (from the taxonomy of human motivation), or "intrinsic reinforcement" with "intrinsic motivation", or "perceptual arousal" with "amotivation". Overall, both models could profit from integrating them: The taxonomy of human motivation from Ryan and Deci (2000) develops from a descriptive psychological theory to a prescriptive educational theory. The ARCS-model from Keller (2010) could advance into an approach which can be used for individualized and competence-based instruction.

Implementation-theories

Implementation-theories are about how interventions have to be designed for optimizing sensitivity to different contexts. Sensitivity concerns important elements of an educational intervention like active ingredients (e.g., learning materials), dosage (e.g., strength, duration), passive ingredients (e.g., perceptual, communication, or acceptance design), (theoretical) sampling, or side effects (Astleitner, 2013a; Lipsey, 1990). Sensitivity in intervention research was considered from a methodological perspective especially in implementation science (Kelly & Perkins,

2012), design-based research (Bakker, 2019), and concerning response to intervention approaches (Jimerson, Burns, & VanDerHeyden, 2016). Such methodological perspectives concern a) the likelihood that an effect of an intervention will be detectable by research design, b) evidence-based strategies (e.g., data-based decisions) to enhance the effectiveness of interventions, c) iterative cycles of testing and improving interventions, or d) designing measurements, trainings, and evaluations in order to optimize intervention effects in real-world-settings.

Why do we need implementation-theories? Within Figure 2, it is illustrated how scientific theories and implementation-theories are interlinked. Traditionally, intervention researchers think that their independent variables (as intervention) produce effects on their dependent variables. The relationship from independent to dependent variables and related mechanisms or processes are described and explained by scientific theories. However, this is only half of the truth: It is not the independent variable which produces an effect on the dependent variable, but the intervention in context. The intervention is strongly influenced by context characteristics. The relationship between independent variables, context characteristics, and the intervention are in the focus of an implementation-theory. So, an educational intervention is not only a test of the scientific theory, but also simultaneously of the implementation-theory. An effect of an intervention can be positive, because of a good scientific theory. An effect of the same intervention can be negative, because of a bad implementation-theory. For example, humor proved to have a positive effect on learning in teaching contexts. However, when somebody is not able to consider the context (e.g., women or men) of being humorous, then positive effects diminished (e.g., Wanzer, Frymier, & Irwin, 2010). Another example is given in Figure 2: If somebody wants to improve learning by increasing motivation, then as intervention the design of a motivating interaction (between students and teachers) is necessary. This motivating interaction is different, for example, for a fear of failure-context in comparison to a success-orientated context. In both cases, different implementation-theories might be used like approaches on the motivation crowding effect (e.g., Frey & Jegen, 2001), the success-failure-ratio (Gottman, Coan, Carrere, & Swanson, 1998), or protection motivation (e.g., Floyd, Prentice-Dunn, & Rogers, 2000). In this case, for the same independent variables, different types of interventions will be designed and effective.

A theory on the dosage of an intervention. Another example of an implementation-theory can be based on the assumption that the same dosage (e.g., task difficulty) of an intervention leads to different effects in different contexts (e.g., groups of learners) (e.g., Astleitner, 2008). The effect of the dosage depends on the relationship between dosage and effect (Lipsey, 1990, p. 143). This dosage-effect-relationship can be classified into four theoretical types:

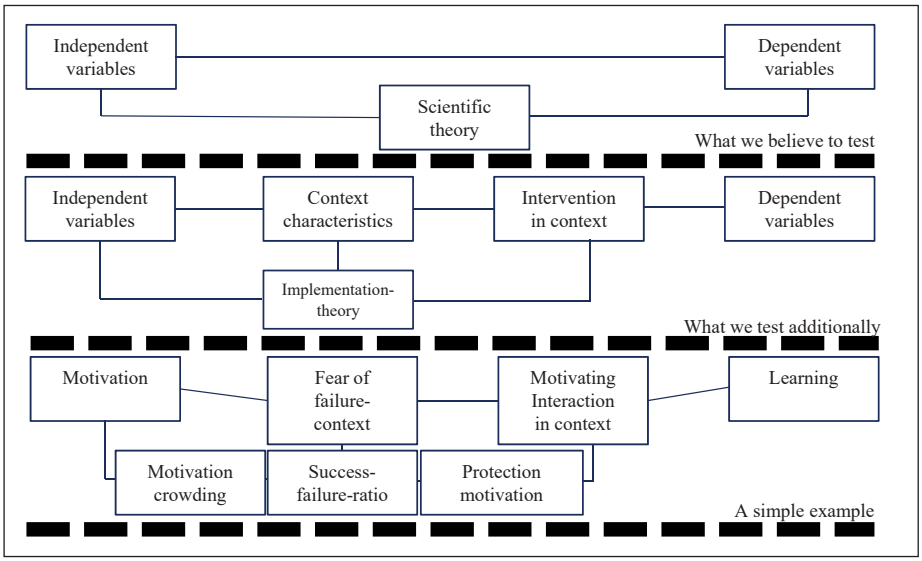

Fig. 2: Scientific theory and implementation-theory.

a) A linear relationship (groups with increasing input needed): This means that an increase (decrease) of the dosage of an intervention will lead to an increase (decrease) in the effect of this intervention. For the group context, this assumption means that interventionists have to invest additional resources (in time, etc.) in order to get better results (e.g., achievements).
b) A step function (the saturated groups): Here, an increase of the dosage of an intervention does not produce any increase in effects for a long time. However, when a certain level of dosage is reached, then there are strong effects. For example, it might be possible that an increase in teacher training will not change the quality of classroom instruction for a long time. However, it might be the case, that such efforts would lead after a certain amount of time to, for example, additional self-organized and highly motivated group building processes which could improve the quality of classroom instruction significantly.
c) A strengthened or weakened relationships (the all-is-possible groups): In this situation, the same dosage has sometimes a strengthened and sometimes a weakened effect. The effect depends not only on the dosage, but also on other factors which often cannot be controlled in educational interventions. In group contexts, the same well-designed teacher education program can, for example, be a large success in one school, and at the same time, be a disaster in another school.
d) A U- or inverted U-relationship (the not too much and not too little groups): Here, a maximum or minimum effect is given at an average dosage of an intervention. For example, in school contexts, it might not be a good advice to change the classroom behaviors of teachers to a large extent when there are problems. It might be more effective, to change some elements, but also keep others. For

example, Chow, Davids, Button, and Renshaw (2015) described learning as a complex interacting system in which the individual practice task and a finely tuned dosage of intervention are essential. Also, McNaughton (2018) compared situations in which there is too much support for students with situations in which there is too little support.

Knowledge about such different types of dosage-effect-relationships is essential when designing interventions, because it allows to optimize intervention effects and to handle resources (in time, money, etc.) more efficiently. When, for example, teachers prefer an average dosage of intervention, then strong interventions should be avoided. In such cases, more in intervention will lead to less effect.

Dark-side-theories

Dark-side-theories describe and explain the negative, non-transparent, or faulty facets of educational interventions. Dark-side-theories are related to general social science approaches like the iceberg model of organizational cultures (Sackmann, 1991), or parallel worlds (Astleitner, 2013b). An explicit dark-side perspective in intervention research can be found, for example, in school reform (e.g., Brooks, 2005), educational leadership (e.g., Polka & Litchka, 2008), transformational leadership (e.g., Tourish, 2013), or organizational behavior (e.g., Griffin & O'Leary-Kelly, 2004).

Why do we need dark-side-theories? The dark-side of educational intervention research was not in foreground of discussions due to some tendencies in current research like the fading out of "negative evidence" and resulting "publication bias" (e.g., Sala & Gobet, 2017). Dark-side-theories could help to find new effective factors in educational intervention research. Recent meta-analyses on the effectiveness of educational interventions often only found small to moderate effects. For example, Scammacca, Roberts, Vaughn, and Stuebing (2015) found a small mean effect size for standard measures of reading of 0.21 although reading in schools represented one of the most important issues in educational intervention research for the last decades.

In addition, there is some specific more or less anecdotal evidence on the dark side in educational fields which could be in the focus of educational intervention research and related theory building. There are, for example, "hidden curriculums" in schools and classrooms (Gordon, Bridglall, & Meroe, 2005), confidential reports on problems in educational practice (Turner, 2017), confessions of ineffective teachers (Owens, 2013), lists of educational errors (Bebell, 2013), documentations of implicit bias in schools (Gullo, Capatosto, & Staats, 2019), or discussions about lies of school reform (Gorski & Zenkov, 2014). However, such soft evidence does not allow to have a profound basis for handling the dark-side of educational interventions, and it does not help to improve the educational profession. In other professional areas like medicine or psychotherapy, there are comprehensive theoretical and empirical approaches on cognitive errors, diagnostic mistakes, or on learning from failures

(Dryden & Neenan, 2011; Howard, 2018). In order that educational intervention research can achieve what these comparable fields of research already accomplished, more dark-side-theories could be helpful.

An example of dark-side-theory on trust development. There is strong evidence that trust represents a core source for interventions in educational contexts (e.g., Bryk & Schneider, 2002). Saunders, Dietz, and Thornhill (2014) found that differing educational interventions are needed to reduce distrust in comparison to build trust. In order to handle such a situation, it would be necessary to have a theory which focuses at the same time on the bright side (trust) and on the dark side (distrust). Exactly such a theory was presented by Lewicki, Tomlinson, and Gillespie (2006) on interpersonal trust development. In particular, within this model of trust development, a dark facet is given, when there is high distrust and low trust, and a bright facet is on high trust and low distrust. From an educational intervention perspective, it could be an important goal to develop people from the dark facet into the bright facet. Such theories which cover at the same time a positive and a negative aspect could be an important starting point for developing dark-side-theories. Such combinations of bright and dark facets concern cognitive processes (e.g., effective and ineffective problem solving), motivational processes (e.g., hope for success and fear of failure) as well as emotional processes (e.g., love and hate).

A Progressive Research Program for Educational Intervention Research

Development-support-, implementation-, and dark-side-theories represent more or less new types of theories in educational intervention research. In order to stimulate research on such theories, a research program must be undertaken in the field educational intervention research. Such a research program should focus on a more intensified consideration of theory building methods and on research strategies which were outlined in this chapter. It consists of:

- concerning development-support-theories
  - the generation of hierarchically organized theories on human development together with a set of educational support strategies,
  - the use of research methods which test the interaction of individual development and support strategies (based on aptitude-treatment-interaction- or complex trait-treatment-interaction studies (e.g., Leutner & Rammsayer, 1995)),
- concerning implementation-theories
  - the exploration of unintended side effects, because they are not on the research agenda of educational intervention research,
  - the optimization of passive ingredients of an intervention, because they are often the reason why well-designed intervention approaches fail, and
- concerning dark-side-theories

- the integration of existing models on bright and dark facets of a phenomenon,
- the use of insider research which allows to explore hidden problems,
- the learning from dark case studies which address negative evidence, and
- the deep analysis of failures or errors in basic research, applied research as well as practice.

Such a comprehensive and sophisticated research program needs time and innovative research strategies for implementation. In order to save resources, a concept can be suggested which could allow to integrate multiple mentioned research strategies simultaneously. Such a strategy concerns amateur researchers and their theories.

Educational Amateur Theories in Intervention Research

Amateur researchers are people which do some kind of research activities in their everyday life. Such an idea is not new. There are concepts like "action research", "field research", "reflective practitioners", or "teachers as researchers" which focus on research done by amateurs outside traditional research institutions. Amateur researchers can, for example, be found in the areas of genealogy, paranormal phenomena, astrology, biology, archeology, volcanology, education in schools, or criminology. From a theoretical perspective, amateur researchers have, especially at the beginning of their activities, "naive theories", "implicit theories", "subjective theories", "practical theories", or "personal theories" which are different from scientific theories as they concern individual everyday assumptions about given phenomena (e.g., Astleitner & Baumgartner, 2015). Based on this background and as an integrative attempt, "educational amateur theories" are all kinds of non-expert assumptions about the characteristics and the relationships of everyday phenomena in the field of education, instruction, training, coaching, or counseling.

An important approach in educational intervention research could be to transform such amateur theories into scientific theories in individuals. An educational intervention would then be a theory transformation process in which a more sophisticated theory is acquired. A sophisticated theory should be true and allow to solve more effectively problems in educational practice. For example, Blackwell, Trzesniewski, and Dweck (2007) used the teaching of theory as an intervention to change implicit theories on intelligence. Or, Tolma, Stoner, Li, Kim, and Engelman (2014) presented an expanded model of the theory of planned behavior in which the following possible elements of amateur theories were included:

- beliefs about consequences of performing the behavior,
- beliefs about significant others' expectations with respect to behavior,
- motivation to comply with significant others' expectations,
- beliefs about anticipated difficulties of performing behavior,
- attitudes,

- subjective norms,
- perceived behavioral control,
- cultural norms,
- perceived susceptibility,
- social modeling,
- intention,
- self-efficacy, or
- fatalism.

The goal of an educational intervention then would be to have a coordinated and equal impact on all of these elements of amateur theories.

References

Addison, R., & Brundrett, M. (2008). Motivation and demotivation of teachers in primary schools: The challenge of change. *Education 3–13: International Journal of Primary, Elementary, and Early Years Education, 36*, 79–94. doi: https://doi.org/10.1080/03004270701733254

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher*, *41*, 16–25. doi: https://doi.org/10.3102/0013189X11428813

Astleitner, H. (2008). Die lernrelevante Ordnung von Aufgaben nach der Aufgabenschwierigkeit [The learning relevant organization of tasks in relation to task difficulty]. In J. Thonhauser (Ed.), *Aufgaben als Katalysatoren von Lernprozessen* (pp. 65–80). Münster: Waxmann.

Astleitner, H. (2010). Methodische Rahmenbedingungen zur Entdeckung der Wirksamkeit von pädagogischen Interventionen [Methodological conditions for identifying the impact of educational interventions]. In T. Hascher & B. Schmitz (Eds.), *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* (pp. 48–62). Weinheim: Juventa.

Astleitner, H. (2013a, September). *A systematic approach on the theoretical quality of educational intervention research: The Interventions Theory Questions (ITQ).* Paper presented at ECER (European Conference on Educational Research), Istanbul. Retrieved from https://www.uni-salzburg.at/fileadmin/multimedia/Erziehungswissenschaft/treatment-validity5.pdf

Astleitner, H. (2013b). *Das Parallelwelt-Phänomen. Sozialwissenschaftliche Grundlagen und Methoden kritischen Denkens* [The parallel world phenomenon. Social science fundamentals and methods of critical thinking]. Münster: Waxmann.

Astleitner, H. (2018a). Multidimensional engagement in learning – An integrated instructional design approach. *Journal of Instructional Research*, *7*, 6–32. doi: https://doi.org/10.9743/JIR.2018.1

Astleitner, H. (2018b). *Spezielle Verfahren sozialwissenschaftlicher Theorieentwicklung* [Special methods of theory building in social research]. Weinheim: Beltz Juventa.

Astleitner, H. (2019, March). *Theories on school effectiveness and ineffectiveness – The next steps*. Powerpoint presentation at the Conference on International Perspectives on School Quality and Teacher Education. Salzburg.

Astleitner, H., & Baumgartner, T. (2015, August). *Transforming identity – The elements and the stability of implicit theories about how love can be developed*. Paper presented at the 8th SELF Biennial International Conference, Kiel. Retrieved from http://www.uni-salzburg.at/fileadmin/multimedia/Erziehungswissenschaft/paperAS_BA6.pdf

Astleitner, H., Kriegseisen, J., & Riffert, F. (2009, September). *Using a multiple evidence model (MUEMO) for testing the effectiveness of educational interventions*. Paper presented at the European Conference of Educational Research (ECER), Vienna. Retrieved from http://seniorenuniversitaet.at/fileadmin/oracle_file_imports/1091171.PDF

Bakker, A. (2019). *Design research in education*. Abingdon, New York: Routledge.

Barger, M. M., & Linnenbrink-Garcia, L. (2017). Developmental systems of students' personal theories about education. *Educational Psychologist*, *52*, 63–83. doi: https://doi.org/10.1080/00461520.2016.1252264

Bebell, C. (2013). *How to fix schools: Educational errors that hurt students, teachers, and schools*. Bloomington: iUniverse.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*, 246–263. doi: https://doi.org/10.1111/j.1467-8624.2007.00995.x

Brooks, J. S. (2005). *The dark side of school reform*. Lanham: Rowman & Littlefield Education.

Bryk, A. S., & Schneider, B. (2002). *Trust in schools. A core resource for improvement*. New York: Russell Sage Foundation.

Carter, L. (2007). *Total instructional alignment. From standards to student success*. Bloomington: Solution Tree.

Cash, R. M. (2017). *Advancing differentiation. Thinking and learning for the 21st century* (2nd ed.). Minneapolis: Free Spirit Publishing.

Chow, J. Y., Davids, K., Button, C., & Renshaw, I. (2015). *Nonlinear pedagogy in skill acquisition*. New York: Taylor & Francis. doi: https://doi.org/10.4324/9781315813042

Creemers, B. P., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, *17*, 347–366. doi: https://doi.org/10.1080/09243450600697242

Creemers, B. P. M., & Kyriakides, L. (2010). Using the Dynamic Model to develop an evidence-based and theory-driven approach to school improvement. *Irish Educational Studies, 29,* 5–23. doi: https://doi.org/10.1080/03323310903522669

Creemers, B. P. M., & Kyriakides, L. (2012). *Improving quality in education. Dynamic approaches to school improvement*. Abingdon, New York: Routledge. doi: https://doi.org/10.4324/9780203817537

DiSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *Journal of the Learning Sciences, 13*, 77–103. doi: https://doi.org/10.1207/s15327809jls1301_4

Dollarhide, C. T., & Lemberger-Truelove, M. E. (Eds.). (2019). *Theories of school counseling for the 21st century*. New York: Oxford University Press.

Dryden, W., & Neenan, M. (2011). *Learning from mistakes in rational emotive behaviour therapy*. Hove, New York: Routledge. doi: https://doi.org/10.4324/9780203553312

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368. doi: https://doi.org/10.1111/j.1745-3984.2001.tb01131.x

Erbacher, T. A., Singer, J. B., & Poland, S. (2015). *Suicide in schools. A practicioner's guide to multi-level prevention, assessment, intervention, and postvention*. New York, Hove: Routledge. doi: https://doi.org/10.4324/9780203702970

Fleener, J. (2016). Addressing educations' most intractable problems. *Emergence: Complexity and Organization, 18*, 1–12. Retrieved from https://journal.emergentpublications.com/article/addressing-educations-most-intractable-problems/

Floyd, D. L., Prentice-Dunn, S., & Rogers, R. W. (2000). A meta-analysis of research on protection motivation theory. *Journal of Applied Social Psychology*, *30*, 407–429. doi: https://doi.org/10.1111/j.1559-1816.2000.tb02323.x

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*, 589–611. doi: https://doi.org/10.1111/1467-6419.00150

Funnell, S. C., & Rogers, P. J. (2011). *Purposeful program theory. Effective use of theories of change and logic models*. San Francisco: Jossey-Bass.

Furlong, M. J., Gilman, R., & Huebner, E. S. (Eds.). (2014). *Handbook of positive psychology in schools* (2nd ed.). New York, Abingdon: Routledge. doi: https://doi.org/10.4324/9780203106525

Goldberg, J. M., Sklad, M., Elfrink, T. R., Schreurs, K. M., Bohlmeijer, E. T., & Clarke, A. M. (2019). Effectiveness of interventions adopting a whole school approach to enhancing social and emotional development: A meta-analysis. *European Journal of Psychology of Education, 34*, 755–782. doi: https://doi.org/10.1007/s10212-018-0406-9

Gordon, E. W., Bridglall, B. L., & Meroe, A. S. (Eds.). (2005). *Supplementary education: The hidden curriculum of high academic achievement*. Lanham: Rowman & Littlefield Education.

Gorski, P. C., & Zenkov, K. (Eds.). (2014). *The big lies in school reform. Finding better solutions for the future of public education*. New York, Abingdon: Routledge. doi: https://doi.org/10.4324/9781315886367

Gottman, J. M., Coan, J., Carrere, S., & Swanson, C. (1998). Predicting marital happiness and stability from newlywed interactions. *Journal of Marriage and the Family*, 60, 5–22. doi: https://doi.org/10.2307/353438

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*, 99–108. doi: https://doi.org/10.1177/1745691611434210

Griffin, R. W., & O'Leary-Kelly, A. (Eds.). (2004). *The dark side of organizational behavior*. San Francisco: Jossey-Bass.

Gullo, G. L., Capatosto, K., & Staats, C. (2019). *Implicit bias in schools. A practitioner's guide*. New York, Abingdon: Routledge. doi: https://doi.org/10.4324/9781351019903

Hanberger, A. (2014). What PISA intends to and can possibly achieve: A critical programme theory analysis. *European Educational Research Journal*, *13*, 167–180. doi: https://doi.org/10.2304/eerj.2014.13.2.167

Hargreaves, D. H. (2001). A capital theory of school effectiveness and improvement. *British Educational Research Journal*, *27*, 487–503. doi: https://doi.org/10.1080/01411920120071489

Howard, J. (2018). *Cognitive errors and diagnostic mistakes: A case-based guide to critical thinking in medicine*. Cham: Springer. doi: https://doi.org/10.1007/978-3-319-93224-8

Humphrey, S. E., & LeBreton, J. A. (Eds.). (2019). *The handbook of multilevel theory, measurement, and analysis*. Washington: American Psychological Association. doi: https://doi.org/10.1037/0000115-000

Irby, B. J., Brown, G., Lara-Alecio, R., & S. Jackson (Eds.). (2013). *The handbook of educational theories*. Charlotte: Information Age Publishing.

Jaccard, J., & Jacoby, J. (2010). *Theory construction and model-building skills*. New York, London: Guilford.

Jackson, R. R., & Lambert, C. (2010). *How to support struggling students*. Alexandria: ASCD.

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (Eds.). (2016). *Handbook of response to intervention. The science and practice of multi-tiered systems of support* (2nd ed.). New York: Springer. doi: https://doi.org/10.1007/978-1-4899-7568-3

Keller, J. M. (2010). *Motivational design for learning and performance. The ARCS model approach*. New York: Springer. doi: https://doi.org/10.1007/978-1-4419-1250-3

Kelly, B., & Perkins, D. F. (Eds.). (2012). *Handbook of implementation science for psychology in education*. New York: Cambridge University Press. doi: https://doi.org/10.1017/CBO9781139013949

Kinshuk (2016). *Designing adaptive and personalized learning environments*. New York, Abingdon: Routledge. doi: https://doi.org/10.4324/9781315795492

Leithwood, K., & Sun, J. (2012). The nature and effects of transformational school leadership: A meta-analytic review of unpublished research. *Educational Administration Quarterly*, *48*, 387–423. doi: https://doi.org/10.1177/0013161X11436268

Leutner, D., & Rammsayer, T. (1995). Complex trait-treatment-interaction analysis: A powerful approach for analysing individual differences in experimental designs. *Personality and Individual Differences*, *19*, 493–511. doi: https://doi.org/10.1016/0191-8869(95)00062-B

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management*, *32*, 991–1022. doi: https://doi.org/10.1177/0149206306294405

Lipsey, M. R. (1990). *Design sensitivity. Statistical power for experimental research*. Newbury Park: Sage.

Lynch, M. (2012). *A guide to effective school leadership theories*. New York: Routledge. doi: https://doi.org/10.4324/9780203181010

Maag Merki, K., Emmerich, M., & Holmeier, M. (2015). Further development of educational effectiveness theory in a multilevel context: From theory to methodology and from empirical evidence back to theory. *School Effectiveness and School Improvement, 26,* 4–9. doi: https://doi.org/10.1080/09243453.2014.938930

Maag Merki, K., Emmerich, M., & Holmeier, M. (Eds.). (2017). *Educational effectiveness theory. Further developments in a multilevel context*. Abingdon: Routledge. doi: https://doi.org/10.4324/9781315231037

Manzano Vázquez, B. (2018). Teacher development for autonomy: An exploratory review of language teacher education for learner and teacher autonomy. *Innovation in Language Learning and Teaching*, *12*, 387–398. doi: https://doi.org/10.1080/17501229.2016.1235171

Mayer, R. E. (1999). Designing instruction for constructivist learning. In C. M. Reigeluth (Ed.), *Instructional-design theories and models* (Vol. II, pp. 141–159). Mahwah: Erlbaum.

McNaughton, S. (2018). *Instructional risk in instruction. Why instruction can fail*. Abingdon, New York: Routledge. doi: https://doi.org/10.4324/9781351129206

Mintrop, R. (2016). *Design-based school improvement. A practical guide for education leaders*. Cambridge: Harvard Education Press.

Olson, M. H., & Hergenhahn, B. R. (2016). *An introduction to theories of learning* (9th ed.). New York, Hove: Routledge. doi: https://doi.org/10.4324/9781315664965

Owens, J. (2013). *Confessions of a bad teacher*. Naperville: Sourcebooks.

Pavé, A. (2012). *Modeling living systems. From cell to ecosystem*. London: ISTE Ltd. doi: https://doi.org/10.1002/9781118569634

Phillips, A. S., & Phillips, C. R. (2016). Behavioral styles of path-goal theory: An exercise for developing leadership skills. *Management Teaching Review*, *1*, 148–154. doi: https://doi.org/10.1177/2379298116639725

Polka, W. S., & Litchka, P. R. (2008). *The dark side of educational leadership*. Lanham: Rowman & Littlefield Education.

Proctor, R. W., & Van Zandt, T. (2018). *Human factors in simple and complex systems* (3rd ed.). Boca Raton: CRC Press.

Reigeluth, C. M., Beatty, B. J., & Myers, R. D. (Eds.). (2017). *Instructional-design theories and models* (Vol. IV). New York, Abingdon: Routledge. https://doi.org/10.4324/9780203056783-1

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, *25*, 197–230. doi: https://doi.org/10.1080/09243453.2014.885450

Rooney, P. K. (2018). A cultural assets model for school effectiveness. *Cambridge Journal of Education*, *48*, 445–459. doi: https://doi.org/10.1080/0305764X.2017.1356266

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*, 54–67. doi: https://doi.org/10.1006/ceps.1999.1020

Sackmann, S. A. (1991). Uncovering culture in organizations. *The Journal of Applied Behavioral Science*, *27*, 295–317. doi: https://doi.org/10.1177/0021886391273005

Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science*, *26*, 515–520. doi: https://doi.org/10.1177/0963721417712760

Sammons, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. Ringwood: B & MBC Distribution Services.

Saunders, M. N., Dietz, G., & Thornhill, A. (2014). Trust and distrust: Polar opposites, or independent but co-existing? *Human Relations*, *67*, 639–665. doi: https://doi.org/10.1177/0018726713500831

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities*, *48*, 369–390. doi: https://doi.org/10.1177/0022219413504995

Scheerens, J. (2013). The use of theory in school effectiveness research revisited. *School Effectiveness and School Improvement*, *24*, 1–38. doi: https://doi.org/10.1080/09243453.2012.691100

Scheerens, J. (2015). Theories on educational effectiveness and ineffectiveness. *School Effectiveness and School Improvement, 26*, 10–31. doi: http://dx.doi.org/10.1080/09243453.2013.858754

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*, 454–499. doi: https://doi.org/10.3102/0034654307310317

Sternberg, R. J. (1986). A triangular theory of love. *Psychological Review, 93,* 119–135. doi: http://dx.doi.org/10.1037/0033-295X.93.2.119

Swanson, R. A., & Chermack, T. J. (2013). *Theory building in applied disciplines*. San Francisco: Berrett-Koehler.

Theoharis, G., & Scanlan, M. (Eds.). (2015). *Leadership for increasingly diverse schools*. New York, Abingdon: Routledge. doi: https://doi.org/10.4324/9781315767574

Tolma, E. L., Stoner, J. A., Li, J., Kim, Y., & Engelman, K. K. (2014). Predictors of regular mammography use among American Indian women in Oklahoma: A cross-sectional study. *BMC Women's Health*, *14*, 101. doi: https://doi.org/10.1186/1472-6874-14-101

Tourish, D. (2013). *The dark side of transformational leadership*. New York: Routledge. doi: https://doi.org/10.4324/9780203558119

Turner, R. (2017). *Classroom confidential. Stories from our public schools*. Seattle: Amazon, CreateSpace Independent Publishing Platform.

Wanzer, M. B., Frymier, A. B., & Irwin, J. (2010). An explanation of the relationship between instructor humor and student learning: Instructional humor processing theory. *Communication Education*, *59*, 1–18. doi: https://doi.org/10.1080/03634520903367238

Wood, S. M., & Peterson, J. S. (Eds.). (2018). *Counseling gifted students*. New York: Springer. https://doi.org/10.1891/9780826136558

2. A Theoretical Perspective on Ineffective Interventions: Malfunctions in Teaching

Hermann Astleitner

Abstract: An important problem in educational intervention research represents the fact that interventions are ineffective. Ineffectiveness means that intended goals are not reached. In educational intervention research, there is a strong tendency to ignore ineffective interventions and related problems. This is also true for teaching. Although, there is ineffective teaching, little is known about different types of ineffectiveness or malfunctions in teaching. In this contribution, there are three main goals related to a dark-side theory. First, it is important to identify a systematic set of teaching malfunctions. Second, it aims at formulating comprehensive macro- and micro-theories about why malfunctions occur. Finally, implications and problems for future educational intervention research are briefly outlined.

Teaching is a highly complex interventionist activity which requires sophisticated professional teacher competences (e.g., Kunter et al., 2013). Successful teaching presupposes to manage goals, contents, instructional methods, learning materials, and assessments more or less simultaneously. In order to establish successful teaching and to support teachers, educational and psychological research produced numerous models and related evidence on effective teaching (e.g., Ko & Sammons, 2013; Kyriakides, Christoforou, & Charalambous, 2013; Schacter & Thum, 2004). Such positive models on teaching were implemented into teacher education and training with significant profits for teachers and their students (e.g., Blömeke, Suhl, & Kaiser, 2011; Carr-Chellman, 2016).

However, all the positive models of good and effective teaching cannot avoid problems and related malfunctions in teaching in our daily classrooms. Such malfunctioning leads to a severe situation for teachers with consequences for their classroom performance and related student learning: They suffer, for example, from reality shock, dropout, stress, burnout, poor recovery, and related health problems (Gluschkoff et al., 2016; Stokking, Leenders, DeJong, & Van Tartwijk, 2003). Handling teaching malfunctions is difficult as teaching is an interventionist activity with uncertain conditions, processes, and outcomes. Teaching is full of complexity, dynamic processes, interactions, situational changes, pedagogical dilemmas, and unsolvable problems (e.g., Lampert, 1985). Therefore, teaching is based on a high probability of ineffectiveness, problems, malfunctions, failures, biases, or errors.

A similar situation might also be given for psychotherapy or medical treatment which are comparable to teaching as they represent interventionist activities in order to solve human problems. However, in interventionist disciplines like psychotherapy or medicine, researchers have started to focus in great detail on ineffectiveness resp. malfunctions or errors in their field. For example, 2016, a whole special issue of the journal Psychotherapy has been focusing on "clinical errors" (Budge, 2016). Also, for example, Howard (2019) has classified a tremendous amount of diagnostic, treatment, preventive, or other types of "medical errors" which are essential for the health of patients.

In past and current research on teaching, such a focus on ineffectiveness as well as problems and malfunctions has not played a significant role. There are several reasons for this situation.

First, it is difficult to find a proper definition of "teaching malfunctions" and synonyms like "ineffective teaching", "teaching problems", "teaching errors", "teaching mistakes", "teaching failures", or "bad teaching". According to the Merriam-Webster Dictionary, "malfunction" is "a failure to operate or function in the normal or correct manner" and an "error" is "an unintentional departure from truth or accuracy", "a breaking of a moral or legal code", or "a false idea or belief" (retrieved from https://www.merriam-webster.com). According to another definition, an "error" is a "failure to carry out a task in the way intended by the person performing it, in the way expected by other people or in a way that achieves the desired objective" (Kletz, 2001, p. V). Considering such definitions, "teaching malfunctions" can be defined as instructional activities which are not supportive or even obstructing in reaching desired educational goals and standards in the classroom. One might argue, that having no well-established definition of teaching malfunctions is not accidental, because there is no need to have a definition of teaching malfunctions and to focus on the dark side of teaching, because it is sufficient to concentrate on the positive side. The dark side is only the opposite, or a low level as well as an inefficient characteristic of the positive side. Here, for example, Jackson (2006) discovered that graduates reflecting on their education describe good teaching and bad teaching in significantly different registers without almost no overlap in the vocabulary with which they describe the two sides of teaching. Also, Raufelder et al. (2016) found that the qualities of bad teachers were not always opposed to those of good teachers. A particular focus on teaching malfunctions seems also to be necessary, because there is a long tradition on special research on human errors in many different fields which were widely overseen in research on teaching (e.g., Strauch, 2018). In addition, there is a more or less new trend in education to focus on the difficult to handle negative "dark side" (see also chapter 1 in this book). For example, Bengtsen and Barnett (2017) used the term "dark" to "comprehend challenges, situations, reactions, aims and goals, which cannot easily be understood and solved by agendas of quality assurance and professionalisation". There is also more or less anecdotal and subjective evidence about "bad teachers" which has to be confronted with more objective scientific methods and results (e.g., Owens, 2013). Finally, there

is the phenomenon of "reporting bias" or "publication bias" in research on teaching which leaded to the fading out of non-significant and negative results (e.g., Dawson & Dawson, 2018). Here, a focus on the negative side of teaching could compensate such bias and allow to gain a more complete and therefore more valid picture of the reality of teaching in our schools.

A second line of arguing for the consideration of teaching malfunctions corresponds with a long tradition in the field of education assuming that people and also teachers can learn from errors and mistakes (e.g., Bryant, 2003) or from unsolved problems in problem-based learning scenarios (e.g., Wedel, Müller, Pfetsch, & Ittel, 2019). For example, Treiber (1984) demonstrated to connect specific behavior patterns showing ineffective teaching with research findings in order to improve teacher education. In a survey from Phelps (2000), teachers reported 43 different teaching mistakes indicating that teaching errors and related fallibility fulfill an important role in classrooms and teacher education. Learning from errors can also be identified within research on the relationship between novice and expert teachers (e.g., Walls, Nardi, von Minden, & Hoffman, 2002). For example, Barbetta, Norona, and Bicard (2005) identified about a dozen of common malfunctions in classroom behavior management. Based on these malfunctions, they have suggested strategies how to improve the situation resp. how to avoid related problems (e.g., "having clear expectations that are enforced and reinforced consistently"). Keith and Frese (2008) did a meta-analysis on the effectiveness of error management trainings (what includes to make errors during training and learn from them). They have found positive and significant effects on performance (d = 0.44). Le Maistre and Paré (2010) found that beginning teachers in daily classrooms have to develop survival strategies (e.g., "satisficing" as having not an optimal but a suffice solution to a problem). Such strategies are erroneous at the beginning, but allow to find better realistic and individual solutions in a step-by-step procedure. A review on learning from errors by Metcalfe (2017) was not related to teaching errors, but delivered profound arguments why also teachers could learn from teaching malfunctions. In analogy, teaching malfunctions can, for example, serve as signposts to correct teaching behaviors, stimulate the remembering of risky contexts, help as dysfunctional response to reduce fear, make it necessary to accommodate to unexpected outcomes, or offset overconfidence. Negative emotional effects of focusing on teaching malfunctions in, for example, teacher education courses can be buffered with emotionally sensitive error management trainings.

Overall, it seems appropriate to say that a focus on teaching malfunctions could a) deliver an expanded basis on how to constitute, measure, and evaluate phenomena which reduce effectiveness in complex and dynamic scenarios like teaching, b) help in getting hidden or unknown knowledge on why things in our classrooms do not work as intended, c) allow to estimate whether successes and failures in teaching are polar opposites or co-existing, d) optimize learning from errors in teacher education settings which should also allow to close the beginner-expert-gap in a more

realistic and individualized way, and e) stimulate deep learning experiences within an emotionally-sensitive scenario.

Is There any Integrative Research on Teaching Malfunctions yet?

Considering that teaching malfunctions represent a valuable field of research, one could ask for already given reviews, meta-analysis, or other research integrating studies on this issue.

Veenman (1984) reviewed research on perceived problems of beginning teachers and identified about 24 problems concerning, for example, classroom discipline, motivating students, dealing with individual differences, and so on. This review identified important problems, but there was no link to teaching malfunctions and related processes. Huang (2002) did a narrative review on errors in English language teaching. This review was mainly focusing on errors of students in learning, but not on teaching. Only few and unorganized issues like the assessment, treatment, and grading errors by teachers concern teaching malfunctions. Kirschner, Sweller, and Clark (2006) reviewed evidence from empirical studies on unguided or minimally-guided learning in classrooms and found that it was less effective than guided learning. They covered however only one important teaching malfunction, namely to realize classroom learning without direct and strong instructional guidance. Hiebert, Morris, Berk, and Jansen (2007) presented an integrative framework on how teachers learn from teaching by testing hypotheses about cause-effect relationships between teaching and learning. Testing hypotheses is a trial-and-error-like approach. However, the authors did not elaborate in detail on errors or malfunctions in teaching. Hattie (2008) analyzed hundreds of meta-analyses and classified methods of teaching from efficient to inefficient. For such classifications, it might be concluded that using teaching methods with low *effect sizes* resp. low impact might be considered as a teaching malfunction. However, one must be cautious on such analyses. For example, Hattie (2008) identified humor of having a very low average impact ($d = 0.04$) in student learning. One cannot conclude that using humor represents a teaching malfunction, because effects of using humor differ considerably in classroom situations. In addition, humor had also not only effects on learning, but also on other factors which are closely related to student achievements like emotions of students (e.g., Bieg, Grassinger, & Dresel, 2017). Scheerens (2015) presented a detailed overview about theories on educational effectiveness and ineffectiveness and asked the question what models have to say about educational ineffectiveness. He has concluded that there are many, but often overseen negative aspects in developing schools and related classroom teaching and that these aspects must be considered "by actively countering implementation failures and side effects, by fostering more realistic expectations on effects and effect sizes among practitioners and policy makers, and by considering alternatives levers for improvement" (p. 27).

In general, such and similar findings make it difficult to identify a comprehensive, conclusive, and integrative perspective on teaching malfunctions within recent research activities.

Objectives and Methods

From a more specific perspective, the mentioned research on teaching malfunctions revealed that there are many deficits and shortcomings. A first objective of this contribution is to develop a systematic classification of different types of teaching malfunctions. Second, a macro and a micro model will be conceptualized which can describe and explain why teaching malfunctions occur. There are such approaches for errors in learning, but not for malfunctions in teaching (e.g., Petkova, 2009).

In order to achieve these objectives, multiple research methods were used in combination. For all goals, it was necessary to conduct a review of literature with an exploratory focus. Due to the limited research status, it was not the orientation to realize a comprehensive and exhaustive review as it is usually done in the case of meta-analyses which focus on effectiveness. Rather, it was the goal to explore and integrate current theoretical and empirical research in order to develop basic concepts and methods on teaching malfunctions (Cooper, 1989). Based on these objectives, scholar.google was used to identify relevant research in scientific journals and books. We used the keywords of "malfunctions" (and related synonyms like "errors", "mistakes", "failures", "shortcomings", or "deficits") and "teaching" (and "instruction" or "classroom behavior"). For specific purposes, we combined these keywords with "theory" (or "model"), "assessment" (or "measurement", "questionnaire", or "scale"), and "teacher education" (or "teacher training"). Overall, more than 60 studies were identified and used in the following sections. In order to support the goal of concept and theory building, theory construction methods based on "focusing concepts" and "causal models" from Jaccard and Jacoby (2010) were applied.

A Taxonomy of Teaching Malfunctions

Within Table 1, a taxonomy of teaching malfunctions is depicted based on a comprehensive and integrative review of research findings. It combines different instructional events, types of teaching malfunctions, definitions, and examples.

The given taxonomy is based on principles of "direct instruction" which focuses on the interaction between teachers and students, the framing of learner performance into goals and tasks, and teachers' activities to support learner performance (Magliaro, Lockee, & Burton, 2005). Direct instruction is – based on Gagné (1985) – focusing on "instructional events" (as interventions or conditions for students learning in the classroom) ranging from gaining attention to enhancing retention and transfer of students, however, with a perspective on malfunctions. The concept of instructional events a) allowed to integrate different paradigm of teaching and

Tab. 1: A Taxonomy of Teaching Malfunctions

Bad teaching events	Types	Definitions	Examples and References
Triggering inattentiveness	Bad disturbance handling	There are classroom disturbances by students which make it difficult to start or continue teaching.	Teachers cannot handle disruptive student behavior effectively. Oliver et al. (2011)
	No perceptual arousal	There are no (unexpected) changes in the learning environment concerning visual or auditory parameters.	Teachers do not vary aspects of their voice like pauses, inflections, or phrasing. Schmidt et al. (1998)
	No inquiry arousal	There are no open questions, unsolved problems, or mysteries which need knowledge-seeking behavior.	Teachers do not assist students in doing experiments for their hypotheses. Yoon et al. (2012)
	Monotony in instructional activities	Teachers use the same teaching-learning activities repeatedly or for a long-time within the lesson.	Teachers do always the same things in classroom. Daschmann et al. (2011)
Establishing diffuse goal orientations	Inadequate goals	Goal statements do not align with requirements in the curriculum.	Teachers state goals that cannot be found within the curriculum. McNeill (2009)
	Unclear goals	Goal statements are missing, non-transparent, unspecified, incomplete, or incomprehensible.	Teachers state as a goal of a lesson that students should learn about some social problems of a country. Seidel et al. (2005)
	Unbalanced goals	Goal statements are non-holistic or focus only on single but not multiple goal areas.	Teacher state cognitive achievement goals without considering motivational or social-emotional goals. Maynard et al. (2017)
	Goal incoherence	Learning processes and products are not set in accordance with the lesson goals.	Teachers state goals at the beginning of the lesson and then never again. Seidel et al. (2005)
	Goal slips	Statements or behaviors related to important goals are changed repeatedly.	Teacher often changes the goals of the lesson. Praetorius et al. (2014)

Bad teaching events	Types	Definitions	Examples and References
Ignoring prerequisite learning	No focus on past learning	Summaries of prerequisite contents or evaluations of past work are missing.	Teachers do not give or evaluate homework. Fernández-Alonso et al. (2015)
	No guiding overviews about new contents	Maps, organizers for gaining an impression or a first model on what must be learned are missing.	Teachers do not use cognitive maps. Dhindsa & Anderson (2011)
	No preparing tasks	There are no tasks for students which activate prior knowledge and link it with new contents.	Teachers do not use a prior knowledge activation strategy. Spires & Donley (1998)
Chaotic information management	No assistance in selecting information	There is no highlighting of important, difficult, or questionable information.	Teachers present information as it were all the same. Kercood & Grskovic (2009)
	Disorganization	There is no clear structure on contents, activities, or in time.	Teachers use unstructured texts in teaching. Hebert et al. (2016)
	No support for integration of new knowledge	There are no elaborative questions, illustrations, or worked examples.	Teachers do not use worked examples for supporting knowledge acquisition. Renkl (2011)
	No focus on note-taking	There are no instructions, strategies, or devices which foster the saving of new information effectively.	Teachers do not use strategic note-taking forms. Boyle (2010)

Bad teaching events	Types	Definitions	Examples and References
Not solving learning problems or errors	Ineffective classroom management	There are no rules in the educational system and no strategies that prevent and correct inappropriate student behavior.	Teachers do not establish rules about how to handle problems. Balli (2011)
	No modeling	There are no activities which demonstrate how to perform desired activities or which articulate the reasoning necessary for the activities.	Teachers do not think out loud during problem-solving. Ness (2016)
	No coaching	There are no activities which monitor, analyze, and regulate the learners' knowledge and skill developments.	Teachers do not monitor the progress of students. Van den Bosch et al. (2017)
	No scaffolding	There are no adjustments of task difficulties.	Teachers do not offer tasks with varying difficulties. Allington et al. (2015)
Preventing task-based learning	No task-based learning environment	There is no usage of a wide variety of tasks for teaching and learning.	Teachers and learners spend little time solving tasks. Fisher (2009)
	No task choice	There is no possibility for students to choose tasks on their own preferences.	Teachers do not allow students to choose different tasks for their own. Morgan (2006)
Nonresponse	No constructive feedback	There is no feedback on the learning process which is explicit, goal-evaluative, specific, positive, and directive-preventive.	Teachers give feedback which was not related to a learning goal. Van den Bergh et al. (2013)
	Bad timing in feedback	Feedback on learning is too delayed or too early, or irregular in time.	Teachers do not give immediate feedback for difficult tasks. Shute (2008)

Bad teaching events	Types	Definitions	Examples and References
Assessment biases	Misalignment	There is no or little connection between assessment, curriculum, and instruction.	Teachers do not follow assessment practices recommended in their coursework. Campbell & Evans (2000)
	No validation of assessment and grades	There are no attempts to learn about and improve the criteria-based quality of assessments and resulting grading.	Teachers use many multiple factors when assessing and grading students in an inconclusive way. McMillan et al. (2002)
Inapplicability of knowledge	No application of knowledge and skills in different contexts.	Knowledge and skills are not applied in varying scenarios or with different experiences.	Teachers do not use examples from daily life environments. Paas & Van Merriënboer (1994)
	No suggestions to think about connections or analogies.	There are no activities in which students have to group problems or invent similar problems.	Teachers do not emphasize the common characteristics of different problems. Fuchs et al. (2003)

learning, b) was widely used and found to be effective in educational contexts, and c) was applied in research activities on teachers and their professional development (e.g., Cronjé, 2006; Krull, Oras, & Sisask, 2007; Martin, Klein, & Sullivan, 2004). These instructional events were used to organize the teaching malfunctions.

A first block of teaching malfunctions is about triggering inattentiveness in classrooms (Kofler, Rapport, & Matt Alderson, 2008). Inattentiveness is about off-task behavior and can be related to the teaching malfunctions of bad disturbance handling, no perceptual arousal, no inquiry arousal, or monotony in instructional activities (e.g., Keller, 1987). For example, such a malfunctioning is given when teachers cannot handle disruptive student behavior effectively what makes it difficult to gain attention for learning.

A second block of teaching malfunctions concerns establishing diffuse goal orientations (Kunst, van Woerkom, & Poell, 2018). It is about not informing students about the goals of learning in a sufficient and compulsory way so that goal setting and goal commitment during learning are negatively affected (e.g., Moeller, Theiler, & Wu, 2012). Related teaching malfunctions include inadequate, unclear, or unbalanced goals as well as goal incoherence or goal slips. An example of such a malfunctioning is about teachers who focus on goals for learning which cannot be found within the given curriculum.

A third block of teaching malfunctions is about ignoring prerequisite learning of students (e.g., Lo & Hew, 2017). In this case, teaching malfunctions are given when there are no foci on past learning, no guiding overviews about new contents of learning, or no preparing tasks. For example, when teachers do not give or evaluate homework or similar preparing tasks, then they fail to stimulate prerequisite learning.

A fourth block of teaching malfunctions refers to a disorganized information management of learning materials within the classroom (Diekema & Olson, 2011). It is about interruptions in the flow of information during presenting, distributing, and storing on- and offline materials (e.g., Mayer, 1999). Here, teaching malfunctions are given when students get no assistance in selecting information, when information is disorganized, when support for the integration of new information is missing, or when there is no focus on note-taking. For example, such a teaching malfunction occurs when teachers present information as it were all the same.

A fifth block of teaching malfunctions is on not solving learning problems or errors of students successfully (Tulis, 2013). It means not or not effectively providing learner guidance in case of learning problems (e.g., Jonassen, 1999). Malfunctions concern ineffective classroom management, and the missing of modeling, coaching, or scaffolding behavior. Such malfunctions are, for example, given when teachers do not establish rules (and consequences) about how to handle learning-related or social problems in classrooms.

A sixth block concerns the problem of preventing task-based activities by students what means not effectively eliciting active student performance during teaching. Related malfunctions are given when no task-based learning environments are

established and when there is no more or less autonomous task choice by students (e.g., Van Merriënboer & Kirschner, 2013). An example of such a malfunction is, for example, given when teachers and learners spend no or only little time on solving tasks.

A seventh block of teaching malfunctions refers to nonresponsive behavior, when there are problems in providing feedback to students. Malfunctions are given when feedback is not constructive or bad in timing (e.g., Havnes, Smith, Dysthe, & Ludvigsen, 2012). Given feedback without focusing on important teaching goals represents an example of this malfunction.

An eighth block concerns assessment biases when judging capabilities and performances of students (Atjonen, 2014). Malfunctions are given when there are a misalignment (between assessment, curriculum, and instruction) and no validation attempts on assessments and related grades (e.g., Zhang & Burry-Stock, 2003). Malfunctions in assessments are, for example, given when teachers use assessment practices which are entirely unknown to students.

A final ninth block is related to the inapplicability of knowledge when enhancing retention and transfer are restricted. Teaching malfunctions in class relate to the lack of application of knowledge and skills in different contexts and no stimulation to think about relationships or analogies in content (e.g., Gentner, Loewenstein, & Thompson, 2003). Such a malfunction is, for example, given when teachers do not embed contents in daily life experiences.

Overall, 28 different types of teaching malfunctions were depicted and classified. Of course, this taxonomy is not exhaustive or covering the full range of research. However, the given taxonomy represents a systematic starting point for further research activities. Within a next step of research, it is important to clarify why such malfunctions in teaching occur.

A Macro- and Micro-Theory about Why Malfunctions Occur

Teaching behavior and related malfunctions are affected by multiple factors which can be summarized within a macro model on teaching malfunctions (Figure 1). Such a macro model focuses on observable processes and products on the school-, classroom-, teacher-, and student-level. Such school/classroom-, and teacher-related factors can be identified when combining more or less general approaches on human erroneous behavior and error management (e.g., Dekker, 2014; Dhillon, 2009; Reason, 1990) as well as research on school-related and instructional effectiveness (e.g., Creemers & Kyriakides, 2012; Stronge, Ward, & Grant, 2011).

First, there are challenges in the teaching profession which require a change of goals and behavior from teachers. These challenges result from changes in politics and corresponding visions, which occur at regular intervals in the educational systems. They are also linked to guidelines or instructions for actions from leadership activities by school principals or other decision makers. In many schools, there is a certain culture about how to deal with malfunctions in teaching. Sometimes such

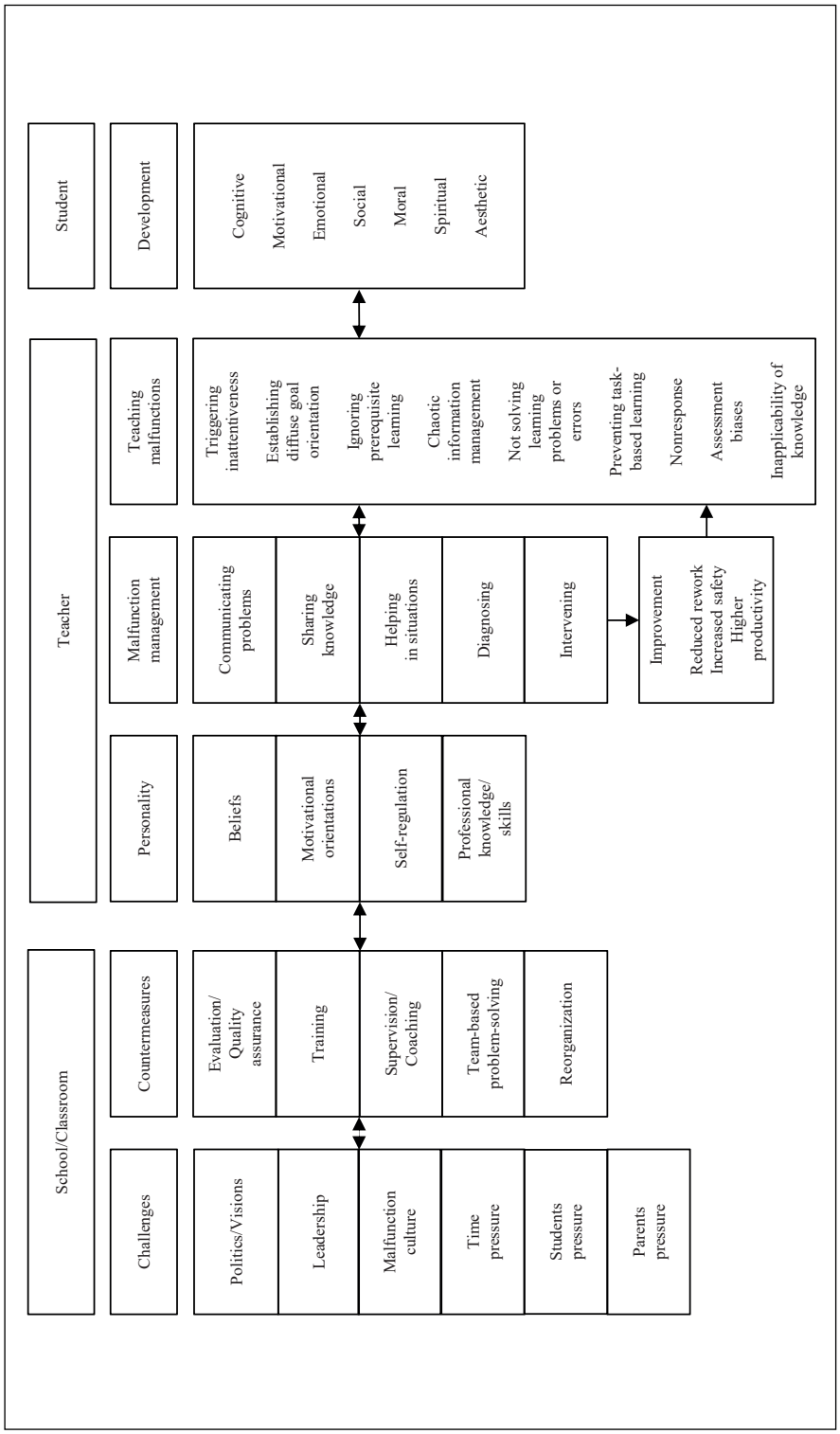

Fig. 1: A macro model on teaching malfunctions.

malfunctioning is simply ignored, sometimes it is being actively processed. Teaching is also acting under pressure. This pressure comes from time restrictions or requirements, the heterogeneity and scale of problems associated with students, and parental wishes and ideas.

Second, whether challenges have a positive or negative effect on teachers, depends on countermeasures which represent all educational and organizational measures to help teachers to avoid or reduce negative effects on their teaching. Such measures concern traditionally used procedures as evaluation resp. quality assurance, training, supervision, and coaching as well as team-based problem-solving. Reorganization concerns restructuring or reforming attempts in school and classroom contexts like workplace learning and systematic reflection in order to avoid teaching malfunctions to reoccur (e.g., Imants, 2002).

Third, it is assumed that the personality of the teacher influences malfunction management, and related teaching malfunctions (e.g., Klassen & Tze, 2014). Personality concerns more or less stable personality traits about teaching and quality of teaching like beliefs, motivational orientations, as well as self-regulation and professional knowledge and skills. For example, beliefs on teaching concern capabilities (skills for executing actions), or contexts (students, parents, other teachers, or administrators which influence the probability of producing certain results in teaching) (Lumpe, Haney, & Czerniak, 2000).

Fourth, even teachers with the most sophisticated personality traits will not be able to avoid teaching malfunctions. When teachers experience malfunctions and related problems in teaching, then it is important that they manage solving such problems. Malfunction management refers to all activities that assist teachers in realizing improvement on problems in teaching (e.g., Lampert, 2001). Improvement is given, for example, when there is reduced rework, increased safety, or higher productivity. Related activities on teaching malfunctions are about communicating problems, sharing knowledge, helping and assisting in difficult situations, diagnosing (assessments on problems), and intervening (implementation of problem-solving attempts).

Fifth, all the mentioned blocks of variables are related directly or indirectly to teaching malfunctions ranging from triggering inattentiveness to producing inapplicability of knowledge. At the current state of theory building and research, it is not possible to relate specific conditions to specific teaching malfunctions. At the moment, it can be mentioned, as indicator of validity, that the process model on teaching malfunctions is comparable or has similarities to other models of teaching effectiveness like the "model of selected organizational characteristics that contribute to student achievement" (Sweetland & Hoy, 2000), or "multilevel dynamic education model of school, teacher, and teacher effectiveness on student learning" (Ding & Sherman, 2006).

Sixth, it is assumed that teaching malfunctions are affecting student development. Student development has a cognitive facet concerning knowledge and skills, but also an affective one. Affective student engagement is about emotional (man-

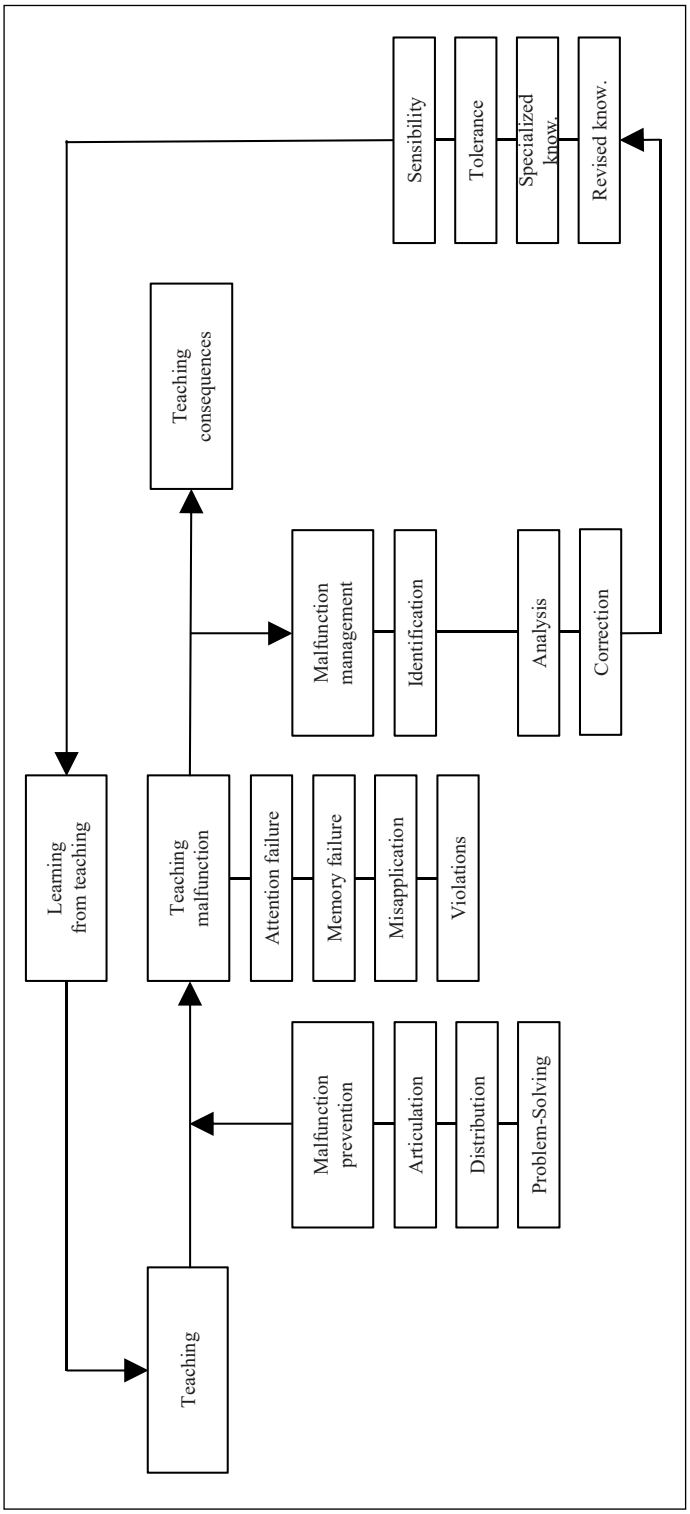

Fig. 2: A micro model on teaching malfunctions.

agement of feelings), moral (handling norms of behavior), social (building relationships), spiritual (focusing on meaning in life), aesthetic (recognizing and creating beauty), and motivational development (cultivating interests) (Martin & Reigeluth, 1999).

This process model represents, up to our best knowledge, a first organized collection of assumptions on teaching malfunctions and its possible causes. The model can be considered as a theory in a very early stage. Of course, such a theory needs refinement and testing based on empirical research.

A first step in refinement is to consider a micro model on teaching malfunctions (see Figure 2). Such a micro perspective focuses on components of actions and related cognitive processes which are involved in teaching malfunctions. The micro-model is based on research and related models from Frese and Keith (2015; on error prevention, management, and associated processes and outcomes), Petkova (2009; on entrepreneurial learning from performance errors), Rach, Ufer, and Heinze (2013; on a process model for learning in error situations), and Reinach and Viale (2006; on a human error framework).

Based on this model, it is assumed that during teaching, there is a high probability that malfunctions occur. Teaching malfunctions lead to consequences for teaching. Such consequences stimulate learning from teaching which influences teaching itself. Teaching consequences concern the effectiveness (ineffectiveness), efficiency (inefficiency), robustness (instability), and (positive and negative) side effects of teaching and teaching malfunctions. Teaching malfunctions can be reduced by malfunction prevention and can be changed in their consequences by malfunction management. Teaching malfunctions are based on cognitive processes concerning attention failures, memory failures, misapplications, or violations. Malfunction prevention is about articulating and distributing information on teaching malfunctions as well as related problem-solving. Malfunction management concerns the identification, analysis, and correction of teaching malfunctions. Malfunction management affects learning for teaching as it produces revised and specialized knowledge on teaching as well as certain cognitive orientations like tolerance and sensibility to teaching malfunctions.

Discussions

Based on these theoretical foundations, it has to be clarified, in future research activities, how and with what kind of assessments, it might be possible to measure teaching malfunctions with high reliability and validity. Again, there are assessments for learning deficits and errors, but not for teaching malfunctions. In addition, research scenarios and designs have to be developed which allow to test for effects of teaching malfunctions and their changes. Research on teaching malfunctions is not a simple task as teaching takes place in complex and dynamic contexts which need sophisticated research strategies. In recent approaches on research on learning and instruction, such a focus on teaching malfunctions and related research strategies is

still missing. Furthermore, it is clear that a focus on teaching malfunctions has to be handled in a sensible way in teacher education as people do not like to be confronted with their own performance problems. An emotionally-sound instructional design for courses on teaching malfunctions in teacher education has to be developed. There are emotional design of instruction approaches for student learning, which could be used for teacher education (e.g., Astleitner, 2000).

Having a list of teaching malfunctions and related process models represents a first starting point for doing empirical research. Such a research on teaching malfunctions can use all traditional social research methods and designs. However, as teaching malfunctions represent an issue which is uncomfortable for those affected, special precautions must be taken. For example, a significant problem is social desirability bias. Teaching malfunctions are socially undesirable and therefore covered up, especially in surveys or interviews of teachers and students. Classroom observations are less prone to such errors when there are implemented ethically, for example, in nonlaboratory settings (e.g., in real instructional situations), with unobtrusive (e.g., as time-on-task in classrooms), unexpected (e.g., without notice), or disguised measures (e.g., with a focus on students). Another, less problematic alternative might be to use ratings of malfunctions with a strong observational character like, for example, the "Framework for Teaching Evaluation Instrument" from Danielson (2013). This instrument has multiple domains which range from "distinguished" to "unsatisfactory" and was conceptualized not to evaluate or test teachers, but to help them learn. Within such ratings, teaching malfunctions are measured in concrete behaviors that should more or less been counted by the raters.

Finally, it has to be stressed again, that classifying teaching as effective or ineffective, is difficult and remains an open question for research resp. intervention research in classrooms. Sometimes good teaching leads to bad results and vice versa (Schoenfeld, 1988). It is also true that is an advantage of a certain type of teaching for some students in learning may be a disadvantage for others (e.g., Shute, 2008). Sometimes, bad teaching can be compensated by media or textbooks (e.g., Van den Ham & Heinze, 2018). Sometimes, negative effects of teaching turn out to change over time into positive ones, and so on. Such a difficult situation should not lead to the conclusion that empirical intervention research is not very helpful in gaining valid knowledge on teaching and teaching malfunctions. On the contrary, it is much more important to do even more targeted research. Such a more targeted research has to deliver multiple evidence on teaching interventions and related malfunctions (e.g., Astleitner, Kriegseisen, & Riffert, 2009).

References

Allington, R. L., McCuiston, K., & Billen, M. (2015). What research says about text complexity and learning to read. *The Reading Teacher*, *68*, 491–501. doi: https://doi.org/10.1002/trtr.1280

Astleitner, H. (2000). Designing emotionally sound instruction: The FEASP-approach. *Instructional Science, 28*, 169–198. doi: https://doi.org/10.1023/A:1003893915778

Astleitner, H., Kriegseisen, J., & Riffert, F. (2009, September). *Using a multiple evidence model (MUEMO) for testing the effectiveness of educational interventions*. Paper presented at the European Conference of Educational Research (ECER). Vienna. Retrieved from http://seniorenuniversitaet.at/fileadmin/oracle_file_imports/1091171.PDF

Atjonen, P. (2014). Teachers' views of their assessment practice. *Curriculum Journal*, *25*, 238–259. doi: https://doi.org/10.1080/09585176.2013.874952

Balli, S. J. (2011). Pre-service teachers' episodic memories of classroom management. *Teaching and Teacher Education*, *27*, 245–251. doi: https://doi.org/10.1016/j.tate.2010.08.004

Barbetta, P. M., Norona, K. L., & Bicard, D. F. (2005). Classroom behavior management: A dozen common mistakes and what to do instead. *Preventing School Failure: Alternative Education for Children and Youth*, *49*, 11–19. doi: https://doi.org/10.3200/PSFL.49.3.11-19

Bengtsen, S., & Barnett, R. (2017). Confronting the dark side of higher education. *Journal of Philosophy of Education*, *51*, 114–131. doi: https://doi.org/10.1111/1467-9752.12190

Bieg, S., Grassinger, R., & Dresel, M. (2017). Humor as a magic bullet? Associations of different teacher humor types with student emotions. *Learning and Individual Differences*, *56*, 24–33. doi: https://doi.org/10.1016/j.lindif.2017.04.008

Blömeke, S., Suhl, U., & Kaiser, G. (2011). Teacher education effectiveness: Quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge. *Journal of Teacher Education*, *62*, 154–171. doi: https://doi.org/10.1177/0022487110386798

Boyle, J. R. (2010). Strategic note-taking for middle-school students with learning disabilities in science classes. *Learning Disability Quarterly*, *33*, 93–109. doi: https://doi.org/10.1177/073194871003300203

Bryant, L. E. (2003). Becoming a better teacher: Learning from our mistakes. *Communication Studies, 54*, 130–132. doi: https://doi.org/10.1080/10510970309363274

Budge, S. L. (2016). To err is human: An introduction to the special issue on clinical errors. *Psychotherapy*, *53*, 255–256. doi: http://dx.doi.org/10.1037/pst0000084

Campbell, C., & Evans, J. A. (2000). Investigation of preservice teachers' classroom assessment practices during student teaching. *The Journal of Educational Research*, *93*, 350–355. doi: https://doi.org/10.1080/00220670009598729

Carr-Chellman, A. A. (2016). *Instructional design for teachers. Improving classroom practice* (2nd ed.). New York: Routledge. doi: https://doi.org/10.4324/9781315773032

Cooper, H. M. (1989). *Integrating research. A guide for literature reviews* (2nd ed.). Newbury Park: Sage.

Creemers, B. P. M., & Kyriakides, L. (2012). *Improving quality in education. Dynamic approaches to school improvement*. Abingdon: Routledge. doi: https://doi.org/10.4324/9780203817537

Cronjé, J. (2006). Paradigms regained: Toward integrating objectivism and constructivism in instructional design and the learning sciences. *Educational Technology Research and Development*, *54*, 387–416. doi: https://doi.org/10.1007/s11423-006-9605-1

Danielson, C. (2013). The Framework for Teaching Evaluation Instrument. Retrieved from https://usny.nysed.gov/rttt/teachers-leaders/practicerubrics/Docs/danielson-teacher-rubric-2013-instructionally-focused.pdf

Daschmann, E. C., Goetz, T., & Stupnisky, R. H. (2011). Testing the predictors of boredom at school: Development and validation of the precursors to boredom scales. *British Journal of Educational Psychology*, *81*, 421–440. doi: https://doi.org/10.1348/000709910X526038

Dawson, P., & Dawson, S. L. (2018). Sharing successes and hiding failures: 'Reporting bias' in learning and teaching research. *Studies in Higher Education*, *43*, 1405–1416. doi: https://doi.org/10.1080/03075079.2016.1258052

Dekker, S. (2014). *The field guide to understanding 'human error'* (3rd ed.). London: CRC Press.

Dhillon, B. S. (2009). *Human reliability, error, and human factors in engineering maintenance: With reference to aviation and power generation*. Boca Raton: CRC Press. doi: https://doi.org/10.1201/9781439803844

Dhindsa, H. S., & Anderson, O. R. (2011). Constructivist-visual mind map teaching approach and the quality of students' cognitive structures. *Journal of Science Education and Technology*, *20*, 186–200. https://doi.org/10.1007/s10956-010-9245-4

Diekema, A. R., & Olsen, M. W. (2011). Personal information management practices of teachers. *Proceedings of the American Society for Information Science and Technology*, *48*, 1–10. doi: https://doi.org/10.1002/meet.2011.14504801189

Ding, C., & Sherman, H. (2006). Teaching effectiveness and student achievement: Examining the relationship. *Educational Research Quarterly*, *29*, 40–51. Retrieved from https://eric.ed.gov/?id=EJ781882

Fernández-Alonso, R., Suárez-Álvarez, J., & Muñiz, J. (2015). Adolescents' homework performance in mathematics and science: Personal factors and teaching practices. *Journal of Educational Psychology, 107*, 1075–1085. https://doi.org/10.1037/edu0000032

Fisher, D. (2009). The use of instructional time in the typical high school classroom. *The Educational Forum, 73*, 168–176. doi: https://doi.org/10.1080/00131720902739650

Frese, M., & Keith, N. (2015). Action errors, error management, and learning in organizations. *Annual Review of Psychology*, *66*, 661–687. doi: https://doi.org/10.1146/annurev-psych-010814-015205

Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., Hosp, M., & Jancek, D. (2003). Explicitly teaching for transfer: Effects on third-grade students' mathematical problem solving. *Journal of Educational Psychology, 95*, 293–305. doi: https://doi.org/10.1037/0022-0663.95.2.293

Gagné, R. M. (1985*). The conditions of learning and theory of instruction* (4th ed.). New York: Holt, Rinehart and Winston.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*, 393–408. doi: https://doi.org/10.1037/0022-0663.95.2.393

Gluschkoff, K., Elovainio, M., Kinnunen, U., Mullola, S., Hintsanen, M., Keltikangas-Järvinen, L., & Hintsa, T. (2016). Work stress, poor recovery and burnout in teachers. *Occupational Medicine*, *66*, 564–570. doi: https://doi.org/10.1093/occmed/kqw086

Hattie, J. (2008). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London, New York: Routledge. doi: https://doi.org/10.4324/9780203887332

Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation*, *38*, 21–27. doi: https://doi.org/10.1016/j.stueduc.2012.04.001

Hebert, M., Bohaty, J. J., Nelson, J. R., & Brown, J. (2016). The effects of text structure instruction on expository reading comprehension: A meta-analysis. *Journal of Educational Psychology*, *108*, 609–629. doi: https://doi.org/10.1037/edu0000082

Hiebert, J., Morris, A. K., Berk, D., & Jansen, A. (2007). Preparing teachers to learn from teaching. *Journal of Teacher Education*, *58*, 47–61. doi: https://doi.org/10.1177/0022487106295726

Howard, J. (2019). *Cognitive errors and diagnostic mistakes. A case-based guide to critical thinking in medicine*. Cham: Springer. doi: https://doi.org/10.1007/978-3-319-93224-8

Huang, J. (2002). Error analysis in English teaching: A review of studies. *Journal of Chung-San Girls' Senior High School*, *2*, 19–34. Retrieved from http://www.academia.edu/download/35855143/03ERROR_ANALYSIS.pdf

Imants, J. (2002). Restructuring schools as a context for teacher learning. *International Journal of Educational Research*, *37*, 715–732. doi: https://doi.org/10.1016/S0883-0355(03)00067-3

Jaccard, J., & Jacoby, J. (2010). *Theory construction and model-building skills. A practical guide for social scientists*. New York: Guilford.

Jackson, M. (2006). "Serving time": The relationship of good and bad teaching. *Quality Assurance in Education*, *14*, 385–397. doi: https://doi.org/10.1108/09684880610703965

Jonassen, D. H. (1999). Designing constructivist learning environments. In C. M. Reigeluth (Ed.), *Instructional-design theories and models. A new paradigm of instructional* theory (Vol. II, pp. 215–239). Mahwah: Erlbaum.

Keith, N., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology*, *93*, 59–69. doi: https://doi.org/10.1037/0021-9010.93.1.59

Keller, J. M. (1987). Strategies for stimulating the motivation to learn. *Performance + Instruction*, *26*, 1–7. doi: https://doi.org/10.1002/pfi.4160260802

Kercood, S., & Grskovic, J. A. (2009). The effects of highlighting on the math computation performance and off-task behavior of students with attention problems. *Education and Treatment of Children, 32*, 231–241. Retrieved from https://www.jstor.org/stable/42900020; doi: https://doi.org/10.1353/etc.0.0058

Kirschner, P., Sweller, J., & Clark, R. E. (2006). Why unguided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experien-

tial learning and inquiry-based learning. *Educational Psychologist*, *41*, 75–86. doi: https://doi.org/10.1207/s15326985ep4102_1

Klassen, R. M., & Tze, V. M. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review*, *12*, 59–76. doi: https://doi.org/10.1016/j.edurev.2014.06.001

Kletz, T. (2001). *An engineer's view on human error* (3rd ed.). Boca Raton: CRC Press.

Ko, J., & Sammons, P. (2013). *Effective teaching: A review of research and evidence*. Reading: CfBT Education Trust. Retrieved from: https://files.eric.ed.gov/fulltext/ED546794.pdf

Kofler, M. J., Rapport, M. D., & Matt Alderson, R. (2008). Quantifying ADHD classroom inattentiveness, its moderators, and variability: A meta-analytic review. *Journal of Child Psychology and Psychiatry*, *49*, 59–69. doi: https://doi.org/10.1111/j.1469-7610.2007.01809.x

Krull, E., Oras, K., & Sisask, S. (2007). Differences in teachers' comments on classroom events as indicators of their professional development. *Teaching and Teacher Education*, *23*, 1038–1050. doi: https://doi.org/10.1016/j.tate.2006.02.001

Kunst, E. M., van Woerkom, M., & Poell, R. F. (2018). Teachers' goal orientation profiles and participation in professional development activities. *Vocations and Learning*, *11*, 91–111. doi: https://doi.org/10.1007/s12186-017-9182-y

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

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143–152. doi: https://doi.org/10.1016/j.tate.2013.07.010

Lampert, M. (1985). How do teachers manage to teach? Perspectives on problems in practice. *Harvard Educational Review*, *55*, 178–195. doi: https://doi.org/10.17763/haer.55.2.56142234616x4352

Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven: Yale University Press.

Le Maistre, C., & Paré, A. (2010). Whatever it takes: How beginning teachers learn to survive. *Teaching and Teacher Education*, *26*, 559–564. doi: https://doi.org/10.1016/j.tate.2009.06.016

Lo, C. K., & Hew, K. F. (2017). A critical review of flipped classroom challenges in K-12 education: Possible solutions and recommendations for future research. *Research and Practice in Technology Enhanced Learning*, *12*, 4. doi: https://doi.org/10.1186/s41039-016-0044-2

Lumpe, A. T., Haney, J. J., & Czerniak, C. M. (2000). Assessing teachers' beliefs about their science teaching context. *Journal of Research in Science Teaching*, *37*, 275–292. doi: https://doi.org/10.1002/(SICI)1098-2736(200003)37:3<275::AID-TEA4>3.0.CO;2-2

Magliaro, S. G., Lockee, B. B., & Burton, J. K. (2005). Direct instruction revisited: A key model for instructional technology. *Educational Technology Research and Development*, *53*, 41–55. doi: https://doi.org/10.1007/BF02504684

Martin, B. L., & Reigeluth, C. M. (1999). Affective education and the affective domain: Implications for instructional-design theories and models. In C. M. Reigeluth (Ed.), *Instructional-design theories and models. A new paradigm of instructional theory* (Vol. II, pp. 485–509). Mahwah: Erlbaum.

Martin, F., Klein, J., & Sullivan, H. (2004). *Effects of instructional events in computer-based instruction*. Chicago: Association for Educational Communications and Technology.

Mayer, R. E. (1999). Designing instruction for constructivist learning. In C. M. Reigeluth (Ed.), *Instructional-design theories and models. A new paradigm of instructional theory* (Vol. II, pp. 141–159). Mahwah: Erlbaum.

Maynard, B. R., Solis, M. R., Miller, V. L., & Brendel, K. E. (2017). Mindfulness-based interventions for improving cognition, academic achievement, behavior, and socioemotional functioning of primary and secondary school students. *Campbell Systematic Reviews*, *13*, 1–144. doi: https://doi.org/10.4073/CSR.2017.5

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, *95*, 203–213. doi: https://doi.org/10.1080/00220670209596593

McNeill, K. L. (2009). Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena. *Science Education*, *93*, 233–268. doi: https://doi.org/10.1002/sce.20294

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489. doi: https://doi.org/10.1146/annurev-psych-010416-044022

Moeller, A. J., Theiler, J. M., & Wu, C. (2012). Goal setting and student achievement: A longitudinal study. *The Modern Language Journal*, *96*, 153–169. doi: https://doi.org/10.1111/j.1540-4781.2011.01231.x

Morgan, P. L. (2006). Increasing task engagement using preference or choice-making: Some behavioral and methodological factors affecting their efficacy as classroom interventions. *Remedial and Special Education*, *27*, 176–187. doi: https://doi.org/10.1177/07419325060270030601

Ness, M. (2016). Learning from K-5 teachers who think aloud. *Journal of Research in Childhood Education*, *30*, 282–292. doi: https://doi.org/10.1080/02568543.2016.1178671

Oliver, R. M., Wehby, J. H., & Reschly, D. J. (2011). Teacher classroom management practices: Effects on disruptive or aggressive student behavior. *Campbell Systematic Reviews*, *7*, 1–55. doi: https://doi.org/10.4073/csr.2011.4

Owens, J. (2013). *Confessions of a bad teacher. The shocking truth from the front lines of American public education*. Naperville: Sourcebooks.

Paas, F. G., & Van Merriënboer, J. J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*, 122–133. doi: https://doi.org/10.1037/0022-0663.86.1.122

Petkova, A. P. (2009). A theory of entrepreneurial learning from performance errors. *International Entrepreneurship and Management Journal*, *5*, 345. doi: https://doi.org/10.1007/s11365-008-0075-2

Phelps, P. H. (2000). Mistakes as vehicles for educating teachers. *Action in Teacher Education*, *21*, 41–49. doi: https://doi.org/10.1080/01626620.2000.10462979

Praetorius, A. K., Nitsche, S., Janke, S., Dickhäuser, O., Drexler, K., Fasching, M., & Dresel, M. (2014). Here today, gone tomorrow? Revisiting the stability of teachers' achievement goals. *Contemporary Educational Psychology*, *39*, 379–387. doi: https://doi.org/10.1016/j.cedpsych.2014.10.002

Rach, S., Ufer, S., & Heinze, A. (2013). Learning from errors: Effects of teachers training on students' attitudes towards and their individual use of errors. *PNA*, *8*, 21–30. Retrieved from https://eric.ed.gov/?id=EJ1054922

Raufelder, D., Nitsche, L., Breitmeyer, S., Keßler, S., Herrmann, E., & Regner, N. (2016). Students' perception of "good" and "bad" teachers – Results of a qualitative thematic analysis with German adolescents. *International Journal of Educational Research*, *75*, 31–44. doi: https://doi.org/10.1016/j.ijer.2015.11.004

Reason, J. (1990). *Human error*. Cambridge: University Press. doi: https://doi.org/10.1017/CBO9781139062367

Reinach, S., & Viale, A. (2006). Application of a human error framework to conduct train accident/incident investigations. *Accident Analysis & Prevention*, *38*, 396–406. doi: https://doi.org/10.1016/j.aap.2005.10.013

Renkl, A. (2011). Instruction based on examples. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 272–295). New York: Routledge.

Schacter, J., & Thum, Y. M. (2004). Paying for high-and low-quality teaching. *Economics of Education Review*, *23*, 411–430. doi: https://doi.org/10.1016/j.econedurev.2003.08.002

Scheerens, J. (2015). Theories on educational effectiveness and ineffectiveness. *School Effectiveness and School Improvement, 26,* 10–31. doi: http://dx.doi.org/10.1080/09243453.2013.858754

Schmidt, C. P., Andrews, M. L., & McCutcheon, J. W. (1998). An acoustical and perceptual analysis of the vocal behavior of classroom teachers. *Journal of Voice*, *12*, 434–443. doi: https://doi.org/10.1016/S0892-1997(98)80052-0

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of well-taught mathematics courses. *Educational Psychologist*, *23*, 145–166. doi: https://doi.org/10.1207/s15326985ep2302_5

Seidel, T., Rimmele, R., & Prenzel, M. (2005). Clarity and coherence of lesson goals as a scaffold for student learning. *Learning and Instruction*, *15*, 539–556. doi: https://doi.org/10.1016/j.learninstruc.2005.08.004

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153–189. doi: https://doi.org/10.3102/0034654307313795

Spires, H. A., & Donley, J. (1998). Prior knowledge activation: Inducing engagement with informational texts. *Journal of Educational Psychology*, *90*, 249–260. doi: https://doi.org/10.1037/0022-0663.90.2.249

Stokking, K., Leenders, F., De Jong, J., & Van Tartwijk, J. (2003). From student to teacher: Reducing practice shock and early dropout in the teaching profession. *European Journal of Teacher Education*, *26*, 329–350. doi: https://doi.org/10.1080/0261976032000128175

Strauch, B. (2018). *Investigating human error: Incidents, accidents, and complex systems*. Abingdon: Routledge. doi: https://doi.org/10.4324/9781315183053

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*, 339–355. doi: https://doi.org/10.1177/0022487111404241

Sweetland, S. R., & Hoy, W. K. (2000). School characteristics and educational outcomes: Toward an organizational model of student achievement in middle schools. *Educational Administration Quarterly*, *36*, 703–729. doi: https://doi.org/10.1177/00131610021969173

Treiber, F. (1984). Ineffective teaching: Can we learn from it? *Journal of Teacher Education*, *35*, 45–47. doi: https://doi.org/10.1177/002248718403500511

Tulis, M. (2013). Error management behavior in classrooms: Teachers' responses to student mistakes. *Teaching and Teacher Education*, *33*, 56–68. doi: https://doi.org/10.1016/j.tate.2013.02.003

Van den Bergh, L., Ros, A., & Beijaard, D. (2013). Teacher feedback during active learning: Current practices in primary schools. *British Journal of Educational Psychology*, *83*, 341–362. doi: https://doi.org/10.1111/j.2044-8279.2012.02073.x

Van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice*, *32*, 46–60. doi: https://doi.org/10.1111/ldrp.12122

Van den Ham, A. K., & Heinze, A. (2018). Does the textbook matter? Longitudinal effects of textbook choice on primary school students' achievement in mathematics. *Studies in Educational Evaluation*, *59*, 133–140. doi: https://doi.org/10.1016/j.stueduc.2018.07.005

Van Merriënboer, J. J., & Kirschner, P. A. (2013). *Ten steps to complex learning* (2nd ed.). New York: Routledge. https://doi.org/10.4324/9780203096864

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research*, *54*, 143–178. doi: https://doi.org/10.3102/00346543054002143

Walls, R. T., Nardi, A. H., von Minden, A. M., & Hoffman, N. (2002). The characteristics of effective and ineffective teachers. *Teacher Education Quarterly*, *29*, 39–48. Retrieved from https://search.proquest.com/openview/9a953346067acbc70d3295f518ba0ebd/1?pq-origsite=gscholar&cbl=48404

Wedel, A., Müller, C. R., Pfetsch, J., & Ittel, A. (2019). Training teachers' diagnostic competence with problem-based learning: A pilot and replication study. *Teaching and Teacher Education*, *86*, 1–14. doi: https://doi.org/10.1016/j.tate.2019.102909

Yoon, H. G., Joung, Y. J., & Kim, M. (2012). The challenges of science inquiry teaching for pre-service teachers in elementary classrooms: Difficulties on and under

the scene. *Research in Science Education*, *42*, 589–608. doi: https://doi.org/10.1007/s11165-011-9212-y

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, *16*, 323–342. doi: https://doi.org/10.1207/S15324818AME1604_4

PART 2.
Design Problems

3. Missing Control Group: The Effect of a Self-Congruence Intervention on Teachers' Volitional Competences and Motive Implementation Strategies

Franz Hofmann & Hermann Astleitner

Abstract: The current study tested the effectiveness of a self-congruence intervention for secondary school teachers on volitional skills without having a control group. In a non-randomized pre-post-design, 26 teachers participated in a 42-hours training program over a one-and-a-half-year period. In order to have additional evidence and to simulate a control group, teachers were divided into two groups according to their goal orientations in daily instruction. In both groups, participants reported better volitional skills after the intervention. However, hypothesized positive effects occurred only on some volitional competences (concentration, coping with failure, and self-sense) and motive implementation strategies (coping and passive avoidance). Within discussion, further theoretical and methodological developments of the intervention and its implementation are suggested. Especially, research strategies when having a missing control group are briefly outlined.

Sometimes in educational intervention research, it is not possible to implement a control group although groups without having an intervention are indispensable for achieving internal validity. Despite their problems, such pre-experimental designs can often be found within evaluation research or field experiments (e.g., Marsden & Torgerson, 2012). Missing control groups are also given within research on teacher education (e.g., De Boer, Timmermans, & Van der Werf, 2018). In this study, we focused on certain widely overlooked motivational issues in teacher education. For gaining new and innovative knowledge on such issues, we accepted a missing control group.

Motivational dispositions and processes play an important role in the teaching profession. Research indicated that motivation affected directly or indirectly, for example, reasons for choosing the teaching profession, commitment and responsibility for the job, job satisfaction and well-being, active participation in trainings, implementation of innovations in schools, preparation for daily instruction, teaching effectiveness in the classroom, and the motivation as well as the learning of students (e.g., Klassen & Tze, 2014; Lam, Cheng, & Choy, 2010; Schiefele, 2017; Sinclair, 2008; Wang, Hall, & Rahimi, 2015; Wayne & Youngs, 2003). Considering such an impact, it does not surprise that numerous study programs, continuing

education approaches, or trainings were implemented to foster teacher's motivation (see Karabenick, Richardson, & Watt, 2014 and Watt, Richardson, & Smith, 2017 for an overview).

The basis for research-based activities on affecting teacher motivation is built on a wide range of strongly varying theoretical frameworks. For example, within the "personal investment theory" (Maehr & Braskamp, 1986), individuals' personal incentives (for accomplishment, recognition, power, and affiliation), perceptions of a sense of the self (e.g., goal directedness, self-esteem, and self-reliance), and perceived options (e.g., opportunities for advancement) were assumed to be related to purposeful actions of teachers. Neves de Jesus and Lens (2005) suggested an integrated "cognitive-motivational model" for teacher's professional motivation in which professional engagement was related to results (successes or failures), (internal/external and stable/unstable) attributions, expectancies (on control, efficacy, and success), intrinsic motivation, and goal values. Gokce (2010) stressed the importance to distinguish between motivation (for promotion, growth, achievement, interest in work, recognition, and responsibility) and related hygiene factors (like job security, pay, status, supervision, working conditions, organizational policy and management, as well as interpersonal relationships). According to Thoonen, Sleegers, Oort, Peetsma, and Geijsel (2011), teacher motivation consists of teacher self-efficacy, internalization of school goals, tolerance of uncertainty, and well-being. Within a factorial approach, Roness (2011) found teacher motivation to be intrinsically (e.g., to like the task of teaching), altruistically (e.g., to help students to achieve their goals), extrinsically (e.g., to get career benefits), and subject-matter (e.g., to have an interest in certain themes and issues) orientated.

A volitional perspective on teacher motivation. Although, many different variables on teacher motivation were identified in research, two main perspectives for development remain. Firstly, many of these models and related research activities are focusing on expectancy-value approaches which are covering mainly the selection of goals or intentions, but not the execution or implementation of actions. The later aspect is about "self-control", "action-control", or (most widely used) "volition" "during which the individual develops strategies and plans in order to ensure that their intention will be enacted" (Milne, Orbell, & Sheeran, 2002, p. 166). In the field of motivational instructional design, for example, Keller (2008) distinguished between motivational strategies (for curiosity, values, and expectancies), implementation strategies (for pre-action planning), and volitional strategies (for supporting self-regulatory actions). Secondly, many of these models do not have a developmental perspective. In a developmental model, motivation is seen as a process that changes over time and that has to be developed step-by-step as a behavior-based competence (e.g., Hennecke & Freund, 2017). For example, Skinner and Beers (2016) presented a developmental model of teacher stress, coping, and everyday resilience with developmental volitional aspects as, for example, reducing impulsivity.

Evidence from research in educational settings stressed the effectiveness of volition and a developmental view for procrastination, dropout, academic perfor-

mance, training transfer, or conflict resolution (e.g., Deimann & Bastiaens, 2010; Duckworth et al., 2015; Seiberling & Kauffeld, 2017). However, although there is strong evidence for the effectiveness of volitional skills, surveys or interventions on teachers' volitional processes only delivered findings with many different overlapping concepts and inconclusive evidence. For example, Capa-Aydin, Sungur, and Uzuntiryaki (2009) identified volition-related skills (e.g., self-instruction or -reaction) as essential components of teacher's self-regulation. Perels, Merget-Kullmann, Wende, Schmitz, and Buchbinder (2009) considered volitional strategies within a training on self-regulation of kindergarten teachers. Erdogan and Senemoglu (2017) found positive effects of a problem-based-learning intervention on prospective teachers' academic achievements, but not on volition-based self-regulation. Lee and Turner (2017) investigated some volitional skills of pre-service teachers, like, for example, goal commitment and effort regulation and found them related to endogenous instrumentality (i.e., believing that a course content is related to a future goal).

Volitional competences, motive implementation strategies, and an intervention on self-congruence. A more systematic and conclusive theoretical approach in which volitional processes play a major role and which was repeatedly used in learning and education represents the personality systems interaction (PSI)-theory (Kuhl, 2000; Kuhl, Kazén, & Koole, 2006). This theory was based on modern trends in behavior-based competence assessment and development (e.g., Leutner, Fleischer, Grünkorn, & Klieme, 2017; Ortner & van de Vijver, 2015) and was already applied in the field of teacher education (e.g., Wagner, Baumann, & Hank, 2016). In respect to volition, there are two main contributions from this theory which we considered as essential for fostering teacher motivation. One is about volitional competences and one about implementation strategies for satisfying motives (on affiliation, achievement, power, and freedom). "Volitional competences" (Forstmeier & Rüddel, 2008, p. 66) concern two basic modes of action control: Self-regulation (as a self-integrating mode of volition) and self-control (as a self-disciplining mode of volition). Self-control includes, for example, goal recollection, forgetfulness prevention, or planning skills. Self-regulation comprises, for example, attentional focusing, self-motivation, or emotion regulation. Motive-related "implementation strategies" (Schüler, Brandstätter, Wegner, & Baumann, 2015, p. 843) concern approach or avoidance behaviors that are associated with the realization of motives like context-sensitive switching, intuitive behavior control, constructive coping, or goal-orientated acting (e.g., Baumann, Kazén, & Kuhl, 2010).

Volitional competences and motive implementation strategies are compounded of numerous different interacting sub-skills what makes it difficult to affect them by educational measures. For educational purposes and training programs, an important goal in fostering volitional competences and strategies is that individuals learn to make decisions about goals and related activities which are congruent with their motives (Rheinberg & Engeser, 2010). Kuhl and Quirin (2011, p. 76) assumed such a "self-congruence" as given, "when specific goals and other conscious thoughts are in line with global personal goals and values".

Being self-congruent for teachers also means not to handle too many goals and values at the same time or to be capable of dealing with multiple goals (e.g., Parker, Martin, Colmar, & Liem, 2012). So, for our study, we assume that volitional processes are related to teachers' goal orientations (on single or multiple goals) and on handling them in a balanced way (e.g., Boekaerts, de Koning, & Vedder, 2006). We also used this assumption in our study for building and comparing two groups of teachers. So, there is a theoretical basis for the simulation of our missing control group.

An intervention on teacher's self-congruence. Based on these assumptions, the first author of this study developed an intervention program on fostering volitional competences and motive implementation strategies related to self-congruence decision making. Within an 18-month period, four two-day training modules were held with an overall duration of about 42 hours. There were two trainers, one of them being the first author of this study[1]; both trainers were equally competent concerning the theoretical background; in phases of working in two groups participants were randomized. Six goals were in the center of the intervention: (a) Teachers should acquire basic psychological knowledge about volitional processes on the PSI-theory, especially on self-regulation and self-control. (b) They should relate basic educational concepts (e.g., empathy) to the PSI-theory and the own personality development. (c) They should be able to evaluate their own skills for implementing the affiliation- and the power-motive into daily classroom instruction. (d) They should know strategies on how to act in a self-congruent way in stressful situations before they start interacting with other people (e.g., students or parents). (e) They should apply their knowledge on personality-related motivational mechanisms for diagnosing activities in educational practice. (f) They should be aware about their volitional competences and motive implementation strategies and know which of them might have to be improved in order to handle stressful situations in the classroom.

In the training modules, online and/or printed materials (tests, Powerpoint slides, and work sheets) covering research findings, practical problems, and self-assessment tools were used. After the first, second, and third training modules, teachers were stimulated to transfer experiences from the modules into daily problem-solving activities in classrooms and other school contexts. Module 1 of the intervention included preparation (with a recording of an instructional unit, a pretest on volitional competences, and a written interview on the professional self-image), and a unit on the basics of the PSI-theory (about motives, goals, personality characteristics, power motives, personal developmental plans) together with a pretest on motive implementation strategies. Module 2 was on the diagnosis of volitional competences and implementation strategies. Teachers learned about the affiliation and the achievement motive, motive congruences, diagnosing the own person and others in a group-based scenario, and finally about giving feedback on diagnostic results. Module 3 focused on educational and instructional consequences. Teachers

1 We thank Dr. Gabriele Salzgeber for her work as the second trainer.

experienced what it means to be empathic (based on the PSI-theory), how to design educational activities on a process- and product-perspective, how to deal with classroom management and self-regulation as well as how to consider interdisciplinary learning goals. Finally, in module 4, teachers learned about the freedom motive, practiced group-based evaluations, and experienced how to make innovations visible.

Purpose and Hypotheses

Our objective was to explore the effect patterns of the self-congruence intervention on teachers' volitional competences and motive implementation strategies. We used a pre-posttest-design to examine this question in two (artificial or simulated) groups of teachers with single or multiple goal orientations. Our exploratory hypotheses were anchored within research on motive congruence and volitional processes (Sheldon, 2014; Thrash, Maruskin, & Martin, 2012).

We hypothesized that a self-congruence intervention should increase positive and decrease negative volitional competences. Positive volitional competences (e.g., self-determination and planning skills) are ones that increase action control and the successful implementation of intentions. Negative volitional competences (appraisals about individual burden and threat) however decrease the enactment of intentions. Self-congruence should lead to a more positive view of being able to handle stressful situations because environmental demands are newly framed and therefore relativized by individual motives. Better self-access should reduce the experience of individual burden and threat. The self-congruence intervention should also increase active implementation strategies (e.g., coping) and decrease passive implementation strategies (e.g., passive avoidance): Self-congruence should have an activation effect, because it facilitates the realization of a series of goals – even they seem contradictory at a first glance – and therefore gain additional volitional resources.

We also hypothesized that both groups of teachers demonstrate increased volitional competences and strategies as they both participated in the intervention. Stimulating self-congruence should help both groups of teachers to effectively handle their individual goal orientations and to acquire equally well volitional competences and motive implementation strategies. We assumed that a focus on self-congruence stimulates the balancing of goals for both groups of teachers. For teachers with single goal orientations, an intervention in self-congruence should expand their goal orientations and therefore activate volitional resources for achieving additional goals. For teachers with multiple goal orientations, a self-congruence intervention should increase the evaluation and balancing of different goal orientations what should allow them to gain volitional capacities.

Method

Participants

In the beginning, a total sample of 48 secondary school teachers from urban schools in Austria and Germany started in the intervention. However, during the intervention period, 22 teachers dropped out, so that data from only 22 female and 4 male participants were included in this study. We have no data about the reasons for dropout. Research indicates that dropout in teacher trainings usually has multiple reasons (e.g., Basit et al., 2006). Therefore, we do not expect any systematic effect from dropout on the internal and external validity of this study (Mitchell & Jolley, 2010). The ages of the remaining teachers ranged from 26 to 57 years (M = 47.35, SD = 7.2) and they had an average of 18.27 years of teaching experience (SD = 7.51). 23.1% of teachers rated themselves as being "very satisfied" with their profession as a teacher. 15.4% of teachers indicated that they consider themselves as "very successful" in the professional activities.

Research Design

Pre-experimental design without control group. In order to get evidence about effect patterns of the self-congruence intervention, a quasi-experimental pre-posttest design was implemented. The pretest on teachers' volitional competences was taken online four weeks before the beginning of the intervention, the posttest was taken 8 to 10 weeks after the last intervention session. Motive implementation strategies were measured (as pretest) at the beginning of the first intervention session and (as posttest) at the beginning of the last session as a paper-pencil test. There was no control group, because it was not possible to establish such a group within the intervention settings. However, in order to get more evidence about the effect patterns of the intervention and to simulate an additional impact assessment, an artificial control group was created. An artificial (or "simulated" or "post hoc") control group allows to have some additional comparison tests where there is no control group available due to practical, ethical, medical, or other reasons (e.g., Lise, Seitz, & Smith, 2015). Before the intervention, teachers were asked about their goal orientations in daily instruction in school. Teachers stated goals on acceptance of other people (e.g., respect or empathy) or on duties (e.g., discipline or punctuality). Then they were divided into two groups: 54% of teachers had one main goal orientation (acceptance or duty) (i.e., single goal orientation group) and 46% had both goal orientations (acceptance and duty) (i.e., multiple goal orientation group).

Acceptance of intervention and measurements. An acceptability rating for teachers served as an indicator for acceptance of the intervention and the measurements. At the end of the intervention, teachers were asked about how difficult it was for them to answer the questions on volitional skills and whether they profited from the contents of the intervention for their daily professional practice. 84.6% of teachers

indicated (mostly or definitely) that answering was easy and 92.3% (mostly or definitely) recognized professional profits from the intervention.

Measures

Teachers' volitional competences were measured with a short form of the Volitional Components Questionnaire (VCQ-S3; German version: Selbststeuerungs-Inventar: SSI-K3) (Kuhl & Fuhrmann, 2004). The questionnaire consisted of five constructs (self-regulation, self-control, volitional development, self-access, and life stress), 13 dimensions, and 52 items. Each of the 13 dimensions was measured with four items using a four-point Likert scale (on agreement: 1 = "not at all" to 4 = "definitely"). The dimensions (each with an example item; translations by the authors of this study) concerned: Self-determination ("In almost everything what I do in everyday life, I feel that I do it voluntarily"), self-motivation ("When my endurance decreases, I usually know exactly how to strengthen my desire for it"), self-relaxation ("I can reduce nervousness very specifically"), planning skill ("When I have to handle many things, I make myself a schedule (i.e., I set what I do when)"), fearless goal orientation ("In order to motivate myself, I often imagine what happens when I do not handle the matter in time"), initiating control ("When something needs to be done, then I start without hesitation"), intention enactment ("I often put off unpleasant things"), powers of concentration ("My thoughts often wander involuntarily from the matter which I am currently working on"), coping with failures ("After unpleasant experiences, I often do not get out of pondering for a whole time"), self-sense ("When I am sad, I lose the sense of what I really want"), integration ("My behavior often seems contradictory, because again and again another side of me emerges"), burden ("Occupation resp. education are currently very stressful for me"), and threat ("A lot has changed in my life that I have to deal with"). Scores on the dimensions were computed by averaging responses across the relevant four items with higher scores indicating greater volitional competences. For competences related to burden and threat, the opposite was true: Lower scores indicated greater competences.

For measuring motive implementation strategies, the Operant Motive Test (OMT) was used (Kuhl, 2013). In the past, the OMT was confronted with numerous reliability and validity tests (e.g., Schüler, Brandstätter, Wegner, & Baumann, 2015). It was also administered in samples of teachers: For example, Baumann, Chatterjee, and Hank (2016) analyzed implementation strategies for the power motive of future teachers. The OMT includes pictures on situations that are representative for major motive themes like affiliation, achievement, power, and freedom. These pictures are accompanied with questions (e.g., "What is important for the person in this situation and what is the person doing?"). Questions address the presence of a motive, but also of five motive implementation strategies for each motive (resulting in a 4 x 5 matrix). For each picture, the coder has to check whether one of four motives is present or not and which types of implementation strategies are present. These implementation strategies represent basic affective sets across all types of motives

and are consummatory (e.g., with activities that are positive, self-congruent, conflict-free, or with intrinsic flow), based on an active approach (e.g., stimulus-bound, but also spontaneous, and with intuitive action), on coping (e.g., active regulation and integration of negative affect), on active avoidance (e.g., strategic, goal orientated, and planned action to avoid negative affect), and on passive avoidance (e.g., by procrastination, fantasizing help from others, or expecting a sudden turn for the better) (Alsleben & Kuhl, 2011). Scores on the dimensions were computed by averaging responses across the relevant four motives with higher scores indicating stronger motive implementation strategies.

Data Analysis

For testing our assumptions, data were analyzed using IBM SPSS 24 with the general linear model (GLM) on repeated measures representing an ANOVA on two-factors with repeated measures on one factor (Winer, Brown, & Michels, 1991, p. 509). Repeated measures analyses with factors time (pretest vs. posttest) and group (single goal vs. multiple goal) were employed to the expected increase of volitional competences. The same procedure was used to test the hypothesized change in motive implementation strategies. For all statistical analysis, the Alpha-level was set at $p < .05$ and no adjustments were made. Partial eta square (η^2_p) served as indicator of small (≤ .01), medium (about .06), or large (≥ .14) effect sizes (Lipsey, 1990, p. 58).

Results

Reliability and Validity

Descriptive information and correlations between study measures from the pretest can be seen in Table 1. Reliability coefficients range from .68 to .90 indicating good reliabilities. For all dimensions of volitional competences, Cronbach's Alpha *(α)* was computed based on four items, for the measurement of planning skills only three items were used due to a negative item-total correlation. For the five motive implementation strategies, reliability coefficients concern interrater agreement between two ratings done by the first author of this study and a trained research assistant[2]. Interrater reliabilities (r_2) represent the correlation between the first and second rating according to the affect level. Correlations between study measures deliver some validity information. In Table 1, correlations reveal that variables between positive volitional competences (from self-determination to integration), negative volitional competences (burden and threat), and motive implementation strategies have many significant relations. However, between these three blocks of variables, nearly all correlations are not significant indicating that they represent different and unique aspects of teachers' volitional skills. Overall, high intra-block- and low in-

2 We thank Julia Maria Keller MA for her work concerning the ratings.

Tab. 1: Descriptive Statistics, Reliability Information, and Correlation Between Study Measures (Pretest)

Variables	M	SD	α r_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Volitional competences																					
1: Self-determination	1.86	0.44	.69	-																	
2: Self-motivation	1.86	0.48	.80	.67**	-																
3: Self-relaxation	1.52	0.64	.84	.60**	.60**	-															
4: Planning skill	2.00	0.51	.75	-.06	.01	.28	-														
5: Fearless goal ori.	2.56	0.48	.73	.27	.50**	.61**	.23	-													
6: Initiating control	1.90	0.66	.82	.43*	.53**	.61**	.43*	.35	-												
7: Intention enact.	1.96	0.58	.76	.38	.51**	.58**	.24	.39*	.64**	-											
8: Powers of concent.	2.09	0.53	.85	.22	.34	.32	.20	.34	.16	.49*	-										
9: Coping with fail.	1.87	0.62	.84	.47*	.46*	.69**	.15	.44*	.44*	.50**	.25	-									
10: Self-sense	1.97	0.74	.90	.56**	.58**	.66**	.22	.20	.57**	.67**	.38	.59**	-								
11: Integration	2.71	0.43	.85	.25	.60**	.43*	.21	.48*	.32	.59**	.49*	.35	.55**	-							
12: Burden	.57	0.59	.89	-.33	-.04	-.08	.15	.05	-.03	.10	-.07	-.01	.07	-.10	-						
13: Threat	.63	0.68	.88	-.24	-.05	.05	.38	.26	-.04	.13	.11	.15	.05	.01	.82**	-					
Motive implementation strategies																					
14: Consummatory	1.50	1.30	.73	-.04	.31	.04	-.06	.35	-.04	.00	.07	-.16	-.04	.25	.12	.04	-				
15: Active approach	3.54	1.24	.68	.19	.35	.33	.13	.03	.39*	-.03	-.06	.29	.25	.04	.00	-.04	.07	-			
16: Coping	1.27	1.19	.72	.08	-.09	.05	-.31	.18	-.10	-.03	.04	.13	-.04	-.02	.09	.03	-.04	-.18	-		
17: Active avoidance	4.27	1.76	.75	-.05	-.31	-.05	.22	-.16	-.17	-.10	-.02	-.02	-.13	-.10	-.34	-.10	-.55**	-.33	-.11	-	
18: Passive avoid.	3.81	1.81	.88	-.04	-.09	-.20	.04	-.30	-.00	.06	-.14	-.24	.02	-.06	.10	.03	-.23	-.33	-.50**	-.16	-

Note. N = 26. α = Cronbach's Alpha, r_2 = Interrater reliability, M = mean, SD = standard deviation, *p ≤ .05, **p ≤ .01.

Tab. 2: Descriptive Statistics and Repeated Measures Analysis of Variance Results for Volitional Competences

	Pretest		Posttest		Main effects						Interaction		
					Group (G)			Time (T)			G x T		
	M	(SD)	M	(SD)	F	p	η^2_p	F	p	η^2_p	F	p	η^2_p
Self-determin.					0.64	.434	.03	3.26	.085	.13	6.25	.020	.22
SG	1.71	0.30	2.13	0.48									
MG	2.09	0.45	2.02	0.64									
Self-motivation					0.52	.480	.02	0.37	.547	.02	.83	.372	.04
SG	1.77	0.35	1.89	0.38									
MG	1.98	0.63	1.96	0.65									
Self-relaxation					3.25	.085	.13	3.08	.093	.12	.49	.491	.02
SG	1.29	0.52	1.50	0.59									
MG	1.77	0.73	1.86	0.62									
Planning skill					0.13	.727	.01	2.23	.150	.09	.37	.551	.02
SG	2.03	0.59	2.13	0.76									
MG	1.88	0.34	2.12	0.60									
Fearless goal ori.					0.24	.630	.01	0.00	.973	.00	0.17	.688	.01
SG	2.54	0.39	2.58	0.45									
MG	2.50	0.59	2.46	0.46									
Initiating control					0.18	.679	.01	1.10	.306	.05	0.34	.563	.02
SG	1.83	0.63	1.79	0.56									
MG	1.98	0.76	1.84	0.53									
Intention enact.					0.79	.384	.04	4.09	.055	.16	2.26	.147	.09
SG	1.83	0.53	2.14	0.42									
MG	2.14	0.67	2.18	0.51									
Powers of conc.					2.22	.151	.09	9.60	.005	.30	0.01	.910	.00
SG	1.90	0.49	2.17	0.59									
MG	2.23	0.53	2.48	0.61									
Coping with fail.					2.35	.140	.10	5.35	.030	.20	0.00	1.00	.00
SG	1.73	0.57	1.98	0.56									
MG	2.09	0.67	2.34	0.73									
Self-sense					4.21	.052	.16	5.29	.031	.19	1.35	.257	.06
SG	1.69	0.65	2.04	0.58									
MG	2.34	0.78	2.46	0.71									
Integration					0.16	.697	.01	1.50	.234	.06	1.50	.234	.06
SG	2.77	0.33	2.62	0.53									
MG	2.61	0.56	2.61	0.61									
Burden					1.24	.277	.05	0.13	.722	.01	0.13	.722	.01
SG	0.46	0.49	0.37	0.36									
MG	0.64	0.62	0.64	0.83									
Threat					0.56	.462	.03	0.85	.366	.04	0.32	.579	.01
SG	0.48	0.57	0.44	0.42									
MG	0.68	0.58	0.52	0.53									

Note. N = 24. SG = single goal group, MG = multiple goal group.

ter-block-correlations indicated good (convergent and discriminant) validity of the measurements.

Intervention Effects on Volitional Competences

Pretest means, posttest means, standard deviations, and results of repeated measures analysis of variance on volitional competences are illustrated in Table 2. Overall and on a descriptive level, mean comparisons showed that the self-congruent intervention increased most of the positive aspects and decreased the negative aspects of volitional competences.

In detail and based on tests, ANOVAs revealed significant increase over time for all participants on the dimensions of powers of concentration ($F(1, 22) = 9.60$, $p = .005$, $\eta^2_P = .30$), coping with failures ($F(1, 22) = 5.35$, $p = .030$, $\eta^2_P = .20$), and self-sense ($F(1, 22) = 5.29$, $p = .031$, $\eta^2_P = .19$). A nearly significant increase over time was found for the dimensions of intention enactment ($F(1, 22) = 4.09$, $p = .055$, $\eta^2_P = .16$), self-determination ($F(1, 22) = 3.26$, $p = .085$, $\eta^2_P = .13$), and self-relaxation ($F(1, 22) = 3.08$, $p = .093$, $\eta^2_P = .12$). For the dimension of self-determination, a significant interaction effect indicated that an increase was only given for the group of teachers with single goal orientations ($F(1, 22) = 6.25$, $p = .020$, $\eta^2_P = .22$).

Analyses also revealed that teachers with multiple goal orientations had only higher self-sense ($F(1, 22) = 4.21$, $p = .052$, $\eta^2_P = .16$) and higher self-relaxation competences ($F(1, 22) = 3.25$, $p = .085$, $\eta^2_P = .13$) in comparison to the group with single goal orientations. No other significant differences between the two groups of teachers were found.

Intervention Effects on Motive Implementation Strategies

In Table 3, descriptive information and hypotheses tests for the motive implementation strategies are illustrated. Again, as in the case of volitional competences, the self-congruence-intervention increased most of active implementation strategies and decreased the passive strategy on an overall and descriptive level. ANOVA tests revealed that the motive implementation strategy of coping increased after the intervention ($F(1, 20) = 6.46$, $p = .019$, $\eta^2_P = .24$). Also, the intervention decreased the implementation strategy of passive avoidance ($F(1, 20) = 9.35$, $p = .006$, $\eta^2_P = .32$). No other significant or nearly significant group or interaction effects were found.

Tab. 3: Descriptive Statistics and Repeated Measures Analysis of Variance Results for Motive Implementation Strategies

	Pretest		Posttest		Main effects						Interaction		
					Group (G)			Time (T)			G x T		
	M	(SD)	M	(SD)	F	p	η^2_p	F	p	η^2_p	F	p	η^2_p
Consummatory					1.25	.278	.06	2.42	.136	.11	.53	.476	.03
SG	1.36	1.36	2.36	1.91									
MG	1.27	1.01	1.64	0.81									
Active approach					1.66	.212	.08	1.11	.304	.05	1.98	.175	.09
SG	2.91	1.04	3.00	1.10									
MG	3.82	1.17	3.18	1.33									
Coping					.12	.733	.01	6.46	.019	.24	.04	.847	.00
SG	1.55	1.21	2.64	1.91									
MG	1.27	1.27	2.55	2.02									
Active avoidance					.78	.388	.04	.13	.723	.01	.01	.943	.00
SG	4.46	1.86	4.73	2.05									
MG	4.00	2.00	4.18	2.04									
Passive avoidance					.77	.391	.04	9.35	.006	.32	1.72	.205	.08
SG	4.18	1.89	1.91	1.04									
MG	3.91	1.87	3.00	1.61									

Note. N = 22. SG = single goal group, MG = multiple goal group.

Discussion

In this study, we explored whether a participation in a self-congruence intervention would change teachers' volitional skills. Indeed and on a descriptive level, nearly all skills changed positively in the desired direction and about one third of the measured volitional competences and motive implementation strategies were higher after the intervention. In addition and as expected, it was found that the intervention produced nearly similar results for the two groups of teachers (with single or multiple goal orientations). These results indicate, for the first time, that a long-term intervention on self-congruence could positively affect teachers' volitional skills. The learning opportunities included in the self-congruence intervention proved to be appropriate and partly successful. These findings were also supported by the positive acceptability ratings at the end of the intervention.

From a theoretical perspective, it was hypothesized that a focus on self-congruence should (a) affect the view of being able to handle stressful situations, (b) lead to an activation effect, and (c) stimulate the balancing of goals. However, on the one hand, we did not measure or test these mediating variables. On the other hand, it is not clear whether these assumed mediating variables fit consistently with the complex PSI-theory. In general, it has to be mentioned that this study was not a test of the PSI-theory within an educational context. Rather, it was an exploration and a preliminary evaluation of effect patterns of an educational intervention that was based on the PSI-theory. The PSI-theory comes from basic research in the field of personality and social psychology, the intervention represents an instructional

package and an applied approach that integrated multiple perspectives from not only psychological research, but also from other research on teacher motivation or on the instructional design approaches (for theory building in applied disciplines, see Funnell & Rogers, 2011; Swanson & Chermack, 2013). So, the effects of the training were not only attributable to the PSI-theory, but also to our assumptions on the mediating variables and on our intervention design theory. An intervention design theory does not focus on the hypotheses in the subject area (e.g., about volitional competences), but on the design of the intervention in order to be effective in specific implementation contexts (e.g., motivating teachers to improve volitional competences).

Limitations. This study was limited in several ways in addition to the missing control group. First, the test of the self-congruence intervention was done, as far as we know, for the first time in the field of teacher education. So, additional evidence is necessary in order to have a more complete picture about the effect patterns of a self-congruence intervention for different teacher characteristics, or classroom settings. For example, Rupprecht, Paulus, and Walach (2017) tested successfully the influence of a mindfulness intervention on the self-regulation of teachers what comes close to our study conditions. Second, the small sample of teachers restricted the power and the accuracy of statistical tests. Although teachers had higher scores after the training, many of these effects were not statistically significant. Given the sample size of our study and an Alpha-level of .05, only very large effects (effect size > 1.0) could reach statistical significance (Lipsey, 1990, p. 143). One might interpret this constellation in a way that we performed a strong test on the self-congruence intervention. Such a strong test might not be appropriate for a first and exploratory test of an intervention, therefore we did not adapt our Alpha-level in order to increase power (Lipsey, 1990, p. 171). Doing so, about 40 percent of our time-related tests reached significance. A more effective way to handle the power problem would be to compute and set optimal sample sizes before the experiment (e.g., Anderson, Kelley, & Maxwell, 2017). Third, our measurements were general in nature, but not context-specific or -sensitive to the teacher-, classroom-, or school-settings. For example, in the field of self-efficacy measurement, there are instruments which measure a general self-efficacy, but also domain- or context-specific ones (e.g., Schwoerer, May, Hollensbe, & Mencl, 2005). In respect to measurements for volitional skills, for example, Wenhold, Elbe, and Beckmann (2009) developed a VCQ-sport for an elite sport context, or Elsborg, Wikman, Nielsen, Tolver, and Elbe (2017) developed a context-specific scale that can be used in exercise-based contexts (e.g., physical activities). Fourth, our study has not delivered evidence about whether volitional skills of teachers affected student learning. One might expect that improving teachers' volitional skills should have positive transfer effects on, for example, classroom management and as a consequence on student learning (e.g., Tessier, Sarrazin, & Ntoumanis, 2010).

Implications for Future Research

In summary, our exploratory and applied research study demonstrated that an intervention on teachers' volitional skills that was based on self-congruence might be promising for future research activities. Future theory building needs to explicate the relationship between self-congruence and volitional skills of teachers in more detail and has to transform more systematically assumptions from basic psychological research into applied theoretical concepts of educational intervention research. Also, future research activities should produce information about transfer effects on students learning and replications of intervention effectiveness before a dissemination of the intervention is advisable. Especially, evidence about effectiveness in time (during, at the end, and after the intervention on the long run), in contexts (in similar, nested, and real-life situations), and on different-intervention-functioning (considering side and interaction effects) has to be gathered in future research activities (e.g., Astleitner, Kriegseisen, & Riffert, 2009). In order to avoid typical methodological problems in teacher education field-research, a pre-calculation of sample sizes for optimal power, control groups or other alternatives, the development of a context-specific measurement of volitional skills, and the testing of transfer effects on the learning of students should be considered.

Having more and better evidence for interventions on teachers' volitional skills should lead in the long run to a more comprehensive integration of volitional aspects in all stages of teacher education. The focus on volition is important for beginners in the teaching profession as it helps to prevent cascade or domino-effects resulting in de-motivation, or dropout (e.g., Bembenutty, White, & Vélez, 2015). For more experienced teachers, volitional skills are important for dealing successfully with rapidly changing educational contexts and their complex demands especially in order to cope actively and not passively working stress (e.g., Aelterman, Vansteenkiste, Van Keer, & Haerens, 2016).

Alternative research designs when having problems with control groups. Finally, our study has low internal validity because no control group without an intervention was included. Therefore, in addition to the intervention, additional factors could contribute to the findings. Using our artificial control group reduces this problem, but did not solve it entirely. In the field of teacher education, it is often difficult to get participants in studies, especially for control groups with no beneficial interventions. Of course, one alternative would be to have multiple intervention groups and to use one or some of them as control group for comparison. Another possibility would be to use sophisticated statistical procedures. For example, West et al. (2008) listed several complex statistical approaches for handling threats to internal validity like multilevel, instrumental variable, missing data, sensitivity, or propensity score analysis. Another practicable alternative would be to consider "nonequivalent dependent variables" which are predicted *not* to change due to an intervention but to response on contextually given internal validity threats (Coryn & Hobson, 2011, p. 33). Other probably less practicable alternatives concern conducting stan-

dard-based "single-subject experimental designs" (e.g., Smith, 2012), "interrupted time series analysis" (e.g., Penfold & Zhang, 2013), or combinations of experimental designs with mixed methods approaches (Edmonds & Kennedy, 2013).

Overall, our study should show that, within educational intervention research, a missing control group is an essential problem. However, especially in practical contexts where it is often not possible to have a fully controlled situation, alternatives to missing control groups exist. Our study showed at least one of these alternatives together with hints on further possibilities which are not always in the consciousness of researchers and practitioners within applied educational settings.

References

Aelterman, N., Vansteenkiste, M., Van Keer, H., & Haerens, L. (2016). Changing teachers' beliefs regarding autonomy support and structure: The role of experienced psychological need satisfaction in teacher training. *Psychology of Sport and Exercise*, *23*, 64–72. doi: https://doi.org/10.1016/j.psychsport.2015.10.007

Alsleben, P., & Kuhl, J. (2011). Touching a person's essence: Using implicit motives as personal resources in counseling. In W. M. Cox & E. Klinger (Eds.), *Handbook of motivational counseling: Goal-based approaches to assessment and intervention with addiction and other problems* (2nd ed., pp. 109–129). Chichester: Wiley & Sons. doi: https://doi.org/10.1002/9780470979952.ch5

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*, 1547–1562. doi: https://doi.org/10.1177/0956797617723724

Astleitner, H., Kriegseisen, J., & Riffert, F. (2009, September). *Using a multiple evidence model (MUEMO) for testing the effectiveness of educational interventions*. Paper presented at the European Conference of Educational Research (ECER). Vienna. Retrieved from http://seniorenuniversitaet.at/fileadmin/oracle_file_imports/1091171.PDF

Basit, T. N., Roberts, L., McNamara, O., Carrington, B., Maguire, M., & Woodrow, D. (2006). Did they jump or were they pushed? Reasons why minority ethnic trainees withdraw from initial teacher training courses. *British Educational Research Journal*, *32*, 387–410. doi: https://doi.org/10.1080/01411920600635411

Baumann, N., Chatterjee, M. B., & Hank, P. (2016). Guiding others for their own good: Action orientation is associated with prosocial enactment of the implicit power motive. *Motivation and Emotion*, *40*, 56–68. doi: https://doi.org/10.1007/s11031-015-9511-0

Baumann, N., Kazén, M., & Kuhl, J. (2010). Implicit motives: A look from personality systems interaction theory. In O. C. Schultheiss & J. C. Brunstein (Eds.), *Implicit motives* (pp. 375–403). New York: Oxford University Press. doi: https://doi.org/10.1093/acprof:oso/9780195335156.003.0013

Bembenutty, H., White, M. C., & Vélez, M. R. (Eds.). (2015). *Developing self-regulation of learning and teaching skills among teacher candidates*. Dordrecht: Springer. doi: https://doi.org/10.1007/978-94-017-9950-8

Boekaerts, M., de Koning, E., & Vedder, P. (2006). Goal-directed behavior and contextual factors in the classroom: An innovative approach to the study of multiple goals. *Educational Psychologist*, *41*, 33–51. doi: https://doi.org/10.1207/s15326985ep4101_5

Capa-Aydin, Y., Sungur, S., & Uzuntiryaki, E. (2009). Teacher self-regulation: Examining a multidimensional construct. *Educational Psychology*, *29*, 345–356. doi: https://doi.org/10.1080/01443410902927825

Coryn, C. L., & Hobson, K. A. (2011). Using nonequivalent dependent variables to reduce internal validity threats in quasi-experiments: Rationale, history, and examples from practice. *New Directions for Evaluation, 131,* 31–39. doi: https://doi.org/10.1002/ev.375

De Boer, H., Timmermans, A. C., & Van der Werf, M. P. C. (2018). The effects of teacher expectation interventions on teachers' expectations and student achievement: Narrative review and meta-analysis. *Educational Research and Evaluation*, *24*, 180–200. doi: https://doi.org/10.1080/13803611.2018.1550834

Deimann, M., & Bastiaens, T. (2010). The role of volition in distance education: An exploration of its capacities. *The International Review of Research in Open and Distributed Learning*, *11*, 1–16. doi: http://doi.org/10.19173/irrodl.v11i1.778

Duckworth, A. L., Shulman, E. P., Mastronarde, A. J., Patrick, S. D., Zhang, J., & Druckman, J. (2015). Will not want: Self-control rather than motivation explains the female advantage in report card grades. *Learning and Individual Differences*, *39*, 13–23. doi: https://doi.org/10.1016/j.lindif.2015.02.006

Edmonds, W. A., & Kennedy, T. D. (2013). *An applied reference guide to research designs. Quantitative, qualitative, and mixed methods*. Los Angeles, CA: Sage.

Elsborg, P., Wikman, J. M., Nielsen, G., Tolver, A., & Elbe, A. M. (2017). Development and initial validation of the volition in exercise questionnaire (VEQ). *Measurement in Physical Education and Exercise Science*, *21*, 57–68. doi: https://doi.org/10.1080/1091367X.2016.1251436

Erdogan, T., & Senemoglu, N. (2017). PBL in teacher education: Its effects on achievement and self-regulation. *Higher Education Research & Development*, *36*, 1152–1165. doi: https://doi.org/10.1080/07294360.2017.1303458

Forstmeier, S., & Rüddel, H. (2008). Measuring volitional competences: Psychometric properties of a short form of the Volitional Components Questionnaire (VCQ) in a clinical sample. *The Open Psychology Journal*, *1*, 66–77. doi: https://doi.org/10.2174/1874350100801010066

Funnell, S. C., & Rogers, P. J. (2011). *Purposeful program theory. Effective use of theories of change and logic models*. San Francisco: Jossey-Bass.

Gokce, F. (2010). Assessment of teacher motivation. *School Leadership and Management*, *30*, 487–499. doi: https://doi.org/10.1080/13632434.2010.525228

Hennecke, M., & Freund, A. M. (2017). The development of goals and motivation. In J. Specht (Ed.), *Personality development across the lifespan* (pp. 257–273). London: Academic Press. doi: https://doi.org/10.1016/B978-0-12-804674-6.00016-8

Karabenick, S. A., Richardson, P. W., & Watt, H. M. (Eds.). (2014). *Teacher motivation: Theory and practice*. New York, Abingdon: Routledge. doi: https://doi.org/10.4324/9780203119273

Keller, J. M. (2008). An integrative theory of motivation, volition, and performance. *Technology, Instruction, Cognition, and Learning*, *6*, 79–104. Retrieved from http://www.oldcitypublishing.com/journals/ticl-home/ticl-issue-contents/ticl-volume-6-number-2-2008/ticl-6-2-p-79-104/

Klassen, R. M., & Tze, V. M. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review*, *12*, 59–76. doi: https://doi.org/10.1016/j.edurev.2014.06.001

Kuhl, J. (2000). The volitional basis of Personality Systems Interaction Theory: Applications in learning and treatment contexts. *International Journal of Educational Research*, *33*, 665–703. doi: https://doi.org/10.1016/S0883-0355(00)00045-8

Kuhl, J. (2013). *Auswertungsmanual für den Operanten Multi-Motiv-Test OMT. IMPART-Test-Manuale [Manual for the Operant Multi-Motive-Test]*. Greven: Sonderpunkt Wissenschaftsverlag.

Kuhl, J., & Fuhrmann, A. (2004). *Selbststeuerungs-Inventar: SSI-K3* [Volitional components questionnaire (short form): VCQ-S3]. Osnabrück: University of Osnabrück.

Kuhl, J., Kazén, M., & Koole, S. L. (2006). Putting self-regulation theory into practice: A user's manual. *Applied Psychology*, *55*, 408–418. doi: https://doi.org/10.1111/j.1464-0597.2006.00260.x

Kuhl, J., & Quirin, M. (2011). Seven steps toward freedom and two ways to lose it. *Social Psychology, 42*, 74–84. doi: https://doi.org/10.1027/1864-9335/a000045

Lam, S. F., Cheng, R. W. Y., & Choy, H. C. (2010). School support and teacher motivation to implement project-based learning. *Learning and Instruction*, *20*, 487–497. doi: https://doi.org/10.1016/j.learninstruc.2009.07.003

Lee, J., & Turner, J. (2017). The role of pre-service teachers' perceived instrumentality, goal commitment, and motivation in their self-regulation strategies for learning in teacher education courses. *Asia-Pacific Journal of Teacher Education*, *45*, 213–228. doi: https://doi.org/10.1080/1359866X.2016.1210082

Leutner, D., Fleischer, J., Grünkorn, J., & Klieme, E. (Eds.). (2017). *Competence assessment in education. Research, models, and instruments*. Cham: Springer. doi: https://doi.org/10.1007/978-3-319-50030-0

Lipsey, M. W. (1990). *Design sensitivity. Statistical power for experimental research*. Newbury Park: Sage.

Lise, J., Seitz, S., & Smith, J. (2015). Evaluating search and matching models using experimental data. *IZA Journal of Labor Economics*, *4*, 16. doi: https://doi.org/10.1186/s40172-015-0031-7

Maehr, M. L., & Braskamp, L. A. (1986). *The motivation factor: A theory of personal investment*. Lexington: Lexington Books.

Marsden, E., & Torgerson, C. J. (2012). Single group, pre-and posttest research designs: Some methodological concerns. *Oxford Review of Education*, *38*, 583–616. doi: https://doi.org/10.1080/03054985.2012.731208

Milne, S., Orbell, S., & Sheeran, P. (2002). Combining motivational and volitional interventions to promote exercise participation: Protection motivation theory and implementation intentions. *British Journal of Health Psychology*, *7*, 163–184. doi: https://doi.org/10.1348/135910702169420

Mitchell, M. L., & Jolley, J. M. (2010). *Research design explained* (7th ed.). Belmont: Wadsworth.

Neves de Jesus, S., & Lens, W. (2005). An integrated model for the study of teacher motivation. *Applied Psychology*, *54*, 119–134. doi: https://doi.org/10.1111/j.1464-0597.2005.00199.x

Ortner, T., & van de Vijver, F. J. R. (Eds.). (2015). *Behavior-based assessment in psychology. Going beyond self-report in the personality, affective, motivation, and social domains.* Boston: Hogrefe. doi: https://doi.org/10.1027/00437-000

Parker, P. D., Martin, A. J., Colmar, S., & Liem, G. A. (2012). Teachers' workplace well-being: Exploring a process model of goal orientation, coping behavior, engagement, and burnout. *Teaching and Teacher Education*, *28*, 503–513. doi: https://doi.org/10.1016/j.tate.2012.01.001

Penfold, R. B., & Zhang, F. (2013). Use of interrupted time series analysis in evaluating health care quality improvements. *Academic Pediatrics*, *13*, 38–44. doi: https://doi.org/10.1016/j.acap.2013.08.002

Perels, F., Merget-Kullmann, M., Wende, M., Schmitz, B., & Buchbinder, C. (2009). Improving self-regulated learning of preschool children: Evaluation of training for kindergarten teachers. *British Journal of Educational Psychology*, *79*, 311–327. doi: https://doi.org/10.1348/000709908X322875

Rheinberg, F., & Engeser, S. (2010). Motive training and motivational competence. In O. C. Schultheiss & J. C. Brunstein (Eds.), *Implicit motives* (pp. 510–548). Oxford: University Press. doi: https://doi.org/10.1093/acprof:oso/9780195335156.003.0018

Roness, D. (2011). Still motivated? The motivation for teaching during the second year. *Teaching and Teacher Education, 27*, 628–638. doi: https://doi.org/10.1016/j.tate.2010.10.016

Rupprecht, S., Paulus, P., & Walach, H. (2017). Mind the teachers! The impact of mindfulness training on self-regulation and classroom performance in a sample of German school teachers. *European Journal of Educational Research*, *6*, 565–581. doi: https://doi.org/10.12973/eu-jer.6.4.565

Schiefele, U. (2017). Classroom management and mastery-oriented instruction as mediators of the effects of teacher motivation on student motivation. *Teaching and Teacher Education*, *64*, 115–126. doi: https://doi.org/10.1016/j.tate.2017.02.004

Schüler, J., Brandstätter, V., Wegner, M., & Baumann, N. (2015). Testing the convergent and discriminant validity of three implicit motive measures: PSE, OMT, and MMG. *Motivation and Emotion*, *39*, 839–857. doi: https://doi.org/10.1007/s11031-015-9502-1

Schwoerer, C. E., May, D. R., Hollensbe, E. C., & Mencl, J. (2005). General and specific self-efficacy in the context of a training intervention to enhance performance expectancy. *Human Resource Development Quarterly*, *16*, 111–129. doi: https://doi.org/10.1002/hrdq.1126

Seiberling, C., & Kauffeld, S. (2017). Volition to transfer: Mastering obstacles in training transfer. *Personnel Review*, *46*, 809–823. doi: https://doi.org/10.1108/PR-08-2015-0202

Sheldon, K. M. (2014). Becoming oneself: The central role of self-concordant goal selection. *Personality and Social Psychology Review*, *18*, 349–365. doi: https://doi.org/10.1177/1088868314538549

Sinclair, C. (2008). Initial and changing student teacher motivation and commitment to teaching. *Asia-Pacific Journal of Teacher Education*, *36*, 79–104. doi: https://doi.org/10.1080/13598660801971658

Skinner, E., & Beers, J. (2016). Mindfulness and teachers' coping in the classroom: A developmental model of teacher stress, coping, and everyday resilience. In K. A. Schonert-Reichl & R. W. Roeser (Eds.), *Handbook of mindfulness in education* (pp. 99–118). New York: Springer. doi: https://doi.org/10.1007/978-1-4939-3506-2_7

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*, 510–550. doi: https://doi.org/10.1037/a0029312

Swanson, R. R., & Chermack, T. J. (2013). *Theory building in applied disciplines*. San Francisco: Berrett-Koehler.

Tessier, D., Sarrazin, P., & Ntoumanis, N. (2010). The effect of an intervention to improve newly qualified teachers' interpersonal style, students' motivation and psychological need satisfaction in sport-based physical education. *Contemporary Educational Psychology*, *35*, 242–253. doi: https://doi.org/10.1016/j.cedpsych.2010.05.005

Thoonen, E. E., Sleegers, P. J., Oort, F. J., Peetsma, T. T., & Geijsel, F. P. (2011). How to improve teaching practices: The role of teacher motivation, organizational factors, and leadership practices. *Educational Administration Quarterly*, *47*, 496–536. doi: https://doi.org/10.1177/0013161X11400185

Thrash, T. M., Maruskin, L. A., & Martin, C. C. (2012). Implicit-explicit motive congruence. In R. M. Ryan (Ed.), *The Oxford handbook of human motivation* (pp. 141–156). New York: Oxford University Press. doi: https://doi.org/10.1093/oxfordhb/9780195399820.013.0009

Wagner, L., Baumann, N., & Hank, P. (2016). Enjoying influence on others: Congruently high implicit and explicit power motives are related to teachers' well-being. *Motivation and Emotion*, *40*, 69–81. doi: https://doi.org/10.1007/s11031-015-9516-8

Wang, H., Hall, N. C., & Rahimi, S. (2015). Self-efficacy and causal attributions in teachers: Effects on burnout, job satisfaction, illness, and quitting intentions. *Teaching and Teacher Education*, *47*, 120–130. doi: https://doi.org/10.1016/j.tate.2014.12.005

Watt, H. M., Richardson, P. W., & Smith, K. (Eds.). (2017). *Global perspectives on teacher motivation*. Cambridge: University Press. doi: https://doi.org/10.1017/9781316225202

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*, 89–122. doi: https://doi.org/10.3102/00346543073001089

Wenhold, F., Elbe, A. M., & Beckmann, J. (2009). Testgütekriterien des Fragebogens VKS zur Erfassung volitionaler Komponenten im Sport [Test standards of the ques-

tionnaire VKS for measuring volitional components in sports]. *Zeitschrift für Sportpsychologie*, *16*, 91–103. doi: https://doi.org/10.1026/1612-5010.16.3.91

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., et al. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*, 1359–1366. doi: https://doi.org/10.2105/AJPH.2007.124446

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

4. Negative Evidence: Fostering Pre-Service Teachers' Competences in Social Research and Related Learning Skills – a Quasi-Experimental Study With Minimal Guidance Interventions

Hermann Astleitner, Michaela Katstaller & Ulrike Greiner

Abstract: The current study evaluated the effect patterns of an intervention within a course program for pre-service teachers in educational research and related learning skills for the first time. 89 pre-service teachers participated in a newly developed four-month course. Within this course, the intervention was implemented during six sessions based on a quasi-experimental pre-post-design. In addition to the training sessions, about half of the participants (intervention group) got additional minimal guidance interventions by using learning supporting maps and questions. All participants reported higher monitoring skills at the end of the intervention. Participants with minimal guidance interventions had higher information and methodological literacies. However, course activities decreased critical thinking and mental modeling showing unintended negative evidence. Further analyses revealed that these negative effects might be related to pre-requisites in research and learning skills. Future research is needed to study the role of thinking and model building as well as related standardized measurements.

When planning an educational intervention, then considering that things go wrong, is well established in the head of researchers. Negative evidence is more probable in complex interventions than positive evidence (e.g., Ioannidis, 2005). Why are then scientific journals full of studies with mainly positive and significant findings and why is there a publication bias or a replication problem in experimental social science (e.g., Francis, 2012)? Anyway, within intervention research negative effects should be expected and handled, even when there is a standard-based and high-quality implementation concerning theoretical background, research ethics and design, or measurements. In addition, there is some kind of a revival to deal with negative effects (e.g., Rozental et al., 2018). Rychetnik, Frommer, Hawe, and Shiell (2002) listed numerous criteria for evaluating evidence on interventions and stressed that if an intervention is not successful, the evidence should help to find out whether concept or theory as well as the implementation of the intervention were faulty. Within the following study, we considered negative evidence to be mainly a problem of research design. We did everything up to our best knowledge to avoid

negative results. However, as our selected general issue of teaching is of highest complexity (e.g., Goldman & Kearns, 1995), we are prepared for having not only positive results.

Teaching is a complex activity which requires all available resources from instructors. Effective teaching also represents a skill, which is both sophisticated and difficult to develop, for teacher education activities. This situation becomes even more challenging when research skills are integrated into teacher education (Munthe & Rogne, 2015). Then, pre-service teachers not only have to acquire skills about teaching or instructional design, but also high-level knowledge and skills about literature review, theory building, research design, data collection, statistical analyses, or project management (Hall, 2009). This challenging task increases cognitive and emotional pressure (due, for example, to "statistics anxiety" (Onwuegbuzie & Wilson, 2003)) on pre-service teachers, which requires solid cognitive and emotional competences. However, considering cognitive competences, for example, Eide, Goldhaber, and Brewer (2004, p. 235) reported evidence that there is a "negative relationship between results of measurement of academic performance and the decision, among colleges graduates, to become a teacher", however, with significant variations based on national and cultural conditions (Boeger, 2016). Concerning emotional competences, for example, Corcoran and Tormey (2012) found that pre-service teachers had levels of emotional competences below the norm. Also, for example, Decker and Rimm-Kaufman (2008, p. 58) found that pre-service teachers rated higher than national norms on neuroticism (which "reflects individuals who are nervous and concerned about their ability to succeed in relation to others"). Such evidence might not be representative, but can increase the sensibility for high diversity in pre-service teachers and the necessity for well-dosed support strategies.

So, it is not surprising that there are problems when fostering pre-service teachers' competences in social research. For example, Hiebert, Morris, Berk, and Jansen (2007) identified four research skills for teachers (i.e., specifying goals, conducting empirical observations, constructing hypotheses, and proposing improvements) and argued that it is difficult to implement such skills successfully into teacher education programs as relevant research findings are missing. Niemi (2008) stressed the necessity to establish highly complex multiple interacting factors (i.e., research competence, evidence-based practice, quality of evidence, delivery and access to evidence, an evaluation culture, and collaborative professional networking) for promoting research-based orientations and attitudes in pre-service teachers. She also pointed out that such orientations are related to different and often contradictory paradigms of research (e.g., qualitative, quantitative, or action-research) that produced additional workload and insecurities for inexperienced pre-service teachers. Van der Linden, Bakx, Ros, Beijaard, and Vermeulen (2012, p. 415) found it difficult to describe and explain the development of pre-service teachers' research competences as "they developed their knowledge and skills more in science-orientated topics and less in research methods and research designs". Van der Linden, Bakx, Ros, Beijaard, and van den Bergh (2015) found positive changes in pre-service

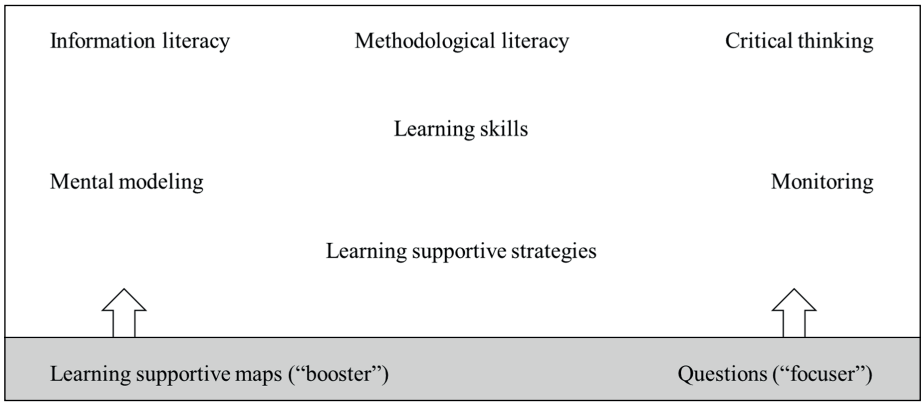

Fig. 1: Fostering competences in educational research.

teachers' knowledge about research during an introductory course. However, their self-efficacy regarding research was related to beliefs which were relatively difficult to develop about abilities to conduct and use research after the course. Haberfellner (2016) found that such beliefs about the usefulness of competences in educational research for pre-service teachers were based on demanding and arduous benefits for graduation work, for daily instruction in schools, for professional career, and for professionalism in the field of teacher education. Finally, Evans, Waring, and Christodoulou (2017) demonstrated that fostering early career teachers' research competences requires the establishment of complex interacting conditions like, for example, research-based collaboration with schools, the use of research-based critical approaches to teaching, the supporting of social-emotional resilience, and the development of metacognitive and self-regulatory skills.

In our study, we responded to such a demanding situation when fostering pre-service teachers' competences in educational research in three ways (see Figure 1).

First, we chose a clear focus and defined such a competence as "ability to understand and apply research-based knowledge" (Groß Ophoff, Schladitz, Lohrmann, & Wirtz, 2014, p. 251) with three essential sub-skills: (1) "Information literacy" as "capacity of people to recognize their information needs, locate and evaluate the quality of information, store and retrieve information, make effective and ethical use of information, and apply information to create and communicate knowledge" (Catts & Lau, 2008, p. 7); (2) "methodological literacy" as the ability to understand and analyze information from descriptive statistical data (Koch, 2011), and (3) "critical thinking" as skill to identify assumptions and to find explanations in theory- and evidence-based research results (Osana & Seymour, 2004). Second, we focused on the process of learning by linking competences to learning supporting strategies based on approaches about "developmental education" (Seel & Hanke, 2015, p. 339) and "constructive alignment" (Biggs, 2014). We considered "mental modeling" (i.e., the construction of cognitive representations on subject areas) and "monitoring"

(i.e., the evaluation and regulation of progress in learning) as important learning skills. For example, Feldon (2010) stressed the important role of mental modeling for scientific problem-solving as it is related to the transfer of information between long-term and working memory, the accuracy of recall, or the encoding of information. In respect to pre-service teachers, for example, Moseley, Desjean-Perrotta, and Utley (2010) analyzed such representations and found incomplete but educationally formable mental models. Monitoring has a long history in the field of teacher education and is about observing one's own performance, confronting it with standards, and deciding about how to improve performance (Rispoli et al., 2017). Monitoring was repeatedly found effective in introductory courses on research methods or on data literacy interventions for pre-service teachers (e.g., Lan, Bradley, & Parr, 1993; Reeves & Honig, 2015). Third, we were aware that fostering pre-service teachers' competences in social research represents a complex and demanding task. As a rule, there is less work-load available in curricula for educational research methods in teacher education than in other fields of social research (e.g., educational science or psychology). Such a situation needs special instructional support for pre-service teachers without overloading or overstraining them. Therefore, we decided to use "minimal guidance interventions" "in form of process- or task-relevant information that is available if learners decide to use it" (Kirschner, Sweller, & Clark, 2006, p. 76). As recent research showed that (sometimes weak) minimal guidance effects can be increased with high levels of practice, we established multiple assignments as core elements of instructional interactions (e.g., Brunstein, Betts, & Anderson, 2009). In addition, we implemented – as experimental condition – learning supporting maps (called "booster") and questions (called "focuser") as instructional support devices. Maps were found effective in learning and represent graphic organizers that deliver visual knowledge on relationships among concepts and processes (Nesbit & Adesope, 2006). Such maps were, for example, successfully used in data-based field activities with pre-service teachers (Francis, 2015). Questions for supporting learning were, for example, an essential element in inquiry-oriented interventions for pre-service teacher's thinking on research (e.g., Gitlin, Barlow, Burbank, Kauchak, & Stevens, 1999) or when designing an experimental activity as inquiry (e.g., Cruz-Guzmán, García-Carmona, & Criado, 2017).

A Course for Pre-Service Teachers on Fostering Competences in Social Research

In order to support pre-services teachers' competences in social research and related learning skills, a course entitled "Theories, concepts, and categories of Educational Science" was developed by the second and the third authors of the study for the first time. Due to high registration numbers, the same course was offered for four different groups of pre-service teachers, each course lasted over the entire duration of the semester. The first course instructor held two courses and had

many years of experience in social research methods and related teaching as well as project work. The second course instructor held the other two courses and had more than two decades of teaching, research, and administration experience in the field of teacher education. The course was part of a study entrance and orientation phase within a bachelor teacher education curriculum at a School-of-Education and had an average work-load of about 50 hours. It lasted about four months and consisted of about 11 one and a half hour sessions. Major course learning outcomes concerned the acquisition and reflection of (1) fundamental concepts of Educational Science in teacher education, (2) declarative knowledge about basic requirements of the teaching profession and (3) basic methods of scientific research in teacher education. From an instructional design perspective, the course was based on a variety of instructional methods consisting of lecture/presentation, Socratic dialogue, demonstration/modeling, guided discussion/debate, and cooperative group learning. Students had to accomplish multiple assignments during or/and after each session. Assignments concerned questions, problem-solving tasks, self-reflections, or group activities. They were based on different educational goal levels ranging from retrieval, comprehension, analysis, knowledge utilization, meta-cognition to self-system thinking (Marzano & Kendall, 2008). For each assignment, students were prepared with theory- and research-based presentations or texts about important research findings. As major course achievement and in order to successfully complete the course, students had to write a research report on a self-chosen topic. Research reports were evaluated on multiple criteria (i.e., paper is free of errors in spelling and grammar; reference list and in-text citations follow citation guidelines; writing style is clear and uses scientific language; concepts are clearly defined; argumentation is based on research findings; statistical data is correctly interpreted; topic's importance and research questions are established conclusively; sections in the paper follow a logical sequence; problems are viewed from multiple and critical perspectives).

Purpose and Hypotheses

Our first objective was to investigate the effect patterns of a course intervention on educational research competences in enhancing pre-service teachers' research and related learning skills. The second objective was to gain insight into the effect patterns of learning supporting maps and questions on these skills. We used a quasi-experimental pre-post-design to examine these questions. In both intervention and control groups, pre-service teachers were confronted with course contents and methods on educational research. In addition, within the intervention group, pre-service teachers received additional learning supporting maps and questions.

With regard to research skills, we first hypothesize that both groups will demonstrate increased information literacy, methodological literacy, and critical thinking in a posttest (in comparison to a pretest). We also hypothesize that the use of learning supporting maps and questions should produce higher research skills within

the intervention group (in comparison to a non-use situation within the control group). We expect the same effects also with regard to learning skills, because within the course many mental modeling supporting elements (e.g., diagrams, tables, or figures) and monitoring assisting activities (e.g., questioning research findings, comparison with criteria and standards, and giving correcting feedback) were established. For both groups of dependent variables, we assume a general positive instructional effect of the course and a specific additional positive instructional effect of the minimal guidance interventions.

The general instructional effect of the course should be achieved by a strong focus on results and methods of school- and teacher-education-research with a demonstrative effect: Pre-service teachers could learn from research papers how to do social research (e.g., Brill & Yarden, 2003). In addition, the focus on school- and teacher-education research had high relevance for pre-service teachers as it is always related to challenges of their professional future. A third effective factor was established by the highly varying instructional methods within the course. Both high relevance and high variation should stimulate motivation and volition for learning and deepen learning experiences. The specific instructional effect of minimal guidance interventions should be achieved by delivering a processual guidance and an assistance in self-control activities (e.g., Astleitner, 2018a).

Positive and negative effects. Above all, we expect that both, the general and the specific instructional effects are positive in nature: They should have a positive resp. increasing effect on research and learning skills. However, we are not quite sure about the effect patterns, because our course was held for the first time and teacher education results were often inconclusive: Even courses for pre-service teachers which were based on a well-balanced instructional design could cause instable, inconsistent, non-significant, or even negative effects (e.g., Cheong, 2010; Lederman, Schwartz, Abd-El-Khalick, & Bell, 2001; Schüle, Besa, Schriek, & Arnold, 2017).

Method

Participants

A total of 89 pre-service teachers from four courses on research methods at a School-of-Education from an Austrian university volunteered to participate in the study. Two courses (one intervention and one control group) were held by the second author of this study, two courses (again one intervention and one control group) by the third author. Participants of these courses were randomly assigned to intervention and control groups: First, participants decided of their own on the participation in one of the groups. Second, the first author of the study decided without knowing the participants which groups were fixed as intervention and control groups. Of the 45 participants in the intervention group, 73% were female. Their ages ranged from 18 to 28 years (M = 21.29, SD = 1.87). Of the 44 pre-service teachers in the control group, 68% were female. These participants were between 18

and 30 years of age (M = 21.31, SD = 2.38). Students within the intervention group got a grade *average* at the end of the courses of 2.84 (SD = 1.09), students within the control group of 2.73 (SD = 1.30) showing, as assumed before the intervention, no ethically questionable disadvantages of the control group. Before the intervention, participants in both groups had to self-evaluate their depth of learning in order to explore relevant group differences. Two items were used: "In today's course session, I learned only superficially, without going into depth" and "In today's course session, I acquired scientifically founded knowledge" on a four-point scale (from "often" (0) to "never" (3)). Both groups had low, but not significantly different values for the first (intervention group: M = 1.29, SD = 1.01; control group: M = 1.20, SD = 0.99; $t(61)$ = 0.34, $p > .05$) and the second evaluation item (intervention group: M = 1.96, SD = 1.00; control group: M = 1.89, SD = 0.93; $t(61)$ = 0.32, $p > .05$). Overall, both groups were comparable with regard to sex, age, and depth of learning (before the intervention).

Design

We employed a quasi-experimental pre-post-design to study the effect patterns of the course and the minimal guidance interventions. The whole experiment was implemented during 6 of the 11 weekly course sessions. It started in the third course session and ended in the eighth session with an overall duration of about five weeks. Within the first session, students in both groups were asked about their depth of learning. The second session started with a pretest for both groups and the first minimal guidance intervention in the intervention group. Within the second and third sessions, minimal guidance was implemented in the intervention group. In the fourth session, the intervention group had minimal guidance and both groups completed the posttest (at the end of the session).

Intervention and control group sessions had the same instructor, the same goals, contents, assignments, time schedules, learning materials (texts on, for example, Hattie, 2009) and instructional methods. Within the intervention group, students had additional minimal guidance instruction. Within the control group, there was no such guidance.

Minimal guidance intervention. Within the intervention group, students received minimal guidance instructions on two sheets of paper as a "booster" and as a "focuser" (see Figure 2).

In the "booster", students were informed on a map that an issue has several components (e.g., concepts) related to arguments, reasons, and evidence. These three elements are depicted as affected by research conditions (e.g., theories, methods for data collection, and study results). Such research conditions are regarded as being influenced by a study of literature (with, for example, questions and related search terms). The booster represents a checklist or guidance that should support the completion of task assignments (e.g., presentations, discussions, or exam preparation) within the course. In the "focuser", students were presented with 10 questions which

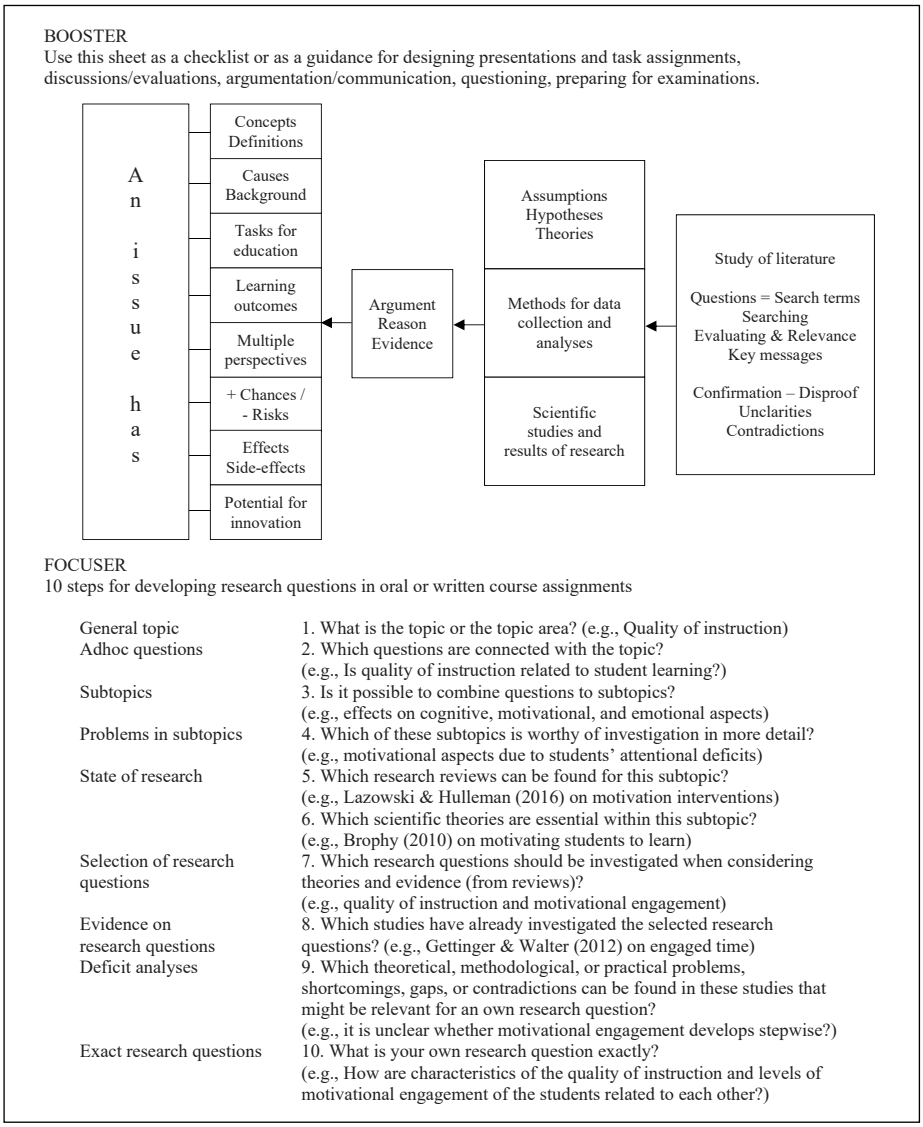

Fig. 2: Minimal guidance interventions.

should support them in developing research questions in oral and written course assignments. The focuser consists of general descriptions (e.g., general topic), related questions (e.g., what is the topic or the topic area?), and examples (e.g., quality of instruction). At the beginning of the intervention, students were informed about the goals of these minimal guidance interventions and about how to use them in order to support their individual learning processes. During all intervention sessions, the instructor asked the students several times to use both, the booster and the focuser, to assist their learning.

Manipulation check and controlling for bias. The training sessions in the intervention and the control group had the same goals, contents, teaching methods and learning materials. For manipulation check, students were asked about their depth of learning and their experienced instructional support in all sessions. Participants had to evaluate in each session the item "In today's course session, I learned how to design scientific questions" (with a four-point scale from "often" to "never"). Answers on this item were summed up for all four intervention sessions (that led to a reduced n due to missing data). Results from a *t-test* showed (nearly) significant higher values in the intervention in comparison to the control group (for the intervention group: $M = 6.94$, $SD = 3.15$; for the control group: $M = 5.28$, $SD = 2.85$; $t(33) = 1.64$, $p = .06$ (one-tailed)). Test results indicate a weak, but consistent manipulation which corresponds with our minimal intervention approach.

As both course instructors instructed one intervention and one control group, experimenter bias (e.g., experimenter might be more supportive to participants who are getting the intervention) must be controlled resp. tested (Mitchell & Jolley, 2010, p. 166). For testing experimenter bias, all participants had to evaluate the item "In today's course session, I was made aware of the different dimensions and aspects of the topics covered" (with a four-point scale from "often" to "never"). Again, answers on this item were summed up for all four intervention sessions. Results from a *t-test* showed no differences between intervention in comparison to the control group (for the intervention group: $M = 8.77$, $SD = 2.81$; for the control group: $M = 9.33$, $SD = 1.65$; $t(33) = -.73$, $p = .24$ (one-tailed)). Test results indicate no experimenter bias throughout the whole period of the intervention.

Measures and Indicators for Reliability and Validity

Within Table 1, descriptive statistics, reliability information and correlations between study measures of the pretest are depicted.

Tab. 1: Descriptive Statistics, Reliability Information, and Correlation between Study Measures (Pretest, 56 < N < 81)

Variables	M	SD	α	1	2	3	4	5
Research competences								
1: Information literacy	9.61	2.08	-	-				
2: Methodological literacy	4.81	1.54	-	.21*	-			
3: Critical thinking	6.60	1.75	-	-.01	.03	-		
Learning skills								
4: Mental modeling	17.66	4.58	.81	-.02	-.03	.07	-	
5: Monitoring	5.91	3.74	.76	-.09	-.07	-.18	.41**	-

Note. α = Cronbach's Alpha, M = mean, SD = standard deviation, *p < .05, **p < .001, one-tailed.

Information literacy was measured by summing up solutions from three curriculum-based tasks. Within the first task, students had to evaluate six research questions whether they meet certain quality criteria (e.g., free of contradictions) or not. Correct answers on all six questions, for example, "Is there a relationship between the stress experiences of teachers and the quality of their instructional behavior in classrooms?" were summed up and ranged from zero to six points with an acceptable difficulty level (with $M = 3.91$, $SD = 1.19$). As the second task, students had to rank five strategies for searching scientific literature (e.g., "Reduce hits by restricting, for example, the year of publication") in an appropriate order (from first to last). Differences between students orders and the optimal order were calculated ($M = 3.46$, $SD = 1.40$). In a third task, students had to decide on three research questions (e.g., "How can classroom videos foster analyzing skills of pre-service teachers?") whether they combine search keywords by using Boolean operators such as "or", "and" or "not" in order to deepen literature search results ($M = 2.25$, $SD = 0.55$). There were low positive *correlations* between the three tasks ($0.07 < r < 0.19$), indicating that they represent common but also separate curriculum-based components of information literacy.

Methodological literacy was measured in a similar way. In a first task, students were confronted with bar charts on three variables (educational level of students based on type of school and parent's education). Based on these charts, students had to evaluate four statements (e.g., "The largest part of adolescents from low or middle educational levels are within apprenticeships") whether they fit the chart information or not ($M = 3.00$, $SD = 0.73$). The second task was similar to the first task but was on a different issue (distribution of students on type and level of schools) with six statements (e.g., "About 45% of students are attending a certain type of secondary school, called Hauptschule"). However, we did not use this task in our analysis, because difficulty levels were to low ($M = 5.13$, $SD = 1.41$) and correlation with other tasks was not positive ($r = -.04$). Within a third task, students had to read an interview text from a qualitative study about conditions for a successful implementation of educational standards in schools (with a length of about 30 lines). Then, students had to evaluate five interpretations (e.g., "Creative things are not testable, and as German language courses are to a large extent based on creative processes, testing educational standards is not appropriate.") whether they fit text passages or not ($M = 1.84$, $SD = 1.20$). Answers on the two selected tasks were correlated ($r = .22$) and summed up for further analysis.

The measurement of critical thinking was also based on three tasks. Within a first task, students had to read an abstract of an empirical study on social participation in inclusive instruction. The text consisted of about 50 lines and represented the basis for five critical statements about theoretical and methodological problems (e.g., "The two basic assumptions are formulated based on theories and related scientific references."). Students had to decide whether these statements are appropriate or not. Correct answers on the five statements were summed up and resulted in acceptable difficulty ($M = 4.47$, $SD = 0.86$). As a second task, students had to

evaluate a short (about four-line) statement about the use of cooperative teaching methods in daily instruction on four dimensions (i.e., type of theory (scientific or everyday theory), precision of concepts (given or not given), empirical verification (personal or with scientific methods), and generalization (for many or few cases)). Again, answers were summed up, showed acceptable difficulty ($M = 2.16$, $SD = 1.45$), and correlated with the first task ($r = .08$). This was not true for an excluded third task on evaluating a text on portfolio-use in the field of teacher education (with five items) which correlated negatively with the second task ($r = -.02$). Again, answers on the two resulting tasks were summed up for further analysis.

Overall, the measurement of research competences was based on few, but complex and curriculum-based tasks with difficulties at a medium level as an indicator of good reliability. Computing internal consistencies as indicators of reliability was not suitable because of the small number of tasks. Correlations between measurements of research competences were low which can be seen as an indicator of discriminating validity (see Table 1). Significant evidence for convergent validity is only given for information and methodological literacy ($r = .21$), but predictive validity indicating correlations between pre- and posttest measurements of all three research competences were significant (r for information literacy = .23, r for methodological literacy = .36, and r for critical thinking = .26, all $p < .05$ (one-tailed)). A further test, which was administered two weeks after the posttest, confirmed significant correlations between the post- and the follow-up-tests and good (predictive) validity (for information literacy: $r = .32$, for methodological literacy: $r = .41$, for critical thinking: $r = .51$) (all $p < .01$ (one-tailed)).

For measuring mental modeling and monitoring, originally 20 items were used based on a five-point-scale (ranging from "never" to "(nearly) always"). Factor and reliability analyses and related criteria led to measurements of seven items for mental modeling and five items for monitoring. Eight items were excluded because of low or ambiguous factor loadings as well as low or negative discrimination indices. Factor analysis indicated good convergent validity based on a stable two-factor solution (mental modeling: $R^2 = .27$, all factor loadings > .52; monitoring: $R^2 = .22$, all factor loadings > .64) and on a significant medium high correlation ($r = .41$) between these two variables (see Table 1). In order to have further predictive validity indicators, the two instruments were administered again two weeks after the posttest. Correlations between the two tests are all high and statistically significant (for mental modeling: $r = .87$; for monitoring: $r = .82$). Reliability analyses of measurements showed good reliability (mental modeling: Cronbach's Alpha = .81; monitoring: Cronbach's Alpha = .76). Mental modeling was measured with items like, for example, "I have tried to integrate the learned with learning experiences from other courses", "I have tried to establish links between different parts of the learning content", or "I worked focused and intensively on the teaching content". For measuring monitoring, items like, for example, "I have assessed myself whether I achieved progress in learning or not", "I have critically evaluated contents by using scientific literature", or "I have made summaries about teaching contents" were used.

Data Analysis

For exploring and testing effect patterns, we analyzed our data using the IBM SPSS 24 General Linear Model (GLM) on a two-factors ANOVA with repeated measures on one factor. Repeated measures analyses were based on factors time (pre- vs. posttest) and group (intervention vs. control group) and their effects on research competences and learning skills. The Alpha-level was set at $p < .05$ and no adaptations were made in order to take the exploratory character of the study into consideration. Partial eta square (η^2_P) served as indicator of effect sizes.

Results

Time and Intervention Effect Patterns

Pre- and posttest means *(M)*, standard deviations *(SD)*, and results of repeated measures analysis of variance on research competences and learning skills are illustrated in Table 2.

Overall and on a descriptive level, mean comparisons on posttests showed no fully consistent effect patterns: Means in the intervention group increased over time for information and methodological literacy, and monitoring, but not for critical thinking and mental modeling. Means in the control group only increased over time for monitoring, but not for all other measures. Means in the posttests were higher in the intervention group in comparison to the control group for information and methodological literacy as well as critical thinking and mental modeling, but not for monitoring.

In detail and based on statistical tests, ANOVAs revealed a significant increase over time for participants in both groups on monitoring ($F = 4.28$, $p = .042$, $\eta^2_P = .06$), a significant decrease on mental modeling ($F = 4.96$, $p = .029$, $\eta^2_P = .07$), and a nearly significant decrease on critical thinking ($F = 3.20$, $p = .079$, $\eta^2_P = .05$). Concerning differences between control and intervention group, analyses revealed a significant increase in information literacy in the intervention group, but a decrease in the control group ($F = 5.32$, $p = .024$, $\eta^2_P = .08$), the same is true for methodological literacy at a level close to significance ($F = 3.32$, $p = .073$, $\eta^2_P = .05$). There was also a nearly significant difference between control and intervention group on critical thinking ($F = 3.29$, $p = .074$, $\eta^2_P = .05$): The intervention produced higher values in the posttest, but higher values in the intervention group were also given at the pretest.

Further Analyses on Negative Results

As results on critical thinking and on mental modeling did not meet hypothetical assumptions, we used variables from pretests in order to explore posttest results (see Table 3).

Tab. 2: Descriptive Statistics and Repeated Measures Analysis of Variance Results for Research and Learning Skills (63 < N < 70)

	Pretest		Posttest		Main effects						Interaction		
					Group (G)			Time (T)			G x T		
	M	(SD)	M	(SD)	F	p	η^2_p	F	p	η^2_p	F	p	η^2_p
Information literacy					0.00	.996	.00	0.06	.815	.00	5.32	.024	.08
CG	10.00	2.04	9.39	2.17									
IG	9.32	1.72	10.07	1.65									
Methodological literacy					1.08	.303	.02	1.98	.164	.03	3.32	.073	.05
CG	5.05	1.58	4.97	1.17									
IG	4.40	1.54	5.03	1.33									
Critical thinking					3.29	.074	.05	3.20	.079	.05	0.52	.474	.01
CG	6.62	1.75	5.97	1.95									
IG	7.03	1.38	6.76	1.53									
Mental modeling					0.94	.336	.01	4.96	.029	.07	0.07	.796	.00
CG	17.42	4.25	16.58	4.43									
IG	18.58	5.26	17.52	5.40									
Monitoring					0.04	.840	.00	4.28	.042	.06	0.66	.421	.01
CG	5.61	3.83	6.61	4.04									
IG	6.06	3.62	6.50	3.43									

Note. CG = control group, IG = intervention group.

Tab. 3: Regression Analyses' Results for Critical Thinking and Mental Modeling (59 < N < 70)

Pretest variables	Critical thinking (Posttest)						Mental modeling (Posttest)					
	Model 1			Model 2			Model 1			Model 2		
	B	SE	p	B	SE	p	B	SE	p	B	SE	p
Information literacy	.05	.13	.735				-.15	.23	.083			
Methodological literacy	-.04	.15	.758				.24	.27	.007	.19	.24	.016
Critical thinking	.27	.15	.044	.31	.13	.011	.06	.24	.492			
Mental modeling	.30	.06	.045	.30	.05	.013	.65	.10	.000	.62	.10	.000
Monitoring	-.00	.07	.991				.23	.13	.020	.24	.12	.010

Note. B = Beta, SE = Standard error, p = p-value.

Within models 1, all five variables from pretests were included in *regression analyses*, within models 2 only variables with significant effects from models 1 remained in model testing. Posttest results on critical thinking were significantly affected by pretest results on critical thinking ($B = .31$, $p = .01$) and on mental modeling ($B = .30$, $p = .01$). Pretest results on mental modeling ($B = .62$, $p = .00$), monitoring ($B = .24$, $p = .01$), and methodological literacy ($B = .19$, $p = .01$) affected significantly posttest results on mental modeling.

Discussions and Implications

In this study, we explored whether participation in a course on research issues and methods together with minimal instructional guidance would improve pre-service teachers' research competences and related learning skills. Our results were mixed, but in line with research findings on limited learning on college campuses (e.g., Arum & Roska, 2011). In respect to information and methodological literacy, we found an increase of competences at the end of the intervention but only when students had additional minimal guidance during the learning process. The intervention also increased the monitoring skills of students (in both groups). However, critical thinking and mental modeling were decreased within the intervention. Deeper analyses of these negative results revealed that students' pretest results or previous competences had a strong effect on both critical thinking and mental modeling. These findings show that both an instructionally well-designed introductory course and minimal guidance interventions can positively affect at least basic research and learning skills of pre-service teachers, like information and methodological literacy as well as monitoring. However, our course and minimal guidance interventions were not effective or even counterproductive in the case of more advanced skills like critical thinking and mental modeling.

Focusing on negative evidence. Fostering pre-service teachers' skills in critical thinking seems to be a difficult task. For example, Temel (2014) did not find effects of traditional teaching methods and problem-based learning on critical thinking dispositions. On the other hand, for example, Kong (2001) found positive effects of a "thinking module" on pre-service teachers' critical thinking dispositions. Also, Cartwright and Noone (2006) suggested establishing "critical imagination" as teaching strategy for fostering critical thinking of pre-service teachers. It engages students in thinking about how things are and how they may be. Such findings in research and our results from further analyses suggest establishing courses on (critical) thinking for pre-service teachers before they start courses on research methods. In addition, teacher educators should know more about the capabilities in (critical) thinking of ongoing pre-service teachers. Therefore, admission tests or other assessments for teacher education study programs should include measurements on critical thinking or, more general, on analytic or reflective thinking skills (e.g., Sternberg, 2017, 2018). However, measuring alone might not be sufficient. Courses on research methods might be designed as a "critical thinking approach" (Jackson, 2015), integrate strategies for "teaching thinking" (Swartz & Perkins, 2016), or be part of whole study programs within an "integrated critical thinking framework" (Dwyer, Hogan, & Stewart, 2014).

Fostering pre-service teachers' skills in mental modeling seems to be equally difficult. For example, Ogan-Bekiroglu (2007) found it hard to change pre-service teachers' flawed or incomplete mental models even with sophisticated methods of "model-based teaching". Wheeldon (2012) found that pre-service teachers did not use models and modeling in their arguments when explaining scientific phe-

nomena. However, fostering mental modeling can be successful when considering multiple conditions: For example, Oh and Oh (2011) listed important topics from research findings which have to be addressed in mental modeling like the meanings of models, purposes of modelling, multiplicity and change of scientific models as well as the use of models in educational settings. Or, Hennissen, Beckers, and Moerkerke (2017) found evidence that pre-service teachers' cognitive schemata had grown from a specific curriculum program in which a strong linkage between theory and practice with sufficient, useful and enforceable assignments and suggestions was realized. In fact, it seems that mental modeling is about theory building that prompts teachers to act as explorers and generators of hypotheses (e.g., Clement, 2000; Cole, 1989). For stimulating pre-service teachers' mental modeling and in order to improve the effectiveness of courses on research methods, theory building methods should be integrated (e.g., Astleitner, 2011, 2018b).

Further limitations. In addition, this study was limited in several ways. First, the relatively small sample size of pre-service teachers might have restrictions on the power of statistical tests, although we also considered larger Alpha-levels for significance testing. However, as in teacher education practice, sample sizes are often limited, other factors to increase design sensitivity and statistical power might additionally be used in future research activities. Such factors concern, for example, high dosage and integrity of treatment, measurement of more proximal effects rather than more distal ones, or statistical variance control with ANCOVA and similar procedures (Lipsey, 1990, p. 171). Second, there were limitations on the design of the study. Internal validity could have been questioned due to missing randomization. We performed a manipulation check and tested experimenter bias and found no problematic differences on our pretests between the different groups. However, we had no tests on self-selection bias and resulting different levels of motivation or work-load between groups. In addition, our instructional manipulation check measured more or less general effects of the intervention, but not whether participants read or used our instructions sufficiently (Oppenheimer, Meyvis, & Davidenko, 2009, p. 867). In future studies, alternatives to randomized controlled trials could be used to mitigate threats to internal validity, like, for example, switching replications (in which an intervention is introduced at different points of time), nonequivalent dependent variables, or multiple control groups (West et al., 2008). Third, more sensitive measurements could have been used. We focused and discussed in detail the reliability and validity of measurements, but we had low correlations between research competences and their items for measuring them. We formulated our items based on a given curriculum, but in future research, such measurements should be expanded and validated by standardized tests. Such tests are available for all research competences in our study, so, for example, an information literacy test (Cameron, Wise, & Lottridge, 2007), a research methods skills assessment (Smith & Smith, 2018), and a critical thinking test for undergraduate students (Hyytinen, Nissinen, Ursin, Toom, & Lindblom-Ylänne, 2015).

In summary, our study represents preliminary evidence that a pre-service teacher course on research methods could succeed in improving basic research and related learning skills. In the study, we combined a focus on competences with related learning skills as suggested in modern principles of competence development. However, we were not successful in relation to advanced or higher-order goal areas (like critical thinking and mental modeling). Further investigation is needed on our negative results and on how to achieve such goals by intensifying thinking and model building together with validation attempts on standardized measurement approaches.

Handling negative results in intervention research. Of course, there are multiple theoretical and methodological reasons why negative results could occur in intervention research. The main reason is that given standards on generating hypotheses and theories, on measuring and manipulating variables, on experimental designs, or writing research proposals were not met (e.g., Mitchell & Jolley, 2010). It is also needless to say that scientific progress is fundamentally based on considering faults and related implications from former research activities. In this case, some kind of sequential testing in series of interventions has taken place. The next interventions should also use experiences from the earlier interventions. The former interventions produce cumulative knowledge and modifications in theory, design, and measurements on the subject in focus. However, these modifications are often more or less conclusive and consistent. Obviously, it is necessary to produce systematically a cumulative science of behavior change within a context of sequential interventions (e.g., Lanovaz et al., 2014).

An important problem in sequential testing concerns the type of theoretical approach. Within such an approach, there must be a sequentially organized developmental model in combination with support strategies for different developmental levels (see chapter 1 in this book). Such a model can then be tested within sequences of (quasi-)experimental tests together with replications and exploratory options. Having then multiple positive and negative evidence, the theoretical model can be adapted. Especially in respect to negative evidence, there is a need to focus research on a step-by-step exploration of reasons for such findings.

Generally, Barlow (2010) pointed out, that sometimes in psychological treatments, people in the experimental group showed greater improvement, but also greater deterioration compared to the control group. Having such evidence, it is, in general, necessary to know more about what intervention is effective for what participant under what contexts. In addition, there should be case study reports of negative effects or single-case experiments immediately after an intervention focusing on an idiographic and nomothetic balance in interpreting effects of interventions. Traditionally, such suggestions concern two lines of intervention research, one on aptitude-treatment-interaction (e.g., Preacher & Sterba, 2019) and one on mixed method research (Edmonds & Kennedy, 2013).

More specifically, Bystedt, Rozental, Andersson, Boettcher, and Carlbring (2014) stressed the necessity to analyze negative effects of interventions on three core

themes: Characteristics of negative effects, causal factors, and approaches on evaluating negative effects. Negative effects might concern short- or long-term effects, no treatment effects, deterioration, or side effects. Causal factors could be associated with inadequately applied methods, potentially harmful interventions, insufficient intervention alliance, failed ethical or professional conduct, discontinuity, or external factors. Comprehensive approaches on evaluating negative effects (with criteria and methods) in social research are still missing. However, many of them might be related to traditional sources of invalidity (like selection, testing, or instrumentation; e.g., Mitchell & Jolley, 2010) and their handling with different experimental designs. Others concern typical standards for intervention research on efficacy, effectiveness, and dissemination (e.g., Flay et al., 2005).

Overall, it seems necessary to do more research on how scientists and all people learn from negative experiences or mistakes. A first helpful step might be to focus in future research activities on error management (see chapter 2 in this book). Frese and Keith (2015) presented a comprehensive approach on error prevention (e.g., zero error tolerance), error management (e.g., acceptance of human error), and related processes (e.g., routines to deal with errors). Within future research activities, such an approach could be applied on intervention research activities and establishing a long-term research perspective on bias in intervention research. Such a perspective could also bring more acceptance of negative findings and more trust into intervention research (e.g., Gorman, 2018).

References

Arum, R., & Roska, J. (2011). *Academically adrift: Limited learning on college campuses.* Chicago: University of Chicago Press. doi: https://doi.org/10.7208/chicago/9780226028576.001.0001

Astleitner, H. (2011). *Theorieentwicklung für SozialwissenschaftlerInnen* [Theory building for social scientists]. Wien: Böhlau, UTB. https://doi.org/10.36198/9783838534619

Astleitner, H. (2018a). Multidimensional engagement in learning – An integrated instructional design approach. *Journal of Instructional Research, 7*, 6–32. doi: https://doi.org/10.9743/JIR.2018.1

Astleitner, H. (2018b). *Spezielle Verfahren sozialwissenschaftlicher Theorieentwicklung* [Special methods of theory building in the social sciences]. Weinheim: Beltz Juventa.

Barlow, D. H. (2010). Negative effects from psychological treatments: A perspective. *American Psychologist*, *65*, 13–20. doi: https://doi.org/10.1037/a0015643

Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of Higher Education*, *1*, 5–22. Retrieved from http://www.herdsa.org.au/herdsa-review-higher-education-vol-1/5-22

Boeger, A. (Ed.). (2016). *Eignung für den Lehrberuf. Auswahl und Förderung [Suitability for the teaching profession. Selection and promotion]*. Wiesbaden: Springer VS. doi: https://doi.org/10.1007/978-3-658-10041-4

Brill, G., & Yarden, A. (2003). Learning biology through research papers: A stimulus for question-asking by high-school students. *Cell Biology Education*, *2*, 266–274. doi: https://doi.org/10.1187/cbe.02-12-0062

Brunstein, A., Betts, S., & Anderson, J. R. (2009). Practice enables successful learning under minimal guidance. *Journal of Educational Psychology*, *101*, 790–802. doi: http://dx.doi.org/10.1037/a0016656

Bystedt, S., Rozental, A., Andersson, G., Boettcher, J., & Carlbring, P. (2014). Clinicians' perspectives on negative effects of psychological treatments. *Cognitive Behaviour Therapy*, *43*, 319–331. doi: https://doi.org/10.1080/16506073.2014.939593

Cameron, L., Wise, S. L., & Lottridge, S. M. (2007). The development and validation of the information literacy test. *College & Research Libraries*, *68*, 229–237. doi: https://doi.org/10.5860/crl.68.3.229

Cartwright, P., & Noone, L. (2006). Critical imagination: A pedagogy for engaging pre-service teachers in the university classroom. *College Quarterly*, *9*(4). Retrieved from https://files.eric.ed.gov/fulltext/EJ835430.pdf

Catts, R., & Lau, J. (2008). *Towards information literacy indicators: Conceptual framework paper.* Retrieved from https://dspace.stir.ac.uk/bitstream/1893/2119/1/cattsandlau.pdf

Cheong, D. (2010). The effects of practice teaching sessions in second life on the change in pre-service teachers' teaching efficacy. *Computers & Education*, *55*, 868–880. doi: https://doi.org/10.1016/j.compedu.2010.03.018

Clement, J. (2000). Model based learning as a key research area for science education. *International Journal of Science Education*, *22*, 1041–1053. doi: https://doi.org/10.1080/095006900416901

Cole, A. L. (1989). Researcher and teacher: Partners in theory building. *Journal of Education for Teaching*, *15*, 225–237. doi: https://doi.org/10.1080/0260747890150304

Corcoran, R. P., & Tormey, R. (2012). How emotionally intelligent are pre-service teachers? *Teaching and Teacher Education*, *28*, 750–759. doi: https://doi.org/10.1016/j.tate.2012.02.007

Cruz-Guzmán, M., García-Carmona, A., & Criado, A. M. (2017). An analysis of the questions proposed by elementary pre-service teachers when designing experimental activities as inquiry. *International Journal of Science Education*, *39*, 1755–1774. doi: https://doi.org/10.1080/09500693.2017.1351649

Decker, L. E., & Rimm-Kaufman, S. E. (2008). Personality characteristics and teacher beliefs among pre-service teachers. *Teacher Education Quarterly*, *35*, 45–64. http://www.jstor.org/stable/23479223

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity*, *12*, 43–52. doi: https://doi.org/10.1016/j.tsc.2013.12.004

Edmonds, W. A., & Kennedy, T. D. (2013). *An applied reference guide to research designs. Quantitative, qualitative, and mixed methods*. Los Angeles, CA: Sage.

Eide, E., Goldhaber, D., & Brewer, D. (2004). The teacher labour market and teacher quality. *Oxford Review of Economic Policy*, *20*, 230–244. doi: https://doi.org/10.1093/oxrep/grh013

Evans, C., Waring, M., & Christodoulou, A. (2017). Building teachers' research literacy: Integrating practice and research. *Research Papers in Education, 32*, 403–423. doi: https://doi.org/10.1080/02671522.2017.1322357

Feldon, D. F. (2010). Do psychology researchers tell it like it is? A microgenetic analysis of research strategies and self-report accuracy along a continuum of expertise. *Instructional Science*, *38*, 395–415. doi: https://doi.org/10.1007/s11251-008-9085-2

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., Mościck, E. K., Schinke, S., Valentine, J. C., & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175. doi: https://doi.org/10.1007/s11121-005-5553-y

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*, 975–991. doi: https://doi.org/10.3758/s13423-012-0322-y

Francis, R. W. (2015). Demonstrating pre-service teacher learning through engagement in global field experiences. *Universal Journal of Educational Research*, *3*, 787–792. doi: https://doi.org/10.13189/ujer.2015.031102

Frese, M., & Keith, N. (2015). Action errors, error management, and learning in organizations. *Annual Review of Psychology*, *66*, 661–687. doi: https://doi.org/10.1146/annurev-psych-010814-015205

Gitlin, A., Barlow, L., Burbank, M. D., Kauchak, D., & Stevens, T. (1999). Pre-service teachers' thinking on research: Implications for inquiry oriented teacher education. *Teaching and Teacher Education*, *15*, 753–769. doi: https://doi.org/10.1016/S0742-051X(99)00015-3

Goldman, S. A., & Kearns, M. J. (1995). On the complexity of teaching. *Journal of Computer and System Sciences*, *50*, 20–31. doi: https://doi.org/10.1006/jcss.1995.1003

Gorman, D. M. (2018). Can we trust positive findings of intervention research? The role of conflict of interest. *Prevention Science*, *19*, 295–305. doi: https://doi.org/10.1007/s11121-016-0648-1

Groß Ophoff, J., Schladitz, S., Lohrmann, K., & Wirtz, M. (2014). Evidenzorientierung in bildungswissenschaftlichen Studiengängen: Entwicklung eines Strukturmodells zur Forschungskompetenz [Evidence-orientation in educational science programs: Development of a structural model of research competence]. In W. Bos, K. Drossel & R. Strietholt (Eds.), *Empirische Bildungsforschung und evidenzbasierte Reformen im Bildungswesen* (pp. 251–276). Münster: Waxmann.

Haberfellner, C. (2016). *Der Nutzen von Forschungskompetenz im Lehramt* [The benefit of research competences in the teaching profession]. Bad Heilbrunn: Julius Klinkhardt.

Hall, E. (2009). Engaging in and engaging with research: Teacher inquiry and development. *Teachers and Teaching: Theory and Practice*, *15*, 669–681. doi: https://doi.org/10.1080/13540600903356985

Hattie, J. A. C. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. Abingdon, New York: Routledge. doi: https://doi.org/10.4324/9780203887332

Hennissen, P., Beckers, H., & Moerkerke, G. (2017). Linking practice to theory in teacher education: A growth in cognitive structures. *Teaching and Teacher Education*, *63*, 314–325. doi: https://doi.org/10.1016/j.tate.2017.01.008

Hiebert, J., Morris, A. K., Berk, D., & Jansen, A. (2007). Preparing teachers to learn from teaching. *Journal of Teacher Education*, *58*, 47–61. doi: https://doi.org/10.1177/0022487106295726

Hyytinen, H., Nissinen, K., Ursin, J., Toom, A., & Lindblom-Ylänne, S. (2015). Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students. *Studies in Educational Evaluation*, *44*, 1–8. doi: https://doi.org/10.1016/j.stueduc.2014.11.001

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi: https://doi.org/10.1371/journal.pmed.0020124

Jackson, S. L. (2015). *Research methods and statistics. A critical thinking approach* (5th ed.). Boston: Cengage Learning.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*, 75–86. doi: https://doi.org/10.1207/s15326985ep4102_1

Koch, U. (2011). *Verstehen Lehrkräfte Rückmeldungen aus Vergleichsarbeiten? Datenkompetenz von Lehrkräften und die Nutzung von Ergebnisrückmeldungen aus Vergleichsarbeiten* [Do teachers understand feedback from comparison work? Data literacy of teachers and the use of results from comparative work]. Münster: Waxmann.

Kong, S. L. (2001, December). *Critical thinking dispositions of pre-service teachers in Singapore: A preliminary investigation*. Paper presented at the meeting of the Australian Association for Research in Education, Fremantle, Australia. Retrieved from https://repository.nie.edu.sg/bitstream/10497/11504/1/AARE-2001-KongSL.pdf

Lan, W. Y., Bradley, L., & Parr, G. (1993). The effects of a self-monitoring process on college students' learning in an introductory statistics course. *The Journal of Experimental Education*, *62*, 26–40. doi: https://doi.org/10.1080/00220973.1993.9943829

Lanovaz, M. J., Rapp, J. T., Maciw, I., Prégent-Pelletier, É., Dorion, C., Ferguson, S., & Saade, S. (2014). Effects of multiple interventions for reducing vocal stereotypy: Developing a sequential intervention model. *Research in Autism Spectrum Disorders*, *8*, 529–545. doi: http://dx.doi.org/10.1016/j.rasd.2014.01.009

Lederman, N. G., Schwartz, R. S., Abd-El-Khalick, F., & Bell, R. L. (2001). Pre-service teachers' understanding and teaching of nature of science: An intervention study. *Canadian Journal of Math, Science & Technology Education*, *1*, 135–160. doi: https://doi.org/10.1080/14926150109556458

Lipsey, M. W. (1990). *Design sensitivity. Statistical power for experimental research*. Newbury Park: Sage.

Marzano, R. J., & Kendall, J. S. (2008). *Designing and assessing educational objectives. Applying the new taxonomy*. Thousand Oaks: Corwin Press.

Mitchell, M. L., & Jolley, J. M. (2010). *Research design explained* (7th ed.). Belmont: Wadsworth.

Moseley, C., Desjean-Perrotta, B., & Utley, J. (2010). The draw-an-environment test rubric (DAET-R): Exploring pre-service teachers' mental models of the environment. *Environmental Education Research*, *16*, 189–208. doi: https://doi.org/10.1080/13504620903548674

Munthe, E., & Rogne, M. (2015). Research based teacher education. *Teaching and Teacher Education*, *46*, 17–24. doi: https://doi.org/10.1016/j.tate.2014.10.006

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, *76*, 413–448. doi: https://doi.org/10.3102/00346543076003413

Niemi, H. (2008). Advancing research into and during teacher education. In: B. Hudson & P. Zgaga (Eds.), *Teacher education policy in Europe: A voice of higher education institutions* (pp. 183–208). Umeå: University of Umeå.

Ogan-Bekiroglu, F. (2007). Effects of model-based teaching on pre-service physics teachers' conceptions of the moon, moon phases, and other lunar phenomena. *International Journal of Science Education*, *29*, 555–593. doi: https://doi.org/10.1080/09500690600718104

Oh, P. S., & Oh, S. J. (2011). What teachers of science need to know about models: An overview. *International Journal of Science Education*, *33*, 1109–1130. doi: https://doi.org/10.1080/09500693.2010.502191

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments – A comprehensive review of the literature. *Teaching in Higher Education*, *8*, 195–209. doi: https://doi.org/10.1080/1356251032000052447

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872. doi: https://doi.org/10.1016/j.jesp.2009.03.009

Osana, H. P., & Seymour, J. R. (2004). Critical thinking in preservice teachers: A rubric for evaluating argumentation and statistical reasoning. *Educational Research and Evaluation*, *10*, 473–498. doi: https://doi.org/10.1080/13803610512331383529

Preacher, K. J., & Sterba, S. K. (2019). Aptitude-by-treatment interactions in research on educational interventions. *Exceptional Children*, *85*, 248–264. doi: https://doi.org/10.1177/0014402918802803

Reeves, T. D., & Honig, S. L. (2015). A classroom data literacy intervention for pre-service teachers. *Teaching and Teacher Education*, *50*, 90–101. doi: https://doi.org/10.1016/j.tate.2015.05.007

Rispoli, M., Zaini, S., Mason, R., Brodhead, M., Burke, M. D., & Gregori, E. (2017). A systematic review of teacher self-monitoring on implementation of behavioral practices. *Teaching and Teacher Education*, *63*, 58–72. doi: https://doi.org/10.1016/j.tate.2016.12.007

Rozental, A., Castonguay, L., Dimidjian, S., Lambert, M., Shafran, R., Andersson, G., & Carlbring, P. (2018). Negative effects in psychotherapy: Commentary and recommendations for future research and clinical practice. *BJPsych open*, *4*, 307–312. doi: https://doi.org/10.1192/bjo.2018.42

Rychetnik, L., Frommer, M., Hawe, P., & Shiell, A. (2002). Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology & Community Health*, *56*, 119–127. doi: http://dx.doi.org/10.1136/jech.56.2.119

Schüle, C., Besa, K. S., Schriek, J., & Arnold, K. H. (2017). Die Veränderung der Lehrerselbstwirksamkeitsüberzeugung in Schulpraktika [The development of student teacher self-efficacy in student teaching field experiences]. *Zeitschrift für Bildungsforschung*, *7*, 23–40. doi: https://doi.org/10.1007/s35834-016-0177-9

Seel, N. H., & Hanke, U. (2015). Erziehung und Persönlichkeit: Personalisation und Individuation [Education and personality: Personalisation and individuation]. In N. M. Seel & U. Hanke, *Erziehungswissenschaft* (pp. 307–480). Berlin: Springer. doi: https://doi.org/10.1007/978-3-642-55206-9_3

Smith, T., & Smith, S. (2018). Reliability and validity of the research methods skills assessment. *International Journal of Teaching and Learning in Higher Education*, *30*, 80–90. Retrieved from https://files.eric.ed.gov/fulltext/EJ1169831.pdf

Sternberg, R. J. (2017). Context-sensitive cognitive and educational testing. *Educational Psychology Review, 30,* 857–884. doi: https://doi.org/10.1007/s10648-017-9428-0

Sternberg, R. J. (2018). Direct measurement of scientific giftedness. *Roeper Review*, *40*, 78–85. doi: https://doi.org/10.1080/02783193.2018.1434715

Swartz, R. J., & Perkins, D. N. (2016). *Teaching thinking. Issues and approaches*. Abingdon, New York: Routledge. doi: https://doi.org/10.4324/9781315626468

Temel, S. (2014). The effects of problem-based learning on pre-service teachers' critical thinking dispositions and perceptions of problem-solving ability. *South African Journal of Education*, *34*, 1–20. doi: https://doi.org/10.15700/201412120936

Van der Linden, W., Bakx, A., Ros, A., Beijaard, D., & van den Bergh, L. (2015). The development of student teachers' research knowledge, beliefs and attitude. *Journal of Education for Teaching*, *41*, 4–18. doi: https://doi.org/10.1080/02607476.2014.992631

Van der Linden, W., Bakx, A., Ros, A., Beijaard, D., & Vermeulen, M. (2012). Student teachers' development of a positive attitude towards research and research knowledge and skills. *European Journal of Teacher Education*, *35*, 401–419. doi: https://doi.org/10.1080/02619768.2011.643401

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., & Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*, 1359–1366. doi: https://doi.org/10.2105/AJPH.2007.124446

Wheeldon, R. (2012). Examining pre-service teachers' use of atomic models in explaining subsequent Ionisation Energy values. *Journal of Science Education and Technology*, *21*, 403–422. doi: https://doi.org/10.1007/s10956-011-9333-0

PART 3.
Measurement Problems

5. Handling Validity Problems in Developmental Measurement Approaches – a Confirmatory Factor Analysis Approach on Student Engagement

Hermann Astleitner

Abstract: Effects patterns in intervention studies depend significantly on the quality of measurements. Important aspects of measurements concern the validity and related approaches on identifying construct validity of dependent variables in educational intervention settings. Problems with construct validity arises especially when multidimensional measurement approaches are used. In this study, confirmatory factor analysis was applied to explore the constructs and different aspects of reliability and validity of a multidimensional and developmental cognitive, motivational, and social-emotional measurement on student engagement in learning. For data collection, 218 undergraduate and graduate students used a 15-item scale as self-assessment instrument. Confirmatory factor analysis showed that a reduced short scale best fitted the given data. Results were discussed in respect to innovative validation issues in educational intervention research like a level-based developmental theoretical perspective and the criteria of validity for change or responsiveness.

Martin (2008) formulated as one essential standard when designing intervention studies, that such studies should use multidimensional instruments to assess diverse dimensions of educational constructs varying, for example, on cognitive and behavioral facets at the same time. A multidimensional perspective represents an important progress within scientific discourses as it allows to get a more complete and therefore more valid picture of a phenomenon. In addition, it stimulates to find, summarize, organize, and integrate scattered research findings into an organized set of relationships resp. a unified theory. It also represents a multi-perspective view on theoretical constructs and related research what is in line with general standards of social research like triangulation (as a combination of multiple research methods in the study of the same construct), or convergent or discriminant validation strategies (Mitchell & Jolley, 2010, p. 164). However, designing and using multidimensional measurement in educational intervention research is not a trivial task. It requires a sophisticated multidimensional theoretical model as well as advanced procedures for measurement development and validity testing.

Theoretical foundation. A first major problem of multidimensional measurement concerns construct validation in general, that is "identifying a construct, defining

it, developing a theory about the structure of the construct (e.g., how many factors are present, how they are related), selecting a means of measuring the construct (e.g., Likert-type scales), and establishing that the measure appropriately represents the construct" (Flake, Pek, & Hehman, 2017, p. 370). In particular, problems in construct validation are about

- the correspondence of theoretical assumptions on dimensions with measurement results (are theoretical dimensions supported by empirical evidence?),
- the uniqueness and independence of dimensions (are the dimensions different from each other?),
- the saturation of dimensions (are all relevant dimensions identified?), or
- the hierarchical status of dimensions (are there higher order dimensions (or factors)?).

Within educational intervention contexts, such multidimensional measurements approaches can, for example, be found on cognition and learning with co-construction of meaning, exploration of different perspectives, error analysis, error communication, reflection on processes and outcomes, feedback seeking behavior, and experimenting (Savelsbergh, van der Heijden, & Poell, 2009). In respect to motivation, multiple dimensions were considered like task motivation, effort, competition, social power, affiliation, social concern, praise, and token (McInerney & Ali, 2006). From an emotional perspective, there are, for example, multidimensional approaches on emotion processing with nonacceptance of emotional responses, difficulties in engaging in goal-directed behavior, impulse control difficulties, lack of emotional awareness, limited access to emotion regulation strategies, and lack of emotional clarity (Gratz & Roemer, 2004).

Developmental perspective. A second significant complication within multidimensional measurements is given in case of a developmental perspective in which developmental steps are organized in a hierarchical order. On the one hand, developmental hierarchies can be found in taxonomies of cognition and learning (e.g., Marzano & Kendall, 2008), on changing motivation from extrinsic to intrinsic (e.g., Ryan & Deci, 2000), on social-emotional aspects like moral reasoning (e.g., Carpendale, 2000), or intercultural sensitivity (e.g., Hammer, Bennett, & Wiseman, 2003). On the other hand, such hierarchically organized development processes were criticized on whether they really exist, how they can be measured, what features define a developmental sequence or what processes underlie developmental change (Lourenço, 2016). Also, Astleitner (2018a, p. 125) identified open questions in existing developmental approaches concerning a) constituent factors of developmental stages (accumulation, networking, deepening, integration, or conversion), b) support mechanisms (changes in reality, norms, information or data, research, training, or coaching) on stimulating developmental changes, c) degrees of accomplishment (fully vs. partially, single vs. combined) of developmental levels, and d)

types of (linear, step-based, delayed, u-related) relationships between developmental factors and stages.

Length of measurement instrument. A third important problem within multidimensional measurements is about how many items are used for measurement. Traditionally within classical test theory, designing measurement instruments is based on the assumption that the larger the number of items, the smaller the amount of error (Drost, 2011, p. 112). Within modern test theory approaches, it is possible to create scales with high precision irrespective of the number of items (e.g., Fraley, Waller, & Brennan, 2000). More or less irrespective of approaches on test theory, single- and multi-item measures were used for even highly complex multidimensional constructs in the past. For example, Littman, White, Satia, Bowen, and Kristal (2006) concluded that two single-item measures for stress were reliable and with validity similar to longer questionnaires. Bergkvist and Rossiter (2007) found no difference in the predictive validity of multi-item in comparison to single-item measures. Postmes, Haslam, and Jans (2013) argued that even a complex construct like social identification can adequately be operationalized with a single item. Laborde, Allen, and Guillén (2016) found a short version of an emotional intelligence questionnaire as a viable alternative to the long version of this instrument.

Goals of the Study

Based on the background of these problems in multidimensional measurement, the general goal of the study is to develop and test a multidimensional measurement instrument in the context of construct validation and related confirmatory factor analysis (e.g., LaNasa, Cabrera, & Trangsrud, 2009). The development of the measurement instrument was based on a theoretical approach on multidimensional and (hierarchically organized) developmental cognitive, motivational, and emotional engagement in learning (Astleitner, 2018b), whereas engagement is about an active involvement and participation in learning activities. Within this approach, it was assumed that cognitive aspects of engagement range hierarchically organized as levels from knowledge, comprehension, convergent thinking, evaluation, to synthesis. Motivational engagements concern hierarchical organized developmental levels like attention, relevance, interest, identification, and intrinsic motivation. Social-emotional engagement is based on the levels of self-assertion, entertainment, belongingness, adaptiveness, and security. For each of these multidimensional levels, there were also related instructional support strategies as developmental conditions which were not considered, as the first important general goal of the study was to explore the validity of a multidimensional measurement instrument and related developmental levels (as dependent variables). Having such a valid instrument would make it possible to test in a next step the validity of developmental conditions (as independent variables).

A first aim of the study was to explore the underlying factors (i.e., theoretical constructs) of the given multidimensional measurements by using confirmatory

factor analysis. Confirmatory factor analysis represents a type of structural equation modeling with observed and latent variables. It is hypothesis-driven, and was used intensively in the psychometric evaluation of test instruments in the past (Brown, 2006). For the three factors of cognitive, motivational, and emotional engagement, it can be assumed that they relate to each other in different ways. First, it can be supposed that the three dimensions are not different from each other and represent one single factor of engagement. Engagement in learning is based on a complex cognitive-affective activity which could reduce the cognitive capacities for distinguishing different multidimensional aspects in information processing. Engagement could therefore be seen as a holistic and integrated way of handling information processing and learning. Such a one-dimensional perspective can be found in a long history of regarding engagement mainly under the focus of instructional time or on academic engaged time for improving student learning (e.g., Fisher & Berliner, 1985). Second, it might be possible that there are two factors in student engagement, a cognitive and a non-cognitive resp. affective one (e.g., Reschly & Christenson, 2012). The cognitive factor subsumes all aspects of student engagement which are related to knowledge and skill application and acquisition (e.g., problem-solving). The non-cognitive factor is about affective resp. motivational and emotional aspects (e.g., connectedness to other learners or issues). Third, it could be assumed that engagement in learning might be related to the assumed three theoretical dimensions of cognitive, emotional, and motivational processes as all three areas represent unique or independent concepts and related research activities (e.g., Reeve & Tseng, 2011).

A second aim of the study is to explore and test the reliability and validity of a short version of a multidimensional measurement on engagement in learning. There are numerous attempts in the literature on measuring cognitive, motivational, or emotional processes in which short versions of measurement instruments were developed. For example, Kroenke, Spitzer, and Williams (2003) evaluated successfully a two-item version of a Patient Health Questionnaire depression module in comparison to a nine-item version. Jackson, Martin, and Eklund (2008) found that a short (9 items) version of a multidimensional assessment of flow had similar model fits, reliabilities, and distributions in comparison to a long 36-item assessment. O'Brien, Cairns, and Hall (2018) developed out from a long version of an engagement scale with 31 items a short version with 12 items in the field of human-computer interaction.

Method

Participants

Participants were 218 undergraduate and graduate students enrolled in two courses at a Department of Educational Science at an Austrian University. The sample size lies within ranges of comparable studies using structural equation modelling and

confirmatory factor analysis: Kline (2015, p. 16) suggested to use a sample-size-to-parameter ratio between 10 to 20: 1 what means that in case of 15 parameters, a sample size between 150 to 300 participants would be optimal. The two courses were lectures on the issue of teaching and learning (53.2 percent of participants) and on the acquisition of complex emotions (46.8 percent) and held every week for about 90 minutes throughout the semester term by the author of the study. Participants' average age was 23.66 (SD = 7.08) years, the majority of participants were female (86.7 percent).

Procedure

Participants were individually tested within a 35 minutes session within a regular lecture unit after about three weeks in the semester. The session was designed as a self-learning and -assessment unit as support for the preparation of the final examinations. Students had to read a text which was adapted to the different contents of the lectures. For the lecture on teaching and learning, students had to read a text on the instructional design of lectures at universities. For the other lecture, there was a text on relationship scripts. Both texts were about 60 to 80 lines in length. In both lectures, about 50 percent of students also had to answer six open questions at the end of the text in order to support self-evaluation. There were no significant differences in overall (summed up) student engagement in groups of students with and without self-evaluation ($t(204) = -0.03$, $p > .05$). Students were instructed to read the texts within a time period of 25 minutes and save the contents of texts as good as possible. After reading the texts, students had about 10 minutes to answer the questions on engagement in learning (together with a knowledge test as preparation for the upcoming final examination).

Measures

For measuring engagement in learning, a 15-items scale by Astleitner (2018b) was used (see all items and *descriptive statistics* within Table 1) for the first time in research activities. Participants were asked to report their experiences during learning. All items had to be answered on a five-point Likert scale ranging from "never" to "(nearly) always". Cognitive engagement concerned the complexity of information processes and products during learning. It ranged from simple (with the item: "can retrieve contents from memory") to complex (item: "can develop new ideas, plan projects, or design products"). Motivational engagement was about the depth of stimulation for actively dealing with information processes and products during learning. It ranged from external (item: "efficaciously concentrates on contents") to internal (item: "engages with contents for their own sake with high satisfaction") sources of stimulation. Social-emotional engagement referred to the emotional attachment with elements of learning. It ranged from distant (item: "experiences feel-

ings of freedom or autonomy in learning") to near (item: "feels secure or sheltered in learning").

Data from Table 1 indicates that most of the items for measuring student engagement have *means* which lie around the midpoint of the scale (i.e., "sometimes", value: 3). Only the items on evaluation and on relevance showed some deviating tendencies. Also, nearly all *standard deviations* were around the value of 1, except for the item on knowledge and comprehension. Overall, these descriptive statistics showed that there were acceptable levels of difficulties and amounts of variance within the items as sufficient basis for further analyses.

Data Analysis

IBM SPSS Statistics 25 (e.g., George & Mallery, 2018) was applied for descriptive data analysis. For testing confirmatory factor analyses models, LISREL 8.8 (Jöreskog & Sörbom, 2006; Jöreskog, Olsson, & Wallentin, 2016) was used. A Kolmogorov-Smirnov test was carried out and distributions of all variables were found to be significantly different from normal ($p < .001$). Thus, PRELIS was used to compute a polyserial correlation matrix and the corresponding asymptotic covariance matrix. For model testing, the weighted least square solution (WLS) was used, because WLS can produce correct standard errors and model fits when non-normality of variables is given. In addition, the variances of the latent variables were set to one, what allows to freely estimate the factor loadings of the items. For evaluating the goodness of fit of the models, different indicators were used like the Chi-Square ($\chi 2$)-value, Non-Normed Fit Index (NNFI), Comparative Fit Index (CFI), and the Standardized Root Mean Square Error of Approximation (RMSEA). Results of $\chi 2/df$ (degrees of freedom) between 2 and 3, NNFI and CFI > .97, and RMSEA values ≤ .05 are considered as good fit (Schermelleh-Engel, Moosbrugger, & Müller, 2003).

Results

Results of the confirmatory factor analyses are depicted in Table 2. In a first step, a confirmatory factor analysis was conducted in which one single underlying factor of student engagement was assumed. In Table 2, fit statistics of the one-factor-model showed that such a model did not fit the given data and cannot be a plausible representation ($\chi 2/df = 5.44$, *NNFI* and *CFI* < .93, *RMSEA* > .05). A deeper inspection of the one-factor model also showed a negative error variance (-.001) for item on intrinsic motivation and modification indices suggested to add 39 error covariances between items. Obviously, results of this first confirmatory factor analysis suggest that there are at least two dimensions within the data on student engagement.

In a second step, a two-factor-solution was tested in which a cognitive dimension and an affective dimension as a combination of motivational and social-emotional engagement were assumed. Again, the model fit was not sufficient ($\chi 2/df = 5.49$, *NNFI* and *CFI* < .93, *RMSEA* > .05), the item on intrinsic motivation had a negative

Tab. 1: Multidimensional Student Engagement Variables (Means (M) and Standard Deviations (SD)) (212 < n < 219)

Items		M	SD
	Cognitive Engagement		
Knowledge:	can retrieve contents from memory.	3.65	0.69
Comprehension:	can summarize, explain, or classify contents.	3.75	0.73
Convergent thinking:	can solve problems by the application of learned procedures.	3.00	0.95
Evaluation:	can find mistakes, criticize, or defend contents based on standards.	2.10	0.96
Synthesis:	can develop new ideas, plan projects, or design products.	2.75	1.06
	Motivational Engagement		
Attention:	efficaciously concentrates on contents.	3.83	0.86
Relevance:	regards contents as personally important.	4.01	0.89
Interest:	reengages voluntarily and repeatedly with contents.	3.53	1.03
Identification:	is committed to goals that are related to contents.	3.06	1.03
Intrinsic motivation:	engages with contents for their own sake with high satisfaction.	3.22	0.96
	Social-emotional Engagement		
Self-assertion:	experiences feelings of freedom or autonomy in learning.	3.17	1.15
Entertainment:	finds joy, fun, or happiness in learning.	2.72	1.05
Belongingness:	feels community or loyalty in learning.	3.22	1.06
Adaptiveness:	is sensible or empathic in relation to elements of learning.	2.89	1.15
Security:	feels secure or sheltered in learning.	2.81	1.13

error variance (-.001) and modifications indices suggested 47 error covariances. Such error covariances indicate that the items or observed variables are measuring something other than the hypothesized factor (Schumacker & Lomax, 2010, p. 165). In addition, there was a very high correlation between the two factors ($r = .98$). All of these indicators showed that student engagement as measured in this study did not have a cognitive and an affective dimension.

In a third step, a three-factor solution with a higher-order factor was tested. Here, it was assumed that the cognitive, motivational, and social-emotional dimensions belonged to the same common construct. A first model-test did not identify a converged solution ($\chi 2/df = 5.56$, *NNFI* and *CFI* < .93, *RMSEA* > .05). After not setting the variance of latent variables to one, a converted, but not acceptable model fit was found ($\chi 2/df = 5.43$, with 50 suggested error covariances). The results of this

Tab. 2: Fit Statistics for the CFA-Models on Student Engagement

Models		χ2	df	NNFI	CFI	RMSEA	p	Changes to fit statistics		
								Δχ2	Δdf	ΔCFI
1.	One factor	489.8	90	0.91	0.92	0.14	0.000	-	-	-
2.	Two factors	488.5	89	0.91	0.92	0.14	0.000	1.3	1	0.00
3.	Three factors – Higher order	484.0	87	0.91	0.92	0.15	0.000	4.5	2	0.00
4.	Three factors	472.1	87	0.91	0.92	0.14	0.000	11.9	0	0.00
5.	Three factors – Reduced with errors	107.1	42	0.95	0.97	0.09	0.000	365.0	45	0.04
6.	Three factors – Strongly reduced without errors	6.0	6	1.00	1.00	0.00	0.424	101.1	36	0.05

model test suggest that the three dimensions did not relate to a common higher-order construct.

In a fourth step, a three-factor solution without a higher-order factor was tested assuming that the cognitive, motivational, and social-emotional dimensions were not held together by a parent construct. Again, the model fit was not acceptable ($\chi 2/df = 5.43$, *NNFI* and *CFI* < .93, *RMSEA* > .05) and modification indices suggested to consider 50 error covariances. Especially, the many identified error covariances suggested that the items were not independent indicators of the underlying constructs.

Therefore, in a fifth step, error terms within but not between dimensions were allowed to correlate. In addition, items with inconsistent values (e.g., negative error variances) were excluded leading to four items per dimension and a significantly improved model fit ($\chi 2/df = 2.55$, *NNFI* and *CFI* > .94, *RMSEA* < .10), however, together with 9 suggested error covariances.

Based on these results, it was concluded to reduce the number of items per dimension and to assume all error terms to be uncorrelated. In addition, for identifying a model with an acceptable fit, no anomalies in values, no suggestions by modification indices, and only statistically significant parameter estimations were used as standards for model evaluation. In addition, troubleshooting tips (e.g., eliminating the bad variables) by Schumacker and Lomax (2010, p. 50, p. 68) were applied to find a good solution. Applying these criteria and tips leaded to an acceptable model fit ($\chi 2/df = 1$, *NNFI* and *CFI* > .97, *RMSEA* < .05). Figure 1 illustrates the resulting reduced factor model of multidimensional student engagement with only 6 out of 15 items. Cognitive engagement is based on knowledge and comprehension, motivational engagement on attention and interest, and social-emotional engagement on self-assertion and entertainment. Correlations between the different dimensions were statistically significant and ranged from $r = .36$ to $r = .88$ ($t > 4.35$) (see right part of Figure 1). On the left part of Figure 1, R^2 as a measure of the strength of the linear relationship is shown. It is usually interpreted as the reliability of the observed variables. For example, comprehension can be seen as the most reliable indicator of cognitive engagement ($R^2 = .85$), attention of motivational engagement ($R^2 = .43$), and entertainment of social-emotional engagement ($R^2 = .75$). Overall, cognitive and social-emotional items showed good and higher reliability in comparison to

motivational engagement. In the center of Figure 1, standardized factor loadings are depicted. They represent validity coefficients, range from .58 to .92, and are all statistically significant ($t > 8.72$). Overall, the reduced factor model showed good model fit and good reliability and validity.

Discussions and Implications

The purpose of this study was to learn about problems and solutions in the measurement of multidimensional constructs which are essential within modern educational intervention research. Within such research, measurements concern hierarchically organized development processes. Such a focus allows to generate adapted educational interventions which are different for different developmental levels. Such an orientation corresponds with contemporary approaches on testing (like "adaptive testing"; Yigit, Sorrel, & de la Torre, 2019), on educational interventions (like "optimized adaptive interventions"; Almirall, Kasari, McCaffrey, & Nahum-Shani, 2018), or on instructional systems design (like "adaptive instructional systems" or "intelligent tutoring systems"; Durlach & Lesgold, 2012).

In this study, a particular multidimensional measurement instrument (from Astleitner, 2018b) on student engagement was tested and evaluated for the first time. Construct validation and related confirmatory factor analyses approaches were used for identifying a short scale with acceptable reliability and validity. Short scales with a low number of items have significant advantages for intervention research as they can be used more efficiently because they require less organizational and time-consuming effort. However, this construct validation brought to light also problems for test development which represent limitations of this study, but also stimulations for future research activities in educational intervention research.

Limitations. The first problem is on representing resp. covering multiple factors. In this study, there was a focus on cognitive, but also motivational as well as social-emotional aspects of learning. The problem here is that, on the one hand, different aspects must be independent from each other. On the other hand, there are interactions between these aspects and related processes. In confirmatory factor analyses approaches, relationships between different factors are measured by correlations between factors (see Figure 1). Recent approaches concern, for example, "bifactor exploratory structural equation modeling to test for a continuum structure" (Howard, Gagné, Morin, & Forest, 2018). However, approaches which are based on correlations do not allow to formulate causal developmental levels and processes, as it was the case in this study. In future studies, especially when they have an interventionist perspective, measurement models with a developmental perspective should be tested via path analysis. Within path analysis, relationships and interactions can be described as mediator and moderator effects. Moderator effects are given, when a variable alters (for example, via zero-order-correlation or interactions) the strength and/or direction between an independent and a dependent variable. Mediator effects are given, when the relationship between indepen-

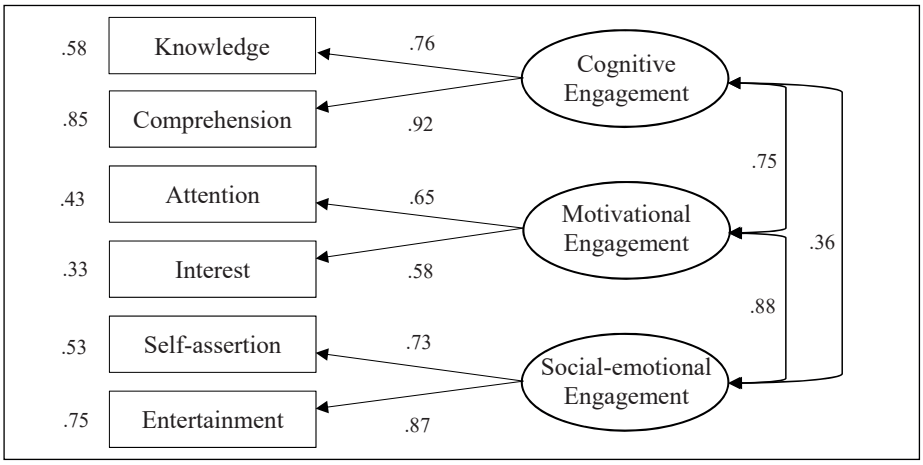

Fig. 1: Reduced factor model of multidimensional student engagement.

dent and dependent variables is caused by a third mediating variable (Little, Card, Bovaird, Preacher, & Crandall, 2007). In general, moderator effects are integrated into structural equation modeling by latent interaction variables through multiplying the related latent variables (Schumacker & Lomex, 2010, p. 333). In specific terms, testing developmental models with structural equation modeling would need approaches based on, for example, "intraindividual variability with repeated measures data" (Hershberger & Moskowitz, 2002), or "growth curves" (Duncan, Duncan, & Strycker, 2006). All these and similar approaches were not used in this study.

The second problem is on the length of the measurement instrument. Within the literature, many different suggestions for the length of a measurement instrument can be found. Within intervention research, this could mean that effect patterns are different for single- and multiple-items measurement approaches. Here, for example, Ang and Eisend (2018) found in their meta-analysis no differences in effect sizes when the dependent variables were measured with single or multiple items. However, Meier (2004) argued that, for intervention research, it is important to have scales which are able to detect a wide range of intervention effects. Such intervention-sensitive measures can be realized by using "intervention item selection rules" like detecting the change of an item's score after an intervention or examining relations between item scores and systematic error sources. Sensitiveness to change or responsiveness concern the ability of a questionnaire to detect relevant resp. important changes due to an intervention. It requires, for example, to identify participants of interventions as improving, worsening, or remaining stable based on external criteria. Then, results on the given measurements can be compared in these groups in order to evaluate responsiveness (Revicki, Hays, Cella, & Sloan, 2008). In this study, there was only a focus on the quantity of items, but not on the quality like sensitiveness to change.

The third problem is on the hierarchical relationships within and between factors. Within developmental models, it is assumed that a higher level cannot be achieved without mastering a lower due to increasing difficulty (e.g., Astleitner, 2008). However, this assumption and similar learning-hierarchy models were questioned in the past (see, for example, Astleitner, 2018a, p. 130; Bergan, 1980). People can reach higher levels of development without fully achieving lower levels. It might be possible that higher levels can be reached, when lower levels are only partially accomplished. It could also be the case that lower and higher levels are achieved more or less simultaneously. Another problem represents the type of relationship: It could be positive or negative linear, but also non-linear like, for example, u or inverted-u. Such complex situations need sophisticated statistical methods for analyses like, for example, "hierarchical analysis" (De Jong, 1999), or "nonlinear modeling" (Dimitruk, Schermelleh-Engel, Kelava, & Moosbrugger, 2007). In this study, structural models were tested without considering the hierarchical status of within and between latent constructs, except the (non-significant) higher order model (see Table 2).

Implications. The first implication of this study is a theoretical one. It is about the type of theory which should guide the design and evaluation of educational interventions and related measurements. Traditionally, there are scientific theories, prescriptive (technological) theories, practical (program) theories, and personal (subjective) theories (Astleitner, 2018a). All of them are relevant for intervention research. Within this study, another type of theory was focused: A multidimensional and level-based developmental theory (see Figure 2). Such a theory combines

- multiple dimensions of a phenomenon (like cognitive, motivational, and social-emotional aspects of student engagement),
- developmental levels within the dimensions (for example, from knowledge to synthesis), and
- level-based intervention strategies for supporting development (for example, varying task-contexts for stimulating convergent thinking).

Such a theory has multiple advantages for intervention research. It allows to have a holistic view on problems, to distinguish and measure individual differences in development, and to use different support strategies which can be adapted to individual differences. Future intervention research has to develop and test such theories in order to find evidence about their capabilities in improving the state-of-the art of educational intervention research.

The second implication of this study concerns a validation of a short scale on student engagement which could be used in further research on multidimensional intervention effects. Of course, further validation attempts are necessary, also because there are many existing well-established measurement instruments on student engagement. For example, Fredricks and McColskey (2012) compared 11 self-report measures on student engagement. The special feature of the measuring

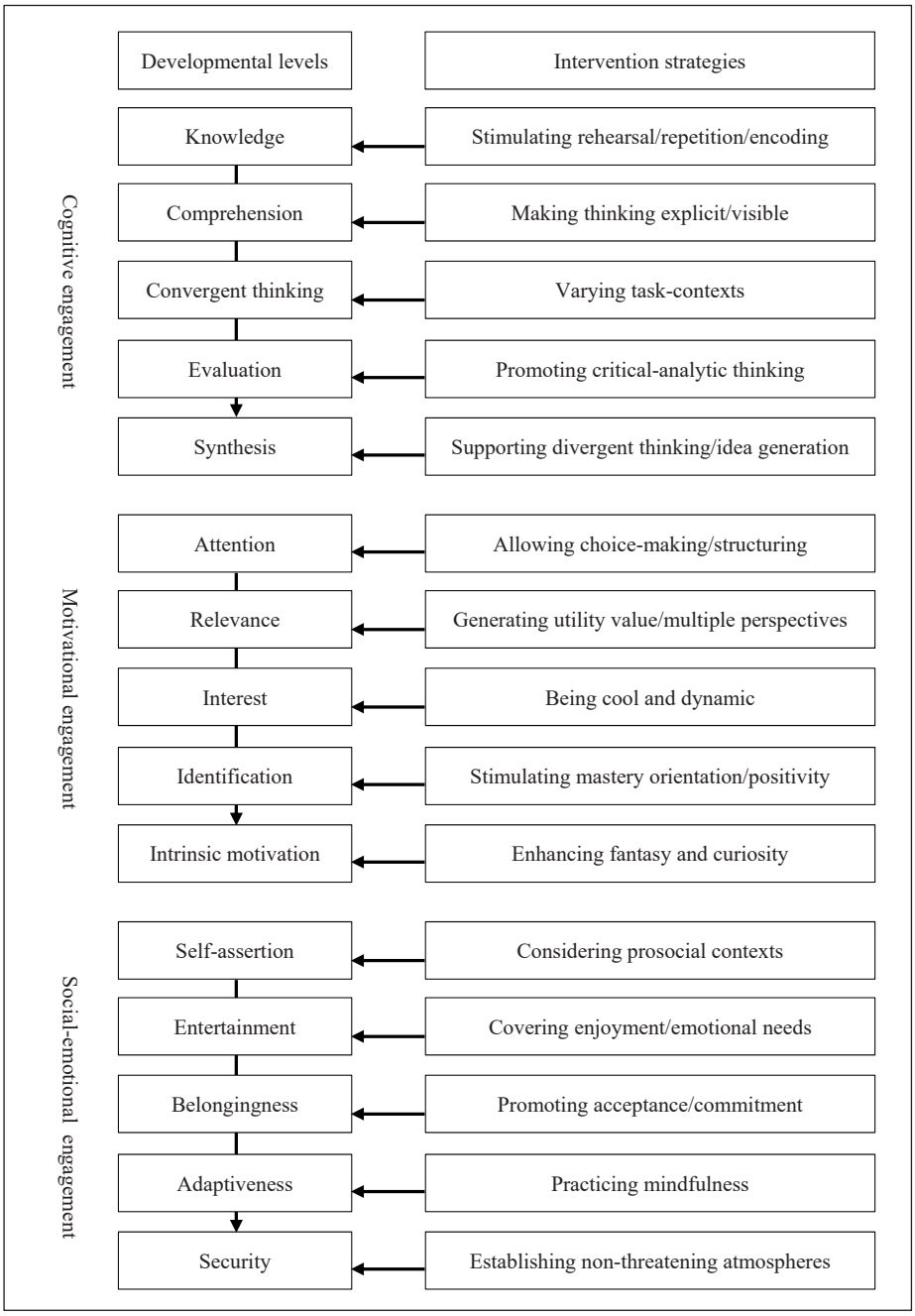

Fig. 2: A multidimensional and level-based developmental approach on student engagement (based on Astleitner, 2018b, p. 13, modified).

instrument which was used in this study is that it is part of an educational developmental or level-based instructional design approach (see Figure 2). Within such an

approach, different hierarchically organized developmental levels (e.g., knowledge and comprehension) are combined with level-based support strategies (e.g., stimulating rehearsal and making thinking visible). Such support strategies represent independent variables when designing educational interventions. In this study, the different support strategies were measured, but not used within statistical analysis. Future research should also compare the scale in this study with other measurements on student engagement.

The third implication might be that in intervention research standards concerning the length of measurement scales should be re-evaluated. The number of items might be less important, instead responsiveness or intervention sensitivity of items should be considered and tested. Within intervention research, the change of a dependent variable is in the focus. A measurement of a variable might be valid, but must not have "validity for change" which means "that a measure shows an observable difference when there is, in fact, a change on the characteristic measured that is of sufficient magnitude to be interesting in the context of application" (Lipsey, 1990, p. 100). Responsiveness is about validity, however, whereas validity concerns the validity of a single score, responsiveness is about the validity of a change score (De Vet, Terwee, Mokkink, & Knol, 2011): It should be evaluated in a longitudinal study in which participants are known to change. Testing responsiveness can be done by comparing changes on the instrument with changes on an important standard or by testing expected mean differences between changes in groups of participants as well as expected correlations between the changes in the scores on the instrument and changes in other instruments. Traditional effect sizes or paired t-tests are not suitable for testing responsiveness, because they are testing the magnitude of the change scores, rather than the validity of the change scores. Future research should clarify the importance of responsiveness in educational intervention research.

References

Almirall, D., Kasari, C., McCaffrey, D. F., & Nahum-Shani, I. (2018). Developing optimized adaptive interventions in education. *Journal of Research on Educational Effectiveness*, *11*, 27–34. doi: https://doi.org/10.1080/19345747.2017.1407136

Ang, L., & Eisend, M. (2018). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research*, *58*, 218–227. doi: https://doi.org/10.2501/JAR-2017-001

Astleitner, H. (2008). Die lernrelevante Ordnung von Aufgaben nach der Aufgabenschwierigkeit [The learning relevant organization of tasks in relation to task difficulty]. In J. Thonhauser (Ed.), *Aufgaben als Katalysatoren von Lernprozessen* (pp. 65–80). Münster: Waxmann.

Astleitner, H. (2018a). Spezielle Verfahren sozialwissenschaftlicher Theorieentwicklung [Special methods of theory building in the social sciences]. Weinheim: Beltz Juventa.

Astleitner, H. (2018b). Multidimensional engagement in learning – An integrated instructional design approach. *Journal of Instructional Research, 7, 6–32*. doi: https://doi.org/10.9743/JIR.2018.1

Bergan, J. R. (1980). The structural analysis of behavior: An alternative to the learning-hierarchy model. *Review of Educational Research*, *50*, 625–646. doi: https://doi.org/10.3102/00346543050004625

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*, 175–184. doi: https://doi.org/10.1509/jmkr.44.2.175

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Carpendale, J. I. (2000). Kohlberg and Piaget on stages and moral reasoning. *Developmental Review*, *20*, 181–205. doi: https://doi.org/10.1006/drev.1999.0500

De Jong, P. F. (1999). Hierarchical regression analysis in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 198–211. doi: https://doi.org/10.1080/10705519909540128

De Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge: University Press. doi: https://doi.org/10.1017/CBO9780511996214

Dimitruk, P., Schermelleh-Engel, K., Kelava, A., & Moosbrugger, H. (2007). Challenges in nonlinear structural equation modeling. *Methodology*, *3*, 100–114. doi: https://doi.org/10.1027/1614-2241.3.3.100

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, *38*, 105–123. Retrieved from https://search.informit.com.au/documentSummary;dn=491551710186460;res=IELAPA

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling* (2nd ed.). Mahwah: Erlbaum.

Durlach, P. J., & Lesgold, A. M. (Eds.). (2012). *Adaptive technologies for training and education*. Cambridge: University Press. doi: https://doi.org/10.1017/CBO9781139049580

Fisher, C. W., & Berliner, D. C. (1985). *Perspectives on instructional time*. New York: Longman.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378. doi: https://doi.org/10.1177/1948550617693063

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*, 350–365. doi: http://doi.org/10.1037/0022-3514.78.2.350

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 763–782). New York: Springer. doi: https://doi.org/10.1007/978-1-4614-2018-7_37

George, D., & Mallery, P. (2018). *IBM SPSS statistics 25 step by step. A simple guide and reference* (15th ed.). New York, Abingdon: Routledge. doi: https://doi.org/10.4324/9781351033909

Gratz, K. L., & Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in emotion regulation scale. *Journal of Psychopathology and Behavioral Assessment*, *26*, 41–54. doi: https://doi.org/10.1023/B:JOBA.0000007455.08539.94

Hammer, M. R., Bennett, M. J., & Wiseman, R. (2003). Measuring intercultural sensitivity: The intercultural development inventory. *International Journal of Intercultural Relations*, *27*, 421–443. doi: https://doi.org/10.1016/S0147-1767(03)00032-4

Hershberger, S. L., & Moskowitz, D. S. (Eds.). (2002). *Modeling intraindividual variability with repeated measures data. Methods and applications*. New York: Psychology Press.

Howard, J. L., Gagné, M., Morin, A. J., & Forest, J. (2018). Using bifactor exploratory structural equation modeling to test for a continuum structure of motivation. *Journal of Management*, *44*, 2638–2664. doi: https://doi.org/10.1177/0149206316645653

Jackson, S. A., Martin, A. J., & Eklund, R. C. (2008). Long and short measures of flow: The construct validity of the FSS-2, DFS-2, and new brief counterparts. *Journal of Sport and Exercise Psychology*, *30*, 561–587. doi: https://doi.org/10.1123/jsep.30.5.561

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.8: Interactive LISREL: Technical support*. Mooresville: Scientific Software.

Jöreskog, K. G., Olsson, U. H., & Wallentin, F. Y. (2016). *Multivariate analysis with LISREL*. Cham: Springer. doi: https://doi.org/10.1007/978-3-319-33153-9

Kline, R. B. (2015). *Principles and practice of structural equation modelin*g (4th ed.). New York: Guilford.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2003). The Patient Health Questionnaire-2: Validity of a two-item depression screener. *Medical Care*, 1284–1292. doi: https://doi.org/10.1097/01.MLR.0000093487.78664.3C

Laborde, S., Allen, M. S., & Guillén, F. (2016). Construct and concurrent validity of the short-and long-form versions of the trait emotional intelligence questionnaire. *Personality and Individual Differences*, *101*, 232–235. doi: https://doi.org/10.1016/j.paid.2016.06.009

LaNasa, S. M., Cabrera, A. F., & Trangsrud, H. (2009). The construct validity of student engagement: A confirmatory factor analysis approach. *Research in Higher Education*, *50*, 315–332. doi: https://doi.org/10.1007/s11162-009-9123-1

Lipsey, M. W. (1990). *Design sensitivity. Statistical power for experimental research*. Newbury Park: Sage.

Little, T. D., Card, N. A., Bovaird, J. A., Preacher, K. J., & Crandall, C. S. (2007). Structural equation modeling of mediation and moderation with contextual factors. In T.D. Little, J.A. Bovaird & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies (*pp. 207–230). Mahwah: Erlbaum. https://doi.org/10.4324/9780203936825

Littman, A. J., White, E., Satia, J. A., Bowen, D. J., & Kristal, A. R. (2006). Reliability and validity of 2 single-item measures of psychosocial stress. *Epidemiology*, 398–403. doi: https://doi.org/10.1097/01.ede.0000219721.89552.51

Lourenço, O. M. (2016). Developmental stages, Piagetian stages in particular: A critical review. *New Ideas in Psychology*, *40*, 123–137. doi: https://doi.org/10.1016/j.newideapsych.2015.08.002

Martin, A. J. (2008). Enhancing student motivation and engagement: The effects of a multidimensional intervention. *Contemporary Educational Psychology*, *33*, 239–269. doi: https://doi.org/10.1016/j.cedpsych.2006.11.003

Marzano, R. J., & Kendall, J. S. (2008). *Designing & assessing educational objectives. Applying the new taxonomy*. Thousand Oaks: Sage.

McInerney, D. M., & Ali, J. (2006). Multidimensional and hierarchical assessment of school motivation: Cross-cultural validation. *Educational Psychology*, *26*, 717–734. doi: https://doi.org/10.1080/01443410500342559

Meier, S. T. (2004). Improving design sensitivity through intervention-sensitive measures. *American Journal of Evaluation*, *25*, 321–334. doi: https://doi.org/10.1177/109821400402500304

Mitchell, M. L., & Jolley, J. M. (2010). *Research design explained* (7th ed.). Belmont: Wadsworth.

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, *112*, 28–39. doi: https://doi.org/10.1016/j.ijhcs.2018.01.004

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology*, *52*, 597–617. doi: https://doi.org/10.1111/bjso.12006

Reeve, J., & Tseng, C. M. (2011). Agency as a fourth aspect of students' engagement during learning activities. *Contemporary Educational Psychology*, *36*, 257–267. doi: https://doi.org/10.1016/j.cedpsych.2011.05.002

Reschly, A. L., & Christenson, S. L. (2012). Jingle, Jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 3–19). New York: Springer. doi: https://doi.org/10.1007/978-1-4614-2018-7_1

Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, *61*, 102–109. doi: https://doi.org/10.1016/j.jclinepi.2007.03.012

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*, 54–67. doi: https://doi.org/10.1006/ceps.1999.1020

Savelsbergh, C. M., van der Heijden, B. I., & Poell, R. F. (2009). The development and empirical validation of a multidimensional measurement instrument for team learning behaviors. *Small Group Research*, *40*, 578–607. doi: https://doi.org/10.1177/1046496409340055

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-

fit measures. *Methods of Psychological Research Online*, *8*, 23–74. Retrieved from https://www.dgps.de/fachgruppen/methoden/mpr-online/

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York, Hove: Routledge.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement*, *43*, 388–401. doi: https://doi.org/10.1177/0146621618798665

6. Pretest Bias: Supporting Undergraduate Learning Through Guided Self-Assessment and Reflective Writing

Hermann Astleitner, Michaela Katstaller, Josef Eisner, Ulrike Greiner & Nomy Dickman

Abstract: Pretests play an important role in estimating the effect patterns of educational interventions. The current study tested the effects of self-assessment and reflective writing activities for undergraduate learning. In a traditional non-randomized pre-post-design, one group of students (the intervention group) had to evaluate their knowledge acquisition activities and to write about their self-image as a scientist and their scientific interests after course sessions. Results showed that participants in the intervention group reported better learning in comparison with the control group without such activities. Learning measurements concerned knowledge acquisition, social research skills and the use of scientific tools, as well as motivational and emotional factors such as joyous exploration and emotional tension handling. However, correlations between variables in the pretests were different from posttests indicating pretest bias. Mediation analysis revealed that knowledge acquisition was related to the use of scientific tools, whereas social research skills were affected by joyous exploration. Discussions focused on multidimensional support of learning, but also on pretest bias in intervention research.

Traditionally, educational intervention research focuses on testing the effects of an intervention between different groups, but also between pre- and posttests. As a rule, participants of an intervention study do not have significant knowledge or skills in a pretest, but should have in a posttest. Kim and Willson (2010, p. 745) argued that pretests can produce bias as they

> may increase (or decrease, depending on the characteristic of the test) scores at posttest not only for the same or similar scales but also for dissimilar scales in which the constructs may be completely different. Whatever the usage of the pretest in research is, the presence of pretest can alter the nature of the intervention and consequently cause problems in measuring the treatment effect per se.

From this situation, it might be concluded that variables and their relationships in the pretest are quite different in comparison to the posttest. When they are different, then it might be probable that they vary in reliability and validity as well as in

correlations between variables. For example, Salgado (1995) and Moscoso (2003) found high variability in the validity coefficients even when measurement settings were held constant and even within settings. Also, Kieffer and MacDonald (2011) found varying reliability coefficients in confrontive coping and distancing as well as seeking social support, escape-avoidance, planful problem solving, and positive reappraisal scales.

Within this study, we explore effect patterns of self-assessment and writing activities within undergraduates learning. We will use pretests and assume that variables in the pretests will correlate differently than in comparison to the posttests. Having this assumption, we will use mediation analysis to explain possible differences. Mediation analysis (about the causal relationships between variables) is closely related to, for example, internal or construct validity and allows also to gain knowledge about problems or bias in validity (e.g., Stone-Romero & Rosopa, 2010). We assume such problems because undergraduate learning is associated with sophisticated educational goals (e.g., Stenlund, 2010).

University undergraduate students should not only become subject matter experts, but should also acquire sophisticated research skills, the latter being anything but trivial. For example, Arum and Roksa (2011) used standardized tests and found that a significant proportion of undergraduate students demonstrated no improvement at the end of their second year in a range of skills such as critical thinking, complex reasoning, or scientific writing. Also, Roohr, Liu, and Liu (2017) found in a longitudinal study that after being in college for one or two years, students did not demonstrate significant learning gains in basic research-related skills. A study from Mathers, Finney, and Hathcoat (2018) revealed that students were making somewhat inexplicable gains in quantitative and scientific reasoning. These improvements however were not related significantly to course assignments. Also, others (see chapter 4 in this book) found that undergraduate (teacher education) students had problems in acquiring higher-order skills like critical thinking or mental modeling although courses were consequently based on principles of effective instructional design.

A clarification needs to be found to explain these findings as well as those resulting from recent meta-analyses (e.g., Hattie, 2015; Schneider & Preckel, 2017) and from a widely established learner-centered paradigm of higher education (Reigeluth, Myers, & Lee, 2017). The most important variable for explaining limited learning is that of self-regulated learning activities based on assessment. Such self-assessment activities have a long tradition in the field of higher education with significant effectiveness for undergraduate learning (e.g., Falchikov & Boud, 1989). For example, Thompson, Pilgrim, and Oliver (2005) found that even relatively simple and easy-to-use self-assessment tools helped first-year students to plan and organize their thoughts and to become more independent and reflective. However, self-assessment activities ranged in their effect patterns: Panadero, Jonsson, and Botella (2017) found only low to medium effects sizes (from 0.23 to 0.65) of self-assessment on different measures of learning. It was found that an important element for effectivity was that students should not only assess their performance, but also should be given

support in order to close the gap between current and desired performance (e.g., Nicol & Macfarlane-Dick, 2006). This support would be guidance (as guidelines, plans, advices, or hints) about what should be done in the future in order to increase reflective processes and learning. To combine self-assessment with learning guidance might produce additional effectiveness in a student's self-regulated learning. This combination was repeatedly found to improve learning significantly (e.g., Lazonder & Harmsen, 2016), whereas minimal guidance was found to be ineffective (e.g., Kirschner, Sweller, & Clark, 2006).

Another important source for improvement of self-assessment effects on learning, requires handling cognitive, motivational, and emotional consequences (e.g., on self-efficacy or anxiety) resulting from good or bad performance. An effective and easy-to-use way of handling such consequences comes from writing activities (e.g., Bazerman, 2007). Especially, "reflective writing" as activity which aims to "take us out of our own narrow range of experience and help us to perceive experiences form a range of viewpoints and potential scenarios" (Bolton, 2010, p. 10) provides comprehensive learning opportunities. Research results report that providing teacher students with opportunities to uncover one's personal beliefs as well as motivational and emotional conditions experienced during their studies can be crucial for enhancing their learning. Reflective writing was found to be effective when teacher students were engaged in exploring their personal epistemology through guided reflection (by using written tasks) (Kyles & Olafson, 2008). It also represents a strong vehicle to help teacher students to become aware of their own self-perceptions and to explore new ways of self-expression, especially when identifying oneself as a research guided teacher (Lea & Stierer, 2011). Reflective writing was also found to be effective for triggering critical consciousness and self-reflection about ones' goals and interests (Brown, 1998). In addition, research on writing in initial teacher education provided evidence about how reflective writing as an educational tool can address students' personal processes of meaning-making (Wittek, Askeland, & Aamotsbakken, 2015). Overall, reflective writing can be seen as a powerful tool to increase personal motivation by uncovering undiscovered strengths and powerful wishes about one's professional development (Cohen-Sayag & Fischl, 2012).

In our study, we responded to such evidence in fostering undergraduates learning in three ways. Firstly, we decided to combine guidance-based self-assessment with reflective writing in order to increase knowledge acquisition and to support positive motivational and emotional processes during learning. This was done in a way where cognitive overload and related negative effects on learning should be avoided (e.g., Mayer, Heiser, & Lonn, 2001). We designed our learning support in a way that it altered with time: In one lesson, students had to do self-assessments and in the next lesson, students applied reflective writing. Secondly, we focused on knowledge acquisition with a summative performance test, but also with formative process variables like the development of social research skills and the use of scientific tools. Formative and summative assessment methods were integrated, not only for increasing explanatory power and validity, but also because there was strong

evidence that undergraduate learning profited from such an integration, even in large classes (e.g., Broadbent, Panadero, & Boud, 2018). Thirdly, we addressed not only cognitive, but also motivational and emotional aspects of learning by considering research findings on multidimensional engagement in learning (e.g., Astleitner, 2018a). From a motivational perspective, we focused on curiosity (and related explorative behavior) as it was found to be essential for undergraduate students and their intrinsically-motivated learning (e.g., MacKinnon, 2017). From an emotional perspective, we considered the handling of emotional tension during learning as it was closely related to motivational-based curiosity and as it covered the negative emotional aspects of learning and related stress (i.e., frustration about not finding suitable solutions) (e.g., Dixon & Kurpius, 2008).

Purpose and Hypotheses

The first objective of this study was to explore effect patterns of combined self-assessment and reflective writing within student learning. The second objective was to learn more about the mechanisms which exist between cognitive and non-cognitive factors and between formative and summative indications of student performance. We used a quasi-experimental pre-posttest-follow up-design to examine these questions. In both intervention and control groups, undergraduate students were confronted with methods related to social research. In addition, within the intervention group, undergraduates were given alternating assignments for self-assessment and reflective writing.

We first hypothesized that both groups would demonstrate increased knowledge acquisition, social research skills, and use of scientific tools (in post- and follow up-tests in comparison to a pretest). In both groups, effective principles of instructional design (from Merrill, 2002) were applied which should produce improved knowledge and skills: Students were engaged in real research problems. Their prior knowledge was activated, new knowledge was demonstrated to them, and they had to apply and integrate the acquired knowledge in order to solve task assignments successfully. These principles subsume important empirical research findings and were found to be effective in fostering undergraduate learning on research methods (Tu & Snyder, 2017) and on motivational aspects of learning (Lee & Koszalka, 2016). We also expected that the implementation of these principles of instructional design should increase successes in learning as well as increasing pleasant learning experiences. This would in turn support emotional tension handling (e.g., Putwain, Sander, & Larkin, 2013).

In addition, it was hypothesized that the use of self-assessment and reflective writing should show an increase in these measures within the intervention group (in comparison to a non-use situation within the control group). Self-assessment should deliver knowledge about the current level of understanding and the related guidance devices should allow a more individualized and therefore optimized learning support. Accurate self-assessments and the use of assessment results selectively

for the optimized selection of new learning activities should increase the knowledge acquisition gained from self-regulated learning as found in a study from Kostons, Van Gog, and Paas (2012). Reflective writing should have an additive effect and stimulate extra cognitive processing and related meaning making on course content and therefore enhance learning (Boals, 2012; Dickman, 2005). In addition, there is evidence that self-assessment (with guidance) and reflective writing had concurrent positive effects concurrently on a broad spectrum of motivational and emotional variables (e.g., Clark, 2012; Kirk, Schutte, & Hine, 2011). Therefore, we expected that improved learning from self-assessment and reflective writing should also increase joyous exploration and emotional tension handling due to an additional learning support. There should also be more success in selecting optimal tasks for learning, deeper and enthralling insights, and more pleasant experiences.

Method

Participants

A total of 48 undergraduate students from five courses on research methods at a School-of-Education and at a Department of Educational Science from an Austrian university volunteered to participate in the study. One course A (with an intervention and a control group) was held by the first author of this study and dealt with methods of single-case research (e.g., Morgan & Morgan, 2009). The other two courses, B and C (with one course as intervention and one course as control group) were held by the second and third authors. These were concerned with theories of educational research as well as on an introduction to scientific work (e.g., Boudah, 2011). Participants within these courses were randomly assigned to intervention and control groups. They were between 19 and 37 years of age ($M = 24.07$, $SD = 4.85$); 64 percent of them were female, 36 percent male. Intervention and control groups did not differ significantly concerning age ($t(43) = .502$, $p > .05$) and sex ($t(43) = 1.203$, $p > .05$). We also asked participants about their style of learning at the beginning of the course with the following items (based on a four-point Likert scale with "never" to "(nearly) always"): "I learn in rational way, structured and with plan", "I need emotional experiences and social contacts in learning", "I let my curiosity and interests guide me", "I give my best to achieve my goals consistently and disciplined", "I build good feelings on the study material and relax before learning", "I critically question what I should learn and test what I'm learning if it's really correct", "I repeat at regular intervals what I have learned so that I will not forget it", "I have great confidence in my abilities" and "I'm satisfied and balanced in terms of my knowledge and skills". Multiple t-tests revealed no significant differences on these items between the intervention and the control group (largest $t(43) = 1.439$, $p > .05$). Overall, both groups were comparable with regard to age, sex, and styles of learning before the intervention.

Design

We employed a quasi-experimental pre-post-design to explore the effect patterns of our instructional interventions. The whole experiment was implemented within a 10-weeks period in an intervention and a control group. At a first session, pretests were taken at the beginning of the course session. Within the next four sessions, the two interventions, which needed overall about 20 to 30 minutes in time, were implemented alternately in the intervention group. Immediately after the end of the intervention, the posttests and about two weeks later, the follow-up-tests were conducted within the course session. Intervention and control groups had the same instructors, goals, schedules, contents, assignments, learning materials and teaching methods. Within the intervention group, students had to perform self-assessments with guidance and reflective writing outside the course sessions. Within the control group, there were no such tasks.

Intervention. Within the intervention group, participants had to accomplish a self-assessment task after the second and fourth course sessions as well as doing reflective writing after the first and third sessions. Both intervention tasks had to be accomplished after the course sessions at home.

Self-assessment with guidance was stimulated by a short questionnaire with the following instructions: "Please assess the dimensions given here and determine for the current time to what extent (from 0 to 100%) your knowledge has been developed in this course. Select in the right column all points that you intend to consider until the next session date for improving your knowledge acquisition. Information relates to books, parts of books, articles, etc., which you need to achieve your goals or accomplish the task in the course. Please work diligently and use about 15 minutes for it". Within this questionnaire, participants had to assess 14 skill dimensions on whether they had a sufficient amount of information to solve a given problem, they could assess the desired information and dispose of information which was based on current international research and so on. For each dimension, three activities that support knowledge acquisition could be selected. Here is an example of a skill dimension and related support activities: Skill: "I dispose of information from a scientifically credible source", activities: "check if authors are scientists", "check if a scientific publication is given" and "check citation rate of the author".

The intervention of reflective writing consisted of two sub tasks and was stimulated with the following instructions: 1. "Please write, as soon as possible after the course, openly about how you see yourself as a researcher? What do you consider as your strengths or development potentials as a researcher? Which goals in research would you like to achieve during the course? Take at least 10 to 15 minutes for this. Deliver this sheet reliably at the next course session. You are also welcome to write on the additional sheets. Please write legibly". 2. "Which topic of today's course did you particularly care about? Which topic would you like to explore more deeply? What exactly do you want to know? How would you proceed from a methodological perspective? Take at least 10 to 15 minutes for this".

Manipulation check and controlling for bias. Manipulation checks showed that participants within the intervention group reacted positively on the intervention manipulations: After the third and the fifth session, participants in the intervention group had to indicate (as part of the intervention) which learning support activities they would apply in the current situation. There were 42 activities from the self-assessment like "searching with scholar.google and google", "examining citations", or "checking the impact of a researcher". After the third session, the *lowest* indicated activities were "printing and filing important studies" and "identifying relevant group of researchers". These two activities were chosen by about 39 percent of the intervention group members. After the fifth session, the *lowest* indicated activities concerned "checking whether study authors are scientists" and "depicting the research history of an issue" with about 52 percent.

In order to control experimenter bias, participants had to rate the courses after the third session by using a teaching effectiveness scale (each of the remaining 11 (out of 14) items based on a seven-point Likert scale from "very low" to "very high"; $\alpha = .69$). Items concerned research orientation within the course, degree of innovation in teaching content, usability of course contents, matching of goals, contents as well as examinations, quality of learning coaching, promoting the autonomy of students, leadership skills of the instructor, teaching method variety and variation, contribution of the course to motivate for study program, and teaching of ethical and legal standards of research. Testing overall (summed up) teaching effectiveness on these items between the intervention and control group revealed no significant differences ($t(41) = -.43$, $p > .05$). We interpret this result as indication of no experimenter bias. In addition, these results indicate, as assumed before the intervention, also no ethically questionable disadvantages of the control group like a lower instructional quality.

Measures and Indications of Reliability and Validity

All dependent variables were measured within the regular course time before the intervention, at the end of the intervention, and as follow-ups. Measurements within the course A of the first author were taken offline (first author); measurements within courses B and C of (second and third authors respectively) were taken online. Table 1 presents reliability information and correlations between study measures of the pre- and posttests.

Knowledge acquisition was measured with self-designed curriculum-based tests. The test in course A had originally twelve (mostly single- or multiple-choice) questions on methods of intervention research. For example, "Internal validity is about a) whether results of an intervention can be generalized about persons, settings, times and so on, b) whether a sample of participants is representative, or c) whether the intervention and nothing else was the cause of change". Or, "Imagine the following research situation: A smoker classifies his smoking situations A, B and C into 'at work', 'in the pub', and 'at home' and, as dependent variable, he estimates

Tab. 1: Reliability Information and Correlation between Study Measures

Pretest, 46 ≥ N ≥ 42

Variables	**α**	**1**	**2**	**3**	**4**	**5**
1: Knowledge acquisition	#	-				
2: Social research skills	.70	.06	-			
3: Use of scientific tools	.67	.05	.27	-		
4: Joyous exploration	.80	.02	-.01	-.17	-	
5: Emotional tension handling	.91	-.43***	-.16	-.10	.44***	-

Note. α = Cronbach's Alpha, ***p < .001, one-tailed; # = Negative average covariance.

Posttest, N = 40

Variables	**α**	**1**	**2**	**3**	**4**	**5**
1: Knowledge acquisition	.73/.82	-				
2: Social research skills	.87	.46***	-			
3: Use of scientific tools	.82	.68***	.62***	-		
4: Joyous exploration	.87	.61***	.74***	.68***	-	
5: Emotional tension handling	.92	.55***	.69***	.53***	.80***	-

Note. α = Cronbach's Alpha, ***p < .001, one-tailed.

the number of smoked cigars. Intervention X consists of paying one Euro for each smoked cigar. Please decide whether this is a) an ABA-design, b) a multiple-baseline-design, c) an ABC-design, or d) no single-case design". The tests in courses B and C consisted of nine tasks on social research methods like "Imagine, one of your classes (experimental group) receives a three-week training to promote critical thinking, the parallel class receives no training (control group). After three weeks, a standardized test measures the critical thinking ability of the students. In the experimental group, the test will be conducted in the second lesson in the morning, and in the control group in the eighth lesson in the afternoon. Results show better performance in the experimental group. How do you evaluate this quasi-experimental study? a) The external validity of this study results for similar school classes is high, b) There are confounded effects between the group variable and the critical thinking test performance, c) There are confounded effects between the group variable and the interfering variables, d) The result of this study can be interpreted conclusively due to its high internal validity"; or "A poor operationalization of a construct like classroom management is an indication that a) a valid test of the research hypotheses is not feasible, b) the research hypotheses were not formulated correctly, c) the underlying theory is not appropriate, d) the construct is not empirically verifiable". Reliability analyses on posttests (with the exclusion of one item from the test in course A) showed acceptable difficulty levels (ranging from .30 to .80 for the test in course A, and from .50 to .74 for tests in courses B and C) and good internal consis-

tency for both knowledge acquisition tests (α for test in course A = .73, for tests in courses B and C = .82).

The measurement of social research skills and the use of scientific tools was orientated on a research self-efficacy scale by Holden, Barker, Meenaghan and Rosenberg (1999). For measuring social research skills, participants had to rate how successful (from 0 to 100%) they were in solving research-based tasks. These tasks were described in five items (with good reliability in the posttest, α = .87) like "choose a research design that will allow to test hypotheses with high reliability and validity" or "select a sample in such a way that representativeness and/or as little as possible disturbing influences are given". The use of scientific tools was measured with three items, the same rating scale and with good reliability (α = .82). Items were, for example, about to "find suitable literature for a specific research question in online databases" or "use computer and software for word processing, literature management, and calculations".

For measuring joyous exploration and emotional tension handling, adapted sub scales from the five-dimensional curiosity scale by Kashdan et al. (2018) were used. Five items for each scale had to be answered on a six-point Likert scale (ranging from 1 = "is not true at all" to 6 = "is entirely true") which resulted in good reliability in the posttests (α for joyous exploration = .87 and for emotional tension handling with the "deprivation sensitivity" subscale = .92). Items like "I view challenging situations at the university as an opportunity to grow and learn" or "I enjoy learning about subjects in science that are unfamiliar to me" were used for measuring joyous exploration. Items such as "I can spend hours on a single scientific problem because I just can't rest without knowing the answer" or "It frustrates me not having all the information I need for a research project" were used for emotional tension handling.

From a measurement validity perspective, it was expected that knowledge acquisition should correlate positively with learning supporting processes like the application of social research skills and the use of scientific tools. Joyous exploration should represent a motivational trigger element in knowledge acquisition, and emotional tension handling an emotional one in undergraduate students (e.g., Mega, Ronconi, & De Beni, 2014). In our study, we found significant correlations between all of these variables supporting the assumption of good validity of related measurements (see the posttest results in Table 1; $r > .45$; $p < .001$).

Pretest-bias or pre- and posttest differences in correlations. However, these correlations were only as expected in the posttest. Within the pretest, most of the correlations were non-significant (see the pretest results in Table 1). At first sight, we attribute these differences between pre- and posttest on the state- and not trait-character of the measurements indicating unstable and learnable personality characteristics. In a next step and considering validity and possible biases, we will try to statistically analyze in more detail mediation processes.

Data Analysis

Data was analyzed using IBM SPSS 25. Repeated measures analyses of variance (ANOVA) and a regression analysis were employed for computing statistics. ANOVAs with the factors time (pre- vs. posttest vs. follow up) and group (intervention vs. control group) were computed to test the expected increase in knowledge acquisition and related variables. We set the Alpha-level for all statistical analyses at $p < .05$. As an indicator of effect size, we used partial eta square (η^2_P) which are considered as large at > .14 (Lipsey, 1990, p. 58). For final mediation analysis, LISREL 8.8 was employed (Jöreskog & Sörbom, 1993).

Results

Pre-, post-, and follow up-test means, standard deviations, and results of repeated measures analyses of variance are illustrated in Table 2.

Overall and on a descriptive level, mean comparisons on post- and follow up-tests showed no fully consistent effect patterns. Means in the intervention group increased over time for social research skills, use of scientific tools, joyous exploration, and emotional tension handling. Also, within this group, means increased for knowledge acquisition from pre- to posttest, but decreased from post- to follow up-test. There were different patterns within the control groups: Means for knowledge acquisition increased over time, but means for social research skills, use of scientific tools, joyous exploration, and emotional tension handling decreased.

ANOVAs revealed highly significant and strong changes over time for participants on the variables of knowledge acquisition ($F(2, 68) = 15.57$, $p = .000$, $\eta^2_P = .31$) and social research skills ($F(2, 66) = 15.24$, $p = .000$, $\eta^2_P = .32$). Three further significant and moderate changes over time occurred regarding use of scientific tools ($F(2, 68) = 3.05$, $p = .054$, $\eta^2_P = .08$), joyous exploration ($F(2, 68) = 4.13$, $p = .020$, $\eta^2_P = .11$), and emotional tension handling ($F(2, 68) = 7.45$, $p = .001$, $\eta^2_P = .18$).

However, there was a strong interaction effect of group and time on knowledge acquisition ($F(2, 68) = 9.09$, $p = .000$, $\eta^2_P = .21$) indicating a higher increase of knowledge acquisition in the intervention than in the control group. Further interaction effects showed a strong increase of social research skills ($F(2, 66) = 29.28$, $p = .000$, $\eta^2_P = .47$), use of scientific tools ($F(2, 68) = 11.00$, $p = .002$, $\eta^2_P = .24$), joyous exploration ($F(2, 68) = 37.09$, $p = .000$, $\eta^2_P = .52$), and emotional tension handling ($F(2, 68) = 22.49$, $p = .000$, $\eta^2_P = .40$) in the intervention group. In contrast, there was a decrease or stabilization in the control group.

As results in the control group were not as positive as expected, we used all five variables from pretests in order to explore posttest results. Overall, regression analysis showed a non-significant model fit ($F(5, 11) = 2.23$, $p = 0.125$). All pretest variables had no significant effects on knowledge acquisition ($t < -1.63$), except the results of the pretest on knowledge acquisition ($B = .70$, $t = 2.72$, $p = .020$).

Tab. 2: Descriptive Statistics and Repeated Measures Analysis of Variance Results for Student Learning ($37 \geq N \geq 34$)

	Pretest		Posttest		Follow up		Main effects						Interaction		
							Group (G)			Time (T)			G x T		
	M	(SD)	M	(SD)	M	(SD)	F	p	η^2_p	F	p	η^2_p	F	p	η^2_p
Knowledge acquisition							9.54	.004	.22	15.57	.000	.31	9.09	.000	.21
CG	.40	.16	.41	.17	.57	.25									
IG	.38	.17	.75	.20	.66	.23									
Social research skills							15.88	.000	.33	15.24	.000	.32	29.28	.000	.47
CG	.55	.23	.46	.05	.50	.06									
IG	.37	.19	.72	.11	.76	.10									
Use of scientific tools							15.43	.000	.31	3.05	.054	.08	11.00	.002	.24
CG	.69	.25	.52	.15	.53	.15									
IG	.72	.16	.76	.13	.79	.12									
Joyous exploration							39.62	.000	.54	4.13	.020	.11	37.09	.000	.52
CG	4.73	.97	3.36	.80	3.29	.89									
IG	4.46	.74	5.08	.48	5.25	.45									
Emotional tension handling							43.15	.000	.56	7.45	.001	.18	22.49	.000	.40
CG	4.63	.99	3.36	.80	3.15	.77									
IG	4.75	.95	5.08	.48	5.16	.62									

Note. M = mean, SD = standard deviation, CG = control group, IG = intervention group.

Mediation analyses and posttest results as based on hot cognitions. Results of a mediation analysis are depicted in Figure 1. Path analysis with LISREL (based on the correlation matrix of the posttest results from Table 1) showed a significant but not perfect fit to the data ($\chi 2 = 9.34$, $df = 5$, $p = .096$, $RMSEA = .15$, $GFI = .91$; for all beta weights *(B): t > 3.53;* no modification indices). Testing a path model revealed that the use of scientific tools had a significant effect on joyous exploration ($B = .36$) and on knowledge acquisition ($B = .68$). Emotional tension handling was related to joyous exploration ($B = .61$) which affected social research skills ($B = .74$). All these relationships are plausible from a theoretical perspective and allow to conclude that the intervention successfully produced significant learning-related processes.

These processes might also explain, why pretest results were different from posttest results: During the pretest, none of the given processes were active or triggered. Therefore, related variables could be on a low level, but also the relationships between variables might be different. During and after the experiment, the variables and relationships were activated and "hot". Before and during the pretest, these variables and relationships were not activated and "cold". Thagard (2006, p. 3) used the concept of "emotional cognition" for classifying "hot thought" as a "thinking that is influenced by emotional factors such as particular emotions, moods, or motivations". The results of the mediation analysis showed that during or at the end of the intervention processes were activated which influenced posttest results. In the pretest situation, such processes were not activated as no intervention was implemented at the point of time. It might be concluded that such active processes change the level and the direction of correlations between variables in the posttest in comparison to the pretest. As evidence for this interpretation, in Table 1 can be seen that there are many significant correlations between variables in the posttest, but many zero-correlations in the pretest.

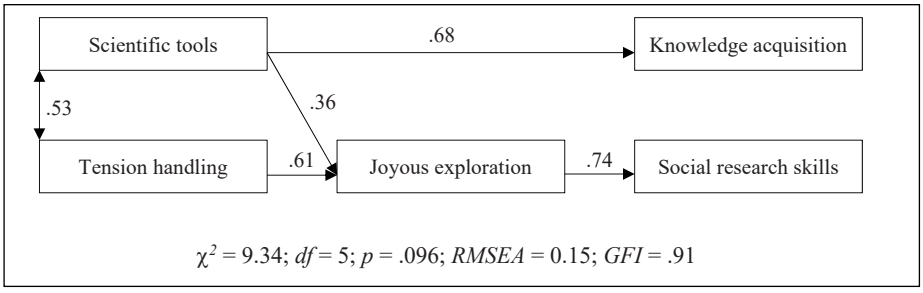

Fig. 1: Mediation analysis on posttest results (n = 40).

Discussions and Implications

In this study, we explored whether self-assessment in combination with reflective writing would improve knowledge acquisition, research skills, and the related motivational and emotional processes of undergraduate students. Our results were positive as we found an expected increase of knowledge acquisition, social research

skills, use of scientific tools, joyous exploration, and emotional tension handling within the intervention group. Within the control group, we found a decrease or stabilization of these variables. Further analyses revealed that knowledge acquisition in the control group was influenced by prior-knowledge. In both groups, the use of scientific tools had an impact on knowledge acquisition, whereas social research skills were affected by joyous exploration which itself was influenced by emotional tension handling and the use of scientific tools. Overall, our study delivers evidence for the positive effects of self-assessment and reflective writing for undergraduate learning and related motivational and emotional processes.

According to our results, using self-assessment with guidance and reflective writing in combination established an effective multidimensional learning support which focuses on cognitive as well as motivational and emotional aspects in learning simultaneously. Such a multidimensional support corresponds with approaches on multiple goals (e.g., learning goals, performance goals, and social reinforcement goals) which undergraduates try to acquire during courses (Valle et al., 2003). A learning support which addresses multiple goals (mainly cognitive goals for self-assessment and non-cognitive goals in reflective writing) covers more individual needs and related development areas than instructional devices which are focusing on single goals. Such a focus on multidimensional aspects seems to be important for learning, as undergraduates' goals and related skills as well as motivational-affective characteristics (e.g., confidence) vary considerably in time and among different clusters of students (e.g., Putwain & Sander, 2016).

An important but unexpected finding was that learning was limited in the control group without self-assessment and expressive writing. Further analysis revealed that prior-knowledge was a crucial factor. Fostering prior-knowledge in courses can be effectively done by "flipped classroom" activities. In such activities, students have to complete preparatory work before attending courses in order to process contents at a deeper level (O'Flaherty & Phillips, 2015). In the past, the flipped classrooms model of instruction was not only related to cognitive learning but also to motivational and social-emotional processes (e.g., Yilmaz, 2017). Therefore, it could be expected for future research that flipped classroom activities could improve joyous exploration and tension handling within a non-intervention scenario.

The statistical and practical significance of the results from an experimental intervention depends on the conditions realized within the control groups. In our control group, we found limited learning and even a decrease in motivational and emotional variables. However, as we tested, overall teaching effectiveness and pretest skills did not differ between control and intervention groups. An explanation for missing effects on motivational and emotional variables might be that traditional courses like our control group do mainly focus on fostering subject-specific knowledge and skills, but not on non-cognitive aspects. Within our groups, we did not consider approaches on the motivational (e.g., the ARCS-model from Keller, 2010) or emotional (e.g., the FEASP-approach from Astleitner, 2000) design of instruction, although such approaches could effectively be integrated in undergrad-

uate courses (e.g., Kim & Keller, 2008). In future undergraduate courses, instructional strategies from such approaches should be implemented, because especially courses on research methods or statistics were found to suffer from motivational and emotional problems like dropout or fear (e.g., Hedges, 2017).

Limitations

Our study was limited in several ways. From a research design perspective, the small sample size might reduce the power of statistical tests and internal validity could suffer from the lack of randomization. However, our results showed strong effects of the intervention which reduces sample sizes for each group needed to attain good power (Lipsey, 1990, p. 143). We had a quasi-experimental design without randomization, but we used manipulation checks, control for bias, and multiple tests to get information on possible validity threats. From a measurement perspective, we had acceptable reliability and validity coefficients.

The problem of pretest bias. Nonetheless, there were large differences between correlations of variables on the pre- and posttest. Pretests were taken about six weeks after the start of the semester, so that certain levels and variabilities of these variables should have been realized. That should allow the finding of conclusive correlations between variables as indicators of validity. This was however not the case in our study. The problem of unstable or changing validity coefficients is well known in social science (e.g., Tisak & Tisak, 1996), but so far, we do not have conclusive explanations for this finding. The results of our mediation analyses revealed that significant learning-related processes were activated by the intervention. We speculated that these processes were not active in the pretest situation. However, future attempts using such or similar measurements might reflect theoretical and methodological conceptions from Roe (2014) on temporal perspectives in validity research.

Astleitner (2018b) suggested several general approaches to explore such problems or bias in measurements:

- First, from a theoretical perspective, it could be beneficial to focus in more detail and based on micro-theories on the relevant variables. Here, especially important is, how participants react on the questions or tasks for measurement and what cognitive and other processes are activated. Another possibility is to consider theories which describe and explain the stepwise time-based development or building of the relevant variables. Finally, it could be helpful to find and apply theories which are capable of explaining different measurement errors. For example, in relation to our study, it has theoretically been explained why the correlations between variables changed dramatically within a few weeks period. An exemplary theoretical explanation could be on "cascading effects" (Patterson, Forgatch, and DeGarmo, 2010) following educational interventions.

- Second, it is obvious, also in our study, that measurements change from situation to situation. Here, there are moments in time, when changes occur more frequently in comparison to less sensible moments. In validation studies, such change-sensitive situations must be identified. Also, the importance of quality criteria for measurements changes over time. For researchers, reliability and validity are always important, but for participants also other criteria like usefulness, or fairness might be important. Finally, measurements are at different situations differently close to participants. In controlled laboratory situations, measurement experiences are different from reality or highly individual introspective experiences. In respect to our study, it could be helpful in future research activities to learn more about the stability or situational variability of such measurements in learning-related situations. A possible research strategy might, for example, be to establish a series of additional repeated measurements. An exemplary option for further research might, for example, be to focus on situation specificity and typical or maximum performance (Patry, 2011).
- Third, another possibility to reduce bias might be not only to measure the relevant variables, but also variables which are, in some way, related to these variables. Having such networks of variables allows to test rival hypotheses. Such tests could deliver information on processes and variables which were not in the original focus, are hidden, might be dark-sided, or represent secondary causes. For example, one might assume that self-efficacy or autonomy are important variables in relation to the motivation of teachers (e.g., Martinek, Hofmann, & Müller, 2018). Others might think, that these variables are less important and outdated, because teachers often have experiences which lead to situations in which they do not significantly develop self-efficacy or autonomy due to the high dynamics and complexities in modern life and classroom management (e.g., Brouwers & Tomic, 2000). What could count more for teacher's motivation and related measurement attempts is about "survival skills" (in conflict resolution, coping with failures, or managing ill-structured learning environments) which are, on the one hand, related to self-efficacy or autonomy in the classroom (e.g., Le Maistre & Paré, 2010). On the other hand, such skills are action-, practice-, or training-orientated what could be much more helpful for teacher's motivation and related measurements as well as interventions than traditional concepts which are attitude-, theory-, and personality-orientated. In relation to our study, also more action-, practice-, or training-orientated measurement approaches could lead to more context-sensitive and therefore more valid measurements.
- Fourth, another widely overlooked issue in measurement and a source for bias is about the polarity or bi-polarity of variables (e.g., Kubzansky, Kubzansky, & Maselko, 2004). This issue concerns questions like, for example, are optimism and pessimism separate constructs or do they represent extreme positive and negative bipolar opposites of the same variable? Answers on such questions play also an important role in the type of measurement scale. For example, Sedlmeier (2006) found that having unipolar versus bipolar scales strongly influenced

participants' ratings and answers. In relation to our study, future research could apply measurements in which important variables are measured as bipolar opposites in comparison to bipolar scales.

Implications for Practice

The findings of this study suggest that higher education activities for undergraduate students should focus more often on the combinations of different learning support approaches. We integrated self-assessment with guidance and reflective writing and established, according to our data, an effective learning environment. Within such an environment, cognitive, motivational and emotional aspects of learning were handled simultaneously realizing a multidimensional support of learning. However, for undergraduate learning, a multidimensional support should not produce an excessive demand on students. Therefore, it is important to acquire more knowledge on conditions when multidimensional support is beneficial or not. Decisions on multidimensional support of learning must become part of approaches on classroom management in undergraduate learning and related university teacher education. We can find classroom management as an important issue in the field of teacher education (e.g., Voss, Wagner, Klusmann, Trautwein, & Kunter, 2017). This is rarely the case in other fields of higher education or of instructional design courses for college teachers. For example, Seeman (2010) related classroom management to preventing disruptive behaviour in college classrooms but not to strategies on decisions about multidimensional learning support. As long as there are no multidimensional approaches on instruction and learning in higher education, a viable way could be to combine, as we did, strategies from different cognitive, motivational, and emotional approaches into course sequences at the same time. In future practical course developments, effective mechanisms of how to adapt such strategies to different learning situations or learner characteristics have to be developed (e.g., Parsons et al., 2018).

References

Arum, R., & Roksa, J. (2011). *Academically adrift. Limited learning on college campuses*. Chicago: University of Chicago Press. doi: https://doi.org/10.7208/chicago/9780226028576.001.0001

Astleitner, H. (2000). Designing emotionally sound instruction: The FEASP-approach. *Instructional Science*, *28*, 169–198. doi: https://doi.org/10.1023/A:1003893915778

Astleitner, H. (2018a). Multidimensional engagement in learning – An integrated instructional design approach. *Journal of Instructional Research, 7, 6–32*. doi: https://doi.org/10.9743/JIR.2018.1

Astleitner, H. (2018b). Das Rahmenkonzept des Grenzübergangs und die Entwicklung und Evaluation von Messverfahren [The conceptual framework of border crossing and the development and evaluation of measurements]. In B. Bütow, J.-L. Patry &

H. Astleitner (Eds.), *Grenzanalysen – Erziehungswissenschaftliche Perspektiven zu einer aktuellen Denkfigur* (pp. 62–78). Weinheim: Beltz Juventa.

Bazerman, C. (Ed.). (2007). *Handbook of research on writing: History, society, school, individual, text.* New York: Erlbaum.

Boals, A. (2012). The use of meaning making in expressive writing: When meaning is beneficial. *Journal of Social and Clinical Psychology*, *31*, 393–409. doi: https://doi.org/10.1521/jscp.2012.31.4.393

Bolton, G. (2010). *Reflective practice. Writing & professional development* (3rd ed.). London: Sage.

Boudah, D. J. (2011). *Conducting educational research: Guide to completing a major project*. Thousand Oaks: Sage. doi: https://doi.org/10.4135/9781483349138

Broadbent, J., Panadero, E., & Boud, D. (2018). Implementing summative assessment with a formative flavour: A case study in a large class. *Assessment & Evaluation in Higher Education*, *43*, 307–322. doi: https://doi.org/10.1080/02602938.2017.1343455

Brouwers, A., & Tomic, W. (2000). A longitudinal study of teacher burnout and perceived self-efficacy in classroom management. *Teaching and Teacher Education*, *16*, 239–253. doi: https://doi.org/10.1016/S0742-051X(99)00057-8

Brown, W.S. (1998). Power of self-reflection through epistemic writing. *College Teaching, 46*, 135–138. doi: https://doi.org/10.1080/87567559809596258

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*, 205–249. doi: https://doi.org/10.1007/s10648-011-9191-6

Cohen-Sayag, E., & Fischl, D. (2012). Reflective writing in preservice teachers' teaching. What does it promote? *Australian Journal of Teacher Education, 37*, 20–36.doi: https://doi.org/10.14221/ajte.2012v37n10.1

Dickman, N. (2005). *Journal writing as a vehicle for reflecting and enhancing learning processes of mathematics teachers in the course of becoming mathematics teacher educators* (in Hebrew, Unpublished doctoral dissertation). Technion – Israel Institute of Technology, Haifa.

Dixon, S. K., & Kurpius, S. E. R. (2008). Depression and college stress among university undergraduates: Do mattering and self-esteem make a difference? *Journal of College Student Development*, *49*, 412–424. doi: https://doi.org/10.1353/csd.0.0024

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59*, 395–430. doi: https://doi.org/10.3102/00346543059004395

Hattie, J. (2015). The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology*, *1*, 79–91. doi: http://dx.doi.org/10.1037/stl0000021

Hedges, S. (2017). Statistics student performance and anxiety: Comparisons in course delivery and student characteristics. *Statistics Education Research Journal*, *17*, 320–336. Retrieved from http://iase-web.org/documents/SERJ/SERJ16(1)_Hedges.pdf. https://doi.org/10.52041/serj.v16i1.233

Holden, G., Barker, K., Meenaghan, T., & Rosenberg, G. (1999). Research self-efficacy: A new possibility for educational outcomes assessment. *Journal of Social Work Education*, *35*, 463–476. doi: https://doi.org/10.1080/10437797.1999.10778982

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International.

Kashdan, T. B., Stiksma, M. C., Disabato, D. D., McKnight, P. E., Bekier, J., Kaji, J., & Lazarus, R. (2018). The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality*, *73*, 130–149. doi: https://doi.org/10.1016/j.jrp.2017.11.011

Keller, J. M. (2010). *Motivational design for learning and instruction. The ARCS model approach*. New York: Springer.

Kieffer, K. M., & MacDonald, G. (2011). Exploring factors that affect score reliability and variability in ways of coping questionnaire reliability coefficients: A meta-analytic reliability generalization study. *Journal of Individual Differences, 32*, 26–38. doi: https://doi.org/10.1027/1614-0001/a000031

Kim, C., & Keller, J. M. (2008). Effects of motivational and volitional email messages (MVEM) with personal messages on undergraduate students' motivation, study habits and achievement. *British Journal of Educational Technology*, *39*, 36–51. doi: https://doi.org/10.1111/j.1467-8535.2007.00701.x

Kim, E. S., & Willson, V. L. (2010). Evaluating pretest effects in pre-post studies. *Educational and Psychological Measurement*, *70*, 744–759. doi: https://doi.org/10.1177/0013164410366687

Kirk, B. A., Schutte, N. S., & Hine, D. W. (2011). The effect of an expressive-writing intervention for employees on emotional self-efficacy, emotional intelligence, affect, and workplace incivility. *Journal of Applied Social Psychology*, *41*, 179–195. doi: https://doi.org/10.1111/j.1559-1816.2010.00708.x

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*, 75–86. doi: https://doi.org/10.1207/s15326985ep4102_1

Kostons, D., Van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction*, *22*, 121–132. doi: https://doi.org/10.1016/j.learninstruc.2011.08.004

Kubzansky, L. D., Kubzansky, P. E., & Maselko, J. (2004). Optimism and pessimism in the context of health: Bipolar opposites or separate constructs? *Personality and Social Psychology Bulletin*, *30*, 943–956. doi: https://doi.org/10.1177/0146167203262086

Kyles, C.R., & Olafson, L. (2008). Uncovering preservice teachers' beliefs about diversity through reflective writing. *Urban Education, 43*, 500–518. doi: https://doi.org/10.1177%2F0042085907304963

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*, 681–718. doi: https://doi.org/10.3102/0034654315627366

Lea, M.R., & Stierer, B. (2011). Changing academic identities in changing academic workplaces: Learning from academics' everyday professional writing practices.

Teaching in Higher Education, 16, 605–616. doi: https://doi.org/10.1080/13562517.2011.560380

Lee, S., & Koszalka, T. A. (2016). Course-level implementation of First Principles, goal orientations, and cognitive engagement: A multilevel mediation model. *Asia Pacific Education Review*, *17*, 365–375. doi: https://doi.org/10.1007/s12564-016-9431-z

Le Maistre, C., & Paré, A. (2010). Whatever it takes: How beginning teachers learn to survive. *Teaching and Teacher Education*, *26*, 559–564. doi: https://doi.org/10.1016/j.tate.2009.06.016

Lipsey, M. W. (1990). *Design sensitivity. Statistical power for experimental research*. Newbury Park: Sage.

MacKinnon, S. L. (2017). "The Curiosity Project": Re-igniting the desire to inquire and transformation through intrinsically-motivated learning and mentorship. *Journal of Transformative Learning*, *4*, 4–21. Retrieved from https://jotl.uco.edu/index.php/jotl/article/view/65

Martinek, D., Hofmann, F., & Müller, F. H. (Eds.). (2018). *Motivierte Lehrperson werden und bleiben: Analysen aus der Perspektive der Theorien der Persönlichkeits-System-Interaktionen und der Selbstbestimmung* [Becoming and staying a motivated teacher: Analyzes from the perspective of personality-system interactions and self-determination]. Münster: Waxmann.

Mathers, C. E., Finney, S. J., & Hathcoat, J. D. (2018). Student learning in higher education: A longitudinal analysis and faculty discussion. *Assessment & Evaluation in Higher Education*, *43*, 1211–1227. doi: https://doi.org/10.1080/02602938.2018.1443202

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology*, *93*, 187–198. doi: https://doi.org/10.1037//0022-0663.93.1.187

Mega, C., Ronconi, L., & De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology*, *106*, 121–131. doi: https://doi.org/10.1037/a0033546

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development*, *50*, 43–59. doi: https://doi.org/10.1007/BF02505024

Morgan, D. L., & Morgan, R. K. (2009). *Single-case research methods for the behavioral and health sciences*. Thousand Oaks: Sage. doi: https://doi.org/10.4135/9781483329697

Moscoso, S. (2003). The within-setting variability of validity in cognitive ability tests. *International Journal of Selection and Assessment*, *11*, 352–355. doi: https://doi.org/10.1111/j.0965-075X.2003.00258.x

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*, 199–218. doi: https://doi.org/10.1080/03075070600572090

O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education*, *25*, 85–95. doi: https://doi.org/10.1016/j.iheduc.2015.02.002

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, *22*, 74–98. doi: https://doi.org/10.1016/j.edurev.2017.08.004

Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., et al. (2018). Teachers' instructional adaptations: A research synthesis. *Review of Educational Research*, *88*, 205–242. doi: https://doi.org/10.3102/0034654317743198

Patry, J. L. (2011). Methodological consequences of situation specificity: Biases in assessments. *Frontiers in Psychology*, *2*, 18. doi: https://doi.org/10.3389/fpsyg.2011.00018

Patterson, G. R., Forgatch, M. S., & DeGarmo, D. S. (2010). Cascading effects following intervention. *Development and Psychopathology*, *22*, 949–970. doi: https://doi.org/10.1017/S0954579410000568

Putwain, D. W., & Sander, P. (2016). Does the confidence of first-year undergraduate students change over time according to achievement goal profile? *Studies in Higher Education*, *41*, 381–398. doi: https://doi.org/10.1080/03075079.2014.934803

Putwain, D., Sander, P., & Larkin, D. (2013). Academic self-efficacy in study-related skills and behaviours: Relations with learning-related emotions and academic success. *British Journal of Educational Psychology*, *83*, 633–650. doi: https://doi.org/10.1111/j.2044-8279.2012.02084.x

Reigeluth, C. M., Myers, R. D., & Lee, D. (2017). The learner-centered paradigm of education. In C. M. Reigeluth, B. J. Beatty & R. D. Myers (Eds.), *Instructional-design theories and models* (Vol. IV, pp. 5–32). New York, Abingdon: Routledge. doi: https://doi.org/10.4324/9781315795478

Roe, R. A. (2014). Test validity from a temporal perspective: Incorporating time in validation research. *European Journal of Work and Organizational Psychology*, *23*, 754–768. doi: https://doi.org/10.1080/1359432X.2013.804177

Roohr, K. C., Liu, H., & Liu, O. L. (2017). Investigating student learning gains in college: A longitudinal study. *Studies in Higher Education*, *42*, 2284–2300. doi: https://doi.org/10.1080/03075079.2016.1143925

Salgado, J. F. (1995). Situational specificity and within-setting validity variability. *Journal of Occupational and Organizational Psychology*, *68*, 123–132. doi: https://doi.org/10.1111/j.2044-8325.1995.tb00577.x

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*, 565–600. doi: http://dx.doi.org/10.1037/bul0000098

Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction*, *16*, 401–415. doi: https://doi.org/10.1016/j.learninstruc.2006.09.002

Seeman, H. (2010). *Preventing disruptive behavior in colleges: A campus and classroom management handbook for higher education*. Lanham: Rowman & Littlefield Education.

Stenlund, T. (2010). Assessment of prior learning in higher education: A review from a validity perspective. *Assessment & Evaluation in Higher Education*, *35*, 783–797. doi: https://doi.org/10.1080/02602930902977798

Stone-Romero, E. F., & Rosopa, P. J. (2010). Research design options for testing mediation models and their implications for facets of validity. *Journal of Managerial Psychology*, *25*, 697–712. doi: https://doi.org/10.1108/02683941011075256

Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge: MIT Press. doi: https://doi.org/10.7551/mitpress/3566.001.0001

Thompson, G., Pilgrim, A., & Oliver, K. (2005). Self-assessment and reflective learning for first-year university geography students: A simple guide or simply misguided? *Journal of Geography in Higher Education*, *29*, 403–420. doi: https://doi.org/10.1080/03098260500290959

Tisak, J., & Tisak, M. S. (1996). Longitudinal models of reliability and validity: A latent curve approach. *Applied Psychological Measurement*, *20*, 275–288. doi: https://doi.org/10.1177/014662169602000307

Tu, W., & Snyder, M. M. (2017). Developing conceptual understanding in a statistics course: Merrill's First Principles and real data at work. *Educational Technology Research and Development*, *65*, 579–595. doi: https://doi.org/10.1007/s11423-016-9482-1

Valle, A., Cabanach, R. G., Núnez, J. C., González-Pienda, J., Rodríguez, S., & Piñeiro, I. (2003). Multiple goals, motivation and academic learning. *British Journal of Educational Psychology*, *73*, 71–87. doi: https://doi.org/10.1348/000709903762869923

Voss, T., Wagner, W., Klusmann, U., Trautwein, U., & Kunter, M. (2017). Changes in beginning teachers' classroom management knowledge and emotional exhaustion during the induction phase. *Contemporary Educational Psychology*, *51*, 170–184. doi: https://doi.org/10.1016/j.cedpsych.2017.08.002

Wittek, A. L., Askeland, N., & Aamotsbakken, B. (2015). Learning from and about writing: A case study of the learning trajectories of student teachers. *Learning, Culture and Social Interaction, 6,* 16–28. doi: https://doi.org/10.1016/j.lcsi.2015.02.001

Yilmaz, R. (2017). Exploring the role of e-learning readiness on student satisfaction and motivation in flipped classroom. *Computers in Human Behavior*, *70*, 251–260. doi: https://doi.org/10.1016/j.chb.2016.12.085

7. Instructional Sensitivity as a Prerequisite for Determining the Effectiveness of Interventions in Educational Research

Alexander Naumann, Stephanie Musow & Michaela Katstaller

Abstract: Student achievement has become a major criterion for evaluating the effectiveness of schooling and teaching. However, valid interpretation and use of test scores in educational contexts require more detailed information about the degree to which the applied test instruments are appropriate to evaluate the intended educational and interventional effects. Instructional sensitivity is the psychometric property of tests or single items to capture effects of classroom instruction. Although instructional sensitivity is a prerequisite for valid inferences on teaching effectiveness, sensitivity is rather assumed than verified in practice. The aim of this chapter is to improve the understanding of instructional sensitivity and its measurement in educational intervention research. Specifically, it first provides an overview of the theoretical framework and relevance of instructional sensitivity. Then, different approaches of measuring instructional sensitivity are outlined and procedures of implementing instructional sensitivity in educational intervention studies are introduced and contrasted with each other. Finally, the role of time spans is discussed and modelling change for short-time and long-time effects in pretest-posttest-follow-up designs is addressed.

Introduction

This chapter aims at embedding instructional sensitivity in the scientific discourse of educational intervention research. Educational intervention research is expected to provide evidence-based insights into the effectiveness of educational measures (e.g., Hascher & Schmitz, 2010). However, evidence-based insights necessitate the availability of instructionally sensitive test instruments for drawing valid conclusions on the effectiveness of educational interventions in schools, higher education, or out-of-school learning activities. Yet, fulfilling such methodological requirements like instructional sensitivity may be challenging in a practice-oriented field like educational intervention research. To foster the methodological foundation of educational intervention studies, we will address the following three issues: (a) the theoretical background and relevance of instructional sensitivity, (b) its measurement, and (c) ways of practical implementation in educational intervention studies.

Throughout the chapter, we will discuss particularities of intervention studies with respect to instructional sensitivity.

Theoretical Background and Relevance of Instructional Sensitivity

While instructional sensitivity received little attention in European countries until recently (Deutscher & Winther, 2018; Naumann, Musow, Aichele, Hochweber, & Hartig, 2019a), the concept has been discussed in the U.S. since the mid-1960s (e.g., Cox & Vargas, 1966). Back then, researchers argued whether traditional item statistics like item difficulty or discrimination were appropriate for selecting items in criterion-referenced testing (e.g., Kosecoff & Klein, 1974). However, the concept has been exposed to essential changes since then. By the end of the 1970s, the main focus shifted from item selection in criterion-referenced testing to issues of validity and test fairness in educational assessments (e.g., Linn & Harnisch, 1981). Essentially, there were two concepts of instructional sensitivity, namely instructional validity and instructional bias. Instructional validity referred to the question to what degree classroom instruction contributes to students' test scores (e.g., Schmidt, Porter, Schwille, Floden, & Freeman, 1983). In contrast, instructional bias referred to differential item functioning for students when they were exposed to different kinds of schooling (e.g., Linn & Harnisch, 1981). Both were seen as essential for drawing inferences on instruction (e.g., Burstein, 1989; Linn, 1983), and consequently, both strands merged in the concept of instructional sensitivity (D'Agostino, Welsh, & Corson, 2007). In 2010, Polikoff defined instructional sensitivity as the psychometric capacity of a test or a single test item of capturing effects of teaching. That is, instructional sensitivity (a) can be seen as a necessary prerequisite for valid test use and interpretation if tests are used for drawing inferences on teaching (Burstein, 1989; Popham, 2007) and (b) can be quantified as a psychometric property of an assessment (Polikoff, 2010). While some researchers have expressed their preferences on whether assessments should be sensitive to the content or to the quality of teaching (e.g., Popham, 2007), today's understanding of instructional sensitivity equally encompasses both aspects of teaching (D'Agostino et al., 2007).

In line with the current *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), instructional sensitivity can be seen as a necessary validity aspect when detecting effects of schooling and teaching. Already in the 1980s, Airasian and Madaus (1983) emphasized the role of instructional sensitivity as an important aspect of construct validity. More specifically, instructional sensitivity was seen as a necessary, though not sufficient requirement for consequential validity (Messick, 1989). Following today's argument-based approach to validity (Kane, 2013), the evaluation of instructional sensitivity provides empirical evidence for a valid use and interpretation of test scores. Unlike other validity aspects such as content or curricular validity aiming at the linkage of tests and items and the intended curriculum, instructional sensitivity refers to the alignment of assessments with the implemented curriculum (Naumann et al., 2019a). Nevertheless, in con-

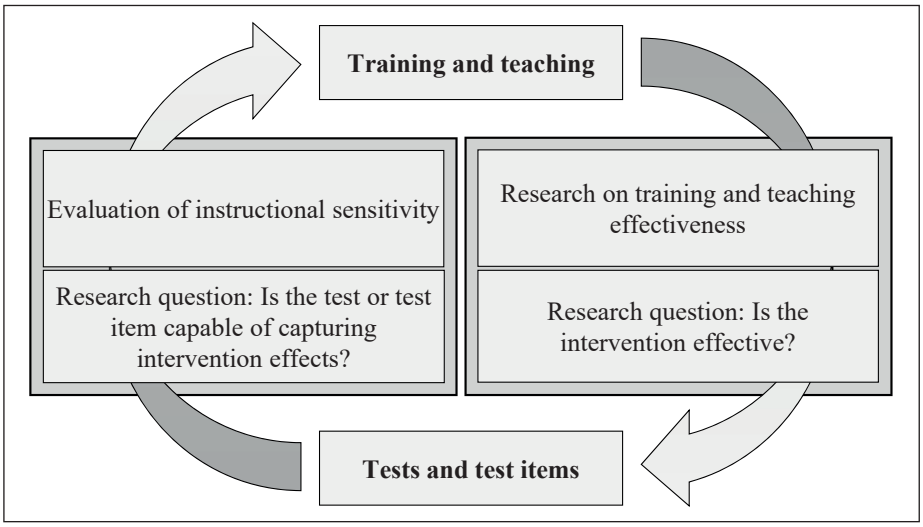

Fig. 1: Relationship of the evaluation of instructional sensitivity and research on education and teaching effectiveness.

trast to the U.S., where discussions of instructional sensitivity have mainly focused on accountability issues (e.g., Popham, 2007), the European discussion puts more emphasis on instructional sensitivity as a central validity aspect in research on educational effectiveness (Naumann, Rieser, Musow, Hochweber, & Hartig, 2019b).

In educational effectiveness research, measures of students' achievement and competencies are the most widespread criteria for evaluating whether or not teaching has been effective (Klieme, 2019). The usual strategy is to use student test scores as dependent variable in (multilevel) regression analyses (Marsh et al., 2012). Applying this strategy requires that the test in principle needs to be instructionally sensitive, that is, capable of capturing effects of teaching. Otherwise, if instructional sensitivity is unclear when evaluating teaching effectiveness, a lack of effects might be either due to ineffective teaching or insensitive assessments (Naumann, Hochweber, & Hartig, 2014; Naumann, Hochweber, & Klieme, 2016). Accordingly, instructional sensitivity needs to be ensured during test development prior to the main effectiveness studies as both explanations remain inextricably confounded otherwise (Naumann et al., 2019a). That is, studies on educational effectiveness have to rely on instruments that are instructionally sensitive to check the degree to which teaching is effective (see right-hand side of Figure 1). However, instructional sensitivity oftentimes is rather assumed than actually investigated empirically (D'Agostino et al., 2007; Naumann et al., 2016).

Two recent studies emphasize practical consequences for inferences on teaching effectiveness stemming from a varying degree of instructional sensitivity. First, Grossman, Cohen, Ronfeldt and Brown (2014) found that tests that operationalize the same construct such as students' achievement in English language arts may show a different extent of instructional sensitivity. Consequently, the test matters whether

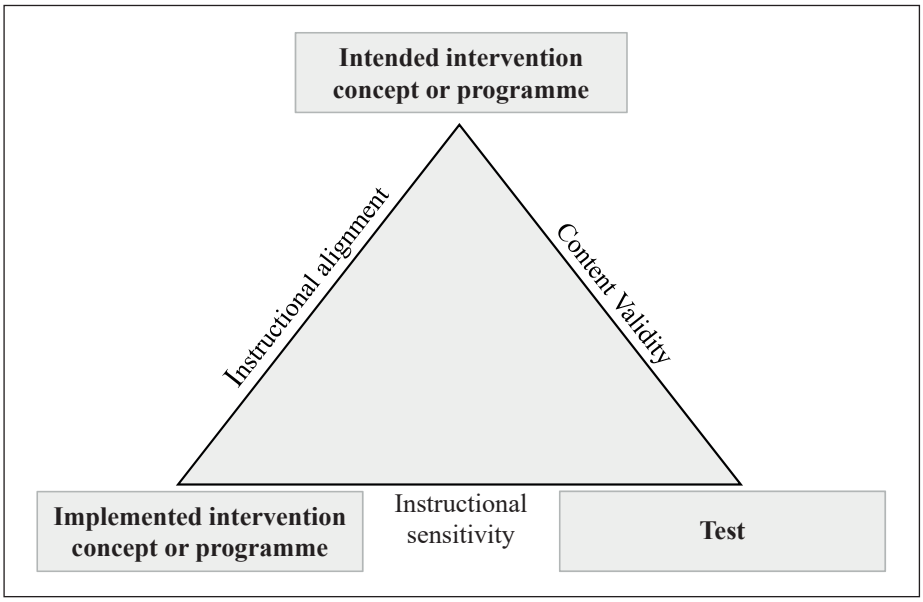

Fig. 2: Ratio of intended intervention concept or programme, implemented intervention concept or programme and test (adapted from Naumann et al., 2019a).

or not certain aspects of teaching are identified to be effective. Second, Naumann and colleagues (2019b) illustrated that a test's instructional sensitivity varies concordantly with the degree of its items' instructional sensitivity. Their results suggest that even slight changes in a test's composition may lead to different conclusions on teaching effectiveness even when the sampled items originate from the same item pool. Taken together, both studies provide empirical evidence that the associations of test scores and the construct(s) of interest, and thus the effect sizes, in an educational effectiveness study depend on the assessment's instructional sensitivity.

The previous considerations correspond to educational intervention research. Students' test scores are also a major criterion for assessing an intervention's success (Hascher & Schmitz, 2010). Accordingly, only if a test and its items are instructionally sensitive, intervention effects – or the lack thereof – can be validly interpreted. Thus, instructional sensitivity may be seen as an essential validity criterion for evaluating an educational intervention to ensure that results are validly interpretable with respect to the intervention's effectiveness. Similar to educational effectiveness research, validly detecting the intervention's effectiveness depends on the extent to which (1) the test itself, (2) the intended intervention concept or programme, and (3) the implemented intervention concept or programme are aligned with each other (Naumann et al., 2019b). Figure 2 depicts the relationship of these three elements as a triad, with each of its sides describing the alignment between two out of the three elements (adapted from Anderson, 2002; Naumann et al., 2019a; Pellegrino, 2002).

A particularity of intervention studies is that both the intended and thus also the implemented curriculum may differ between experimental and control group. The

degree depends on the overall intervention concept. While in some interventions the content may differ completely between experimental and control group (e.g., uninstructed vs. instructed students), there may be an overlap in the intervention content across groups (e.g., same content, but different teaching method; Decristan et al., 2015a). Accordingly, the alignment of intended and implemented intervention concept or programme can either be seen as indicating instructional alignment or treatment adherence within each of the intervention or control conditions, respectively. Treatment adherence is necessary for valid test score interpretation to avoid that the causes of potential effects remain unclear.

Analogously, the alignment of test and intended intervention concept or programme provides arguments for content validity (Hartig, Frey, & Jude, 2012). Empirical evidence for content validity may be given, for instance, by content reconciliation of the test material and formal documents of the intervention concept or programme (Naumann et al., 2019b). If the degree of content validity differs substantially between the experimental and control conditions, valid interpretation of results may be impaired, for example, due to a lack of test fairness. Finally, the actual implemented intervention concept or programme is crucial for the intervention's contribution to students' performance on the test. Accordingly, the alignment of the test and the implemented intervention concept or programme is of special interest, that is, instructional sensitivity. Only if the test is capable of capturing potential intervention effects, results can be validly interpreted. Yet, while the test has to be sensitive to the intervention, it should not favor the intervention conditions compared to the control group. Thus, researchers are required to investigate instructional sensitivity prior to the intervention, for example, in an intervention's pilot study. To provide an understanding of how to achieve this requirement in practice, we will first provide an overview on the measurement of instructional sensitivity hereafter and then propose ways of implementation in educational interventions.

Measuring Instructional Sensitivity

In the course of the last decades, different approaches have been developed to profoundly evaluate instructional sensitivity: (1) item statistics (e.g., Cox & Vargas, 1966; Linn & Harnisch, 1981; Robitzsch, 2009), (2) approaches relating test scores and item responses to instructional measures (e.g., Ing, 2018; Muthén et al., 1995; Ruiz-Primo et al., 2012), and (3) expert ratings (Chen, 2012; Popham, 2007; Popham & Ryan, 2012). Although expert ratings on instructional sensitivity appear beneficial due to economic reasons, they have not been sufficiently evaluated yet. Thus, we will focus on approaches based on actual student tests and item response data in the following section and discuss expert ratings later.

As mentioned before, evaluation of instructional sensitivity should take place prior to the main study to prevent confounding of effectiveness and sensitivity. When evaluating instructional sensitivity, the underlying assumption is that teaching is effective to check whether an instrument is sensitive or not (left-hand side of

Figure 1). That is, instructional sensitivity can be seen as a relational concept that describes the psychometric capacity of a test or a single item of capturing effects of classroom instruction under the condition that teaching is effective (Naumann et al., 2019b).

There are different procedures available for empirical investigation of instructional sensitivity. These procedures can be classified based upon (a) whether they address absolute or relative sensitivity (Naumann, Hartig, & Hochweber, 2017) and (b) their perspective on instructional sensitivity (Naumann et al., 2016). In the following, we will give a brief overview of the resulting framework for measuring instructional sensitivity and how it relates to commonly applied research designs in educational intervention studies.

The Framework for Measuring Instructional Sensitivity

When evaluating the instructional sensitivity of test items, we can distinguish two kinds of sensitivity measures: absolute and relative measures (Naumann et al., 2017). Absolute measures capture an item's overall sensitivity, while relative measures capture the degree to which an item's sensitivity deviates from the test's sensitivity. Absolute and relative sensitivity can be evaluated from each of the three perspectives on instructional sensitivity within the framework. In educational intervention practice, however, absolute measures are usually of more interest.

The three perspectives relate to different variance sources which function as the basis for the investigation of instructional sensitivity (see left-hand side of Figure 3). Naumann and colleagues (2016) label these perspectives (a) the Time Points-Perspective, (b) the Groups-Perspective, and (c) the Time Points- and Groups-Perspective. Each perspective relates to a specific research design that targets the same variance source in the evaluation of the intervention effectiveness (right-hand side of Figure 3).

Time Points-Perspective. The Time Points-Perspective refers to the capacity of a test or an item of differentiating students' learning progress at different points in time. For example, scores of instructionally sensitive tests are expected to increase over time (Baker, 1994). Also, items are expected to get easier over time (Cox & Vargas, 1966). More precisely: An item is considered to be instructionally sensitive, if there is a change in item difficulty between a pretest and a posttest. To investigate instructional sensitivity following a Time Points-Perspective, the Pretest-Posttest-Difference Index (PPDI; Cox & Vargas, 1966) is the most widespread approach. PPDI quantifies instructional sensitivity simply as the difference in item difficulty between posttest and the pretest. With regard to educational intervention studies, the research design underlying this perspective is a one group pre-posttest design.

Groups-Perspective. A second perspective is the groups-perspective (Naumann et al., 2016). This perspective refers to the sensitivity aspect that students should show different performances due to their learning group allocation. To investigate

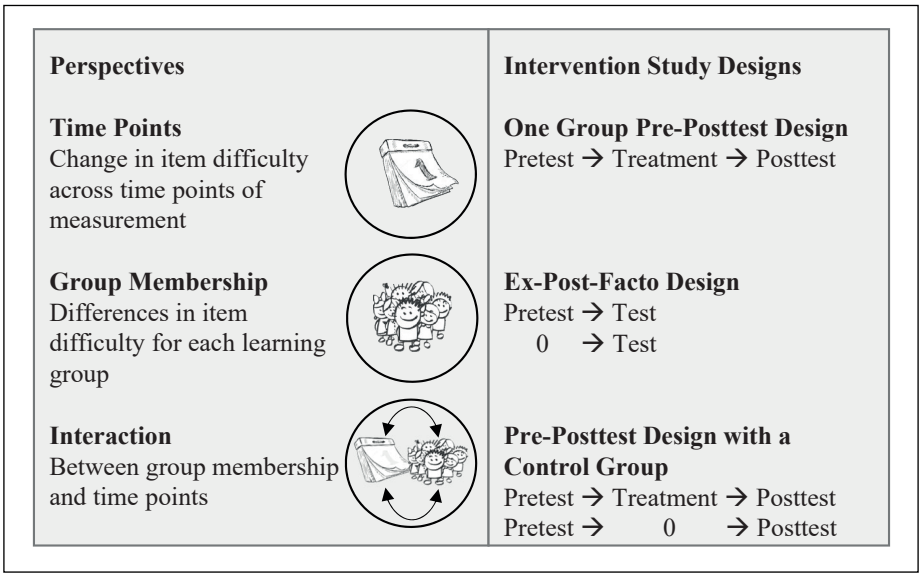

Fig. 3: Three perspectives to evaluate instructional sensitivity and different intervention study designs (Naumann et al., 2016, 2017).

the instructional sensitivity from a Groups-Perspective, analyses of the variation of test scores or item difficulty across groups can be carried out (e.g., Naumann et al., 2017). In educational intervention studies, the research design corresponding to this perspective is a cross-sectional Ex-Post-Facto design. Therefore, assessments in intervention studies are instructionally sensitive, if students' performances in the experimental group significantly differ from the students' performances in the control group.

Time Points and Groups-Perspective. The third perspective refers to the interaction of measurement occasions and learning group membership. Thus, it is called the Time Points and Groups-Perspective (Naumann et al., 2016). The Time Points and Groups-Perspective utilizes information about the group-specific change in test scores or item difficulty respectively. When taking on this perspective, it is optional whether change is modelled as change scores or via covariance-analytic approaches (Naumann et al., 2019b). Accordingly, instructional sensitivity can, for example, be assessed using a longitudinal multilevel Item Response Theory model (LMLIRT; Naumann et al., 2017) or regression of test scores on teaching characteristics while adjusting for prior learning prerequisites (e.g., Polikoff, 2016). In the context of educational intervention studies, the underlying research design corresponds to a pretest-posttest design with a control group. Based on the Time Points- and Groups-Perspective, instruments can be considered as instructionally sensitive in two ways: A test's or an item's (a) global sensitivity describes the average change in item difficulty or test scores across learning groups between measurement occasions, while (b) differential sensitivity indicates the variance of group-specific change in item difficulty or test scores, respectively (Naumann et al., 2016). The

two facets play specific roles in educational intervention studies. The higher the global sensitivity, the higher effect sizes from pretest to posttest can be found in the evaluation of the intervention. If tests or items are capable of differentiating learning progress between the treatment groups, we find indication of differential sensitivity. The higher the instrument's differential sensitivity, the better its capacity to detect specific intervention effects in comparison to the control condition.

Overall, we would like to emphasize that the perspective on instructional sensitivity should fit the research design that is used to evaluate the effectiveness of the intervention. That is, if a test or an item is sensitive from one perspective, it is not necessarily sensitive from another perspective (Naumann et al., 2014). For this reason, it is important to carefully decide whether the analyses are based on Time Points, Groups, or Time Points and Groups-Perspective, respectively. With regard to educational intervention studies, (quasi-)experimental pre-posttest designs with a control group oftentimes are regarded as the gold standard. Therefore, in the following, we will focus on a Time Points and Groups-Perspective to point out ways of ensuring instructional sensitivity in educational intervention studies.

Implementation in Educational Intervention Studies

There are various ways of implementing and ensuring instructional sensitivity in educational intervention studies. The easiest way is to use instruments whose instructional sensitivity has been proven in previous studies (Naumann et al., 2019a). However, a major requirement is that the current study needs to be similar to the research question(s) and the research design of the previous studies. This implies that the new study needs to be comparable with respect to (a) the target population, (b) the intended curriculum, and (c) the assessment design. Otherwise, if previous studies targeted students with learning opportunities differing from those in the current study or provided evidence for instructional sensitivity from a different perspective, it may be hard to assume that the instruments will as well be sensitive in the current study.

A second way of ensuring instructional sensitivity is resorting to expert ratings (Popham, 2007; Popham & Ryan, 2012). Expert ratings on instructional sensitivity are beneficial as they are comparably easy to implement. In principle, expert ratings do not need any empirical student test data or item responses. For this reason, expert ratings appear as a very economical method. However, there are currently only few empirical studies on expert ratings concerning instructional sensitivity (e.g., Chen, 2012; Musow, Naumann, Ruiz-Primo, Hartig, & Hochweber, 2019b). On the one hand, studies found that experts tend to classify more items as sensitive than statistical approaches (Chen, 2012; Musow, Naumann, Hochweber, & Hartig, 2019a; Musow et al., 2019b); on the other hand, recent work by Musow and colleagues (2019a; 2019b) indicates that raters and statistics may coincide depending on the kind of rating and the group of experts recruited. In summary, further research is needed before making a final recommendation.

Lastly, educational interventions may conduct pilot studies for ensuring instructional sensitivity of instruments. Ensuring instructional sensitivity in the context of pilot studies appears beneficial when no instruments whose instructional sensitivity has already been proven are available, and when the aim is to validate instructional sensitivity empirically. Ideally, pilot studies are conducted in samples with similar or at least comparable learning opportunities as the sample of the main study. Then, the aforementioned statistical methods can be applied to the item response data for determining instruments' instructional sensitivity.

In many scenarios, however, sample sizes and/or expertise in elaborate statistical methods may not be sufficient for implementing sophisticated approaches like the aforementioned LMLIRT model. While the LMLIRT model is methodologically sound, its implementation lacks user-friendliness as it requires advanced knowledge in Bayesian estimation and corresponding software packages. Thus, in the following, we will provide a screening procedure that is easy to implement and still allows for the evaluation of instructional sensitivity from a Time Points and Groups-Perspective.

A Screening Procedure for Instructional Sensitivity

Our screening procedure follows comparable methodological principles as the two versions of the LMLIRT model. While the LMLIRT model either utilizes (a) estimates of group-specific change in IRT item difficulty parameters or (b) baseline-adjusted posttest IRT item parameters as a basis for measuring global and differential sensitivity (Naumann et al., 2019b), the proposed screening procedure resorts to Classical Test Theory (CTT) item difficulties. Compared to the latent variable models, the main drawbacks are that the proposed sensitivity measures are purely descriptive, covariance structures between measurement occasions are neglected, and that CTT item difficulties are prone to measurement error. In other words, the observed CTT item difficulties capture an item's true difficulty plus – to some degree – measurement error (e.g., Rost, 2004). The practical advantage is that CTT item difficulties are easy to compute using standard software and applicable in many scenarios that are common to educational intervention studies. Analogous to the LMLIRT model, we will provide two versions of our screening procedure, one suitable for the change-score approach and the other one appropriate for the covariance-analytical approach. The choice of the approach depends on how change will be modelled in the main study. Both versions essentially require three steps for evaluating instructional sensitivity from a Time Points and Groups-Perspective.

Change-Score Approach. The change-score approach requires repeated measurements of the same item. When following the change-score approach, the first step is to calculate CTT item difficulties separately for the treatment and the control condition at pretest and posttest, respectively. If there is a hierarchical data structure with multiple learning groups (e.g., classes) within each condition, we calculate item difficulties for each learning group. Second, we compute the difference in

item difficulty between pretest and posttest for each learning group. Conceptually, this corresponds to group-specific PPDI values. Finally, mean and variation of the group-specific change in item difficulty serve as indicators for absolute global and differential sensitivity. Mean values may range from –1 to 1, with zero indicating that an item is not globally sensitive. Similarly, the higher the variation in group-specific change in item difficulty, the higher the item's differential sensitivity. In the simplest case, when there is only one treatment and one control group, we use the difference in PPDI between the two groups as a measure of differential sensitivity.

Covariance-Analytical Approach. The covariance-analytical approach does not require repeated measurements of the same item, yet, it also does not preclude them. When following the covariance-analytical approach, the first step is to calculate CTT item difficulties at posttest separately for the treatment and the control condition. If there is a hierarchical data structure with multiple learning groups (e.g., classes) within each condition, we calculate item difficulties for each learning group. Second, we regress the group-specific posttest item difficulties on covariates that account for prior learning prerequisites, for example, prior achievement. Then, the residual variance in group-specific item difficulty serves as an indicator for an item's differential sensitivity. If residual variance is near zero, the item under investigation can be considered as not differentially sensitive. In the simplest case, when there is only one treatment and one control group, we use the difference in item difficulty between the two groups as a measure of differential sensitivity. In contrast to the change-score approach, there is no measure of global sensitivity (cf. Naumann et al., 2019b).

Illustrative Data Example. For illustration of the proposed methods, we use data from the study "Individual support and adaptive learning environments in primary school" (IGEL; Decristan et al., 2015b). IGEL was a quasi-experimental intervention study in grade-level three of German primary school science education. More specifically, IGEL was a cluster-randomized controlled trial using a pretest-posttest-follow-up assessment design. Participation was voluntary. First, all participating teachers were trained in the content area of floating and sinking. Then, teachers were assigned to the treatment conditions or the control condition, respectively. Randomization was carried out at the school level. Teachers within the treatment conditions received training in one of three adaptive teaching methods, that is, formative assessment, peer-learning, or scaffolding. Teachers within the control condition received training in parental counseling, which was not expected to show effects on students in the course of the IGEL intervention. After training, teachers implemented the teaching methods in a pre-structured curriculum on floating and sinking in class. The curriculum was adapted from an empirically evaluated primary school inquiry-based science education unit (Hardy, Jonen, Möller, & Stern, 2006; Möller, Jonen, Hardy, & Stern, 2002). It consisted of two consecutive teaching units with five lessons each. The first teaching unit was devoted to the concept of density, while the second one was devoted to the concepts of buoyancy force and displacement. All classes were checked for adherence to the intended curriculum (Adl-Amini, Decristan, Hondrich, & Hardy, 2014). For detailed results regard-

ing the IGEL intervention, see Decristan and colleagues (2015a, 2015b) as well as Hondrich, Hertel, Adl-Amini and Klieme (2016). Our exemplary analyses focus on data from the first teaching unit, as the assessments framing that teaching unit have been extensively investigated for their global and differential instructional sensitivity before, using both the change-score and the covariance-analytic versions of the LMLIRT model (see Naumann et al., 2019b).

The data used for analyses comprises about 1045 students in 54 classes (M_{age} = 8.8 years, SD_{age} = 0.5, 50% female) who participated in the pre- and posttests of students' conceptual understanding of floating and sinking. Students' conceptual understanding served as the main outcome for judging the interventions' effectiveness in fostering students' learning. Corresponding assessments took place with an average time lag of three weeks between pretest and posttest. The tests were administered in classroom-wide assessments by trained personnel. To ensure students' understanding of the tasks, each task was read aloud and visualized using projectors. Then, students had the opportunity to respond to the task. The pretest comprised sixteen items while the posttest consisted of thirteen items, with seven items in common to both measurement occasions. The items were either adapted from previous work done by Hardy and colleagues (2006, 2010), the German TIMSS 2007 science assessment (Bos et al., 2008), or self-constructed. All items were (re)worded to be appropriate for grade level three. Response formats comprised multiple-choice and open-ended tasks. Scoring followed previous research on students' conceptual understanding of floating and sinking (Hardy et al., 2006; Kleickmann et al., 2010). All items fit the partial-credit model (PCM; Masters, 1982).

For our analyses, we split polytomous items into separate dichotomous step indicators. Analyses were carried out using R 3.6.1 (R Core Team, 2019). Markov-Chain-Monte Carlo sampling for the LMLIRT models was conducted via RStan (Stan Development Team, 2019) in a Bayesian framework using vague priors. For details on the technical implementation of the LMLIRT models, see Naumann and colleagues (2017, 2019b). In the covariance-analytic approaches, we adjusted sensitivity measures for students' prior achievement and students' cognitive abilities.

In the change-score approach, calculated group-specific change in CTT item difficulty is highly correlated with latent LMLIRT change estimates, with Pearson correlation ranging from –.95 to –.65 across items (*Mean r* = –.88). Table 1 shows results for items' global and differential sensitivity obtained from the change-score CTT procedure and the change-score LMLIRT model. While the LMLIRT model identifies all repeatedly-administered items as globally sensitive with Bayesian Credible Intervals not comprising zero, the CTT approach seems to identify the items as less globally sensitive. For example, the second step indicator within item 13 appears comparably insensitive. One reason for this finding is that item 13 was very difficult at both measurement occasions, which cannot be captured by the CTT approach in an adequate way. With respect to differential sensitivity, the CTT change-score approach indicates at least some variation in change across groups, while the LMLIRT model identifies more items as differentially sensitive.

Results for the covariance-analytic approach are shown in Table 2. Baseline-adjusted measures of group-specific posttest CTT item difficulty are highly correlated with latent LMLIRT baseline-adjusted estimates, with Pearson correlation ranging from .97 to .79 across items (*Mean r* = .93). Similar to the scenario using change-scores, the CTT approach suggests fewer items to be differentially sensitive compared to the LMLIRT model.

In summary, results support the use of both CTT screening procedures for approximating items' global and differential sensitivity. When comparing LMLIRT and CTT measures of global and differential sensitivity, one has to keep in mind that the measures obtained from the different approaches have different metrics. In the CTT approaches, global sensitivity is expressed in terms of average change between pretest and posttest across groups in the proportion of students who get an item correct, while differential sensitivity describes the degree these proportions vary across groups, expressed in standard deviations. In both cases, the underlying metric is percent correct. In contrast, the LMLIRT models provide measures on a logit scale, with variation across groups expressed as variance. Accordingly, the values from the CTT approaches may appear lower than or even different from those from the LMLIRT models, especially for very easy and very difficult items as they are usually more prone to measurement error.

When screening for instructional sensitivity, we generally recommend excluding such items for which CTT sensitivity measures take on the value zero. However, we do not recommend only selecting items with high global and differential sensitivity values. Depending on the item content, we also recommend considering items with lower sensitivity indices if these items capture hard-to-learn facets of the achievement construct. Nevertheless, we would like to emphasize that the screening procedures may help avoiding the selection of insensitive items, yet they are not ideal for a deeper analysis of the extent of an item's sensitivity.

The Role of Time Spans and how Change is Modelled

Usually when planning an intervention, the question arises whether and to what extent effects are to be expected during a specific period of time. When evaluating instructional sensitivity, a similar question arises: How sensitive are the items in a specific time span? In addition, if data from more than two measurement points are available, there is more than one option to conceptualize change values. To date, there is only little knowledge on the role of time and the ways of modelling it when measuring instructional sensitivity. Yet, when planning an intervention, time plays an important role with regard to the expectation on its effectiveness (Kauffeld, 2010). That is, researchers usually have hypotheses on what intervention effects are expected in which period of time. Consequently, instruments' sensitivity must fit the time span that is covered by the intervention programme.

In addition, pretest-posttest-follow-up design utilize multiple measurement occasions, each associated with specific expectations on effect sizes. When dealing

Tab. 1: Change-score Approach: Item sensitivity results for repeatedly-administered IGEL-items

		LMLIRT model				CTT Screening Procedure	
		Global Sensitivity			Differential Sensitivity	Global Sensitivity	Differential Sensitivity
Item	Cat	M	(SD)	95% BCI	M (SD)		
2	2.1	–4.01	(.16)	[–4.33, –3.70]	.11 (.12)	0.64	0.13
3	3.1	–3.86	(.15)	[–4.16, –3.57]	.06 (.08)	0.65	0.13
4	4.1	–1.83	(.16)	[–2.15, –1.51]	.51 (.21)	0.30	0.20
	4.2	–1.62	(.27)	[–2.13, –1.07]	.52 (.41)	0.13	0.13
5	5.1	–0.99	(.16)	[–1.30, –0.67]	.55 (.26)	0.17	0.19
6	6.1	–0.98	(.14)	[–1.25, –0.70]	.37 (.20)	0.18	0.17
9	9.1	–2.80	(.14)	[–3.08, –2.53]	.15 (.15)	0.50	0.15
13	13.1	–1.94	(.25)	[–2.42, –1.45]	.55 (.33)	0.17	0.15
	13.2	–1.82	(.42)	[–2.65, –1.01]	.62 (.50)	0.06	0.08

Note. M = posterior mean; SD = standard deviation of the posterior mean; BCI = Bayesian credible interval; Cat = score category.

with multiple measurement occasions, there are multiple ways of conceptualizing change values. For example, Steyer, Eid, and Schwenkmezger (1997) distinguish two types of change values in latent variable models: The change values type I imply that values are calculated with reference to the initial value, that is, in a pretest-posttest-follow-up design, the pretest serves as reference. In contrast, change values type II quantify change relative to the nearest measurement points. While Naumann and colleagues' (2017) LMLIRT model is capable of providing both types of measures for group-specific change, the practical implications of the choice of type I and/or type II change values and their appropriateness in different contexts of evaluating instructional sensitivity have not been discussed yet.

To illustrate possible practical implications when modelling change values in educational intervention studies, three prototypical examples with information on the students' performance development of the experimental and control groups are depicted in Figure 4 (adapted from Kauffeld, 2010). In Figure 4, diagram a shows the condition "sensitive items capture effects of the ideal type of intervention", diagram b represents the condition "sensitive items capture effects of a successful intervention", and diagram c shows the condition "sensitive items capture effects of a successful intervention with later development". These three diagrams in Figure 4 underline that depending on the type of change values (type I or type II) chosen, the change values vary differently. Accordingly, it is important to specify whether we are interested in the sensitivity of test items for short-time effects (pretest – posttest) or for long-time effects (pretest – follow-up test, posttest – follow-up test). We thus recommend checking sensitivity of the pretest, posttest and/or follow-up accordingly.

Tab. 2: Covariance-Analytical Approach: Item sensitivity results for all IGEL-items

		LMLIRT Model Differential Sensitivity	CTT Screening Procedure Differential Sensitivity
Item	**Cat**	**M (SD)**	
1	1.1	0.43 (0.14)	0.18
	1.2	0.62 (0.20)	0.16
2	2.1	0.21 (0.13)	0.13
3	3.1	0.07 (0.07)	0.10
4	4.1	0.34 (0.12)	0.18
	4.2	1.19 (0.43)	0.14
5	5.1	0.62 (0.21)	0.18
6	6.1	0.54 (0.19)	0.18
	6.2	1.37 (0.56)	0.11
7	7.1	0.28 (0.11)	0.16
	7.2	0.28 (0.16)	0.12
8	8.1	0.40 (0.17)	0.16
9	9.1	0.24 (0.14)	0.11
	9.2	0.74 (0.29)	0.12
10	10.1	0.26 (0.13)	0.13
11	11.1	0.20 (0.10)	0.13
12	12.1	1.24 (0.40)	0.14
	12.2	1.95 (0.73)	0.12
13	13.1	0.75 (0.28)	0.14
	13.2	0.93 (0.43)	0.08

Note. M = posterior mean; SD = standard deviation of the posterior mean; BCI = Bayesian credible interval; Cat = score category.

Concluding Remarks

In the present chapter, we first provided a brief overview on the concept of instructional sensitivity and then pointed out differences and communalities in its role in educational effectiveness research and educational intervention studies. After presenting common ways of measuring instructional sensitivity, we proposed a screening procedure based on CTT that allows for approximating the absolute instructional sensitivity of single items in situations where more complex approaches are not feasible, for example, when sample sizes are small. Finally, we discussed the role of time lapses in the context of instructional sensitivity. We are confident that the ideas presented in this book chapter help fostering the valid use and interpretation of test scores in the context of educational intervention studies. Again, we would like to point out that the screening procedures presented in this chapter can

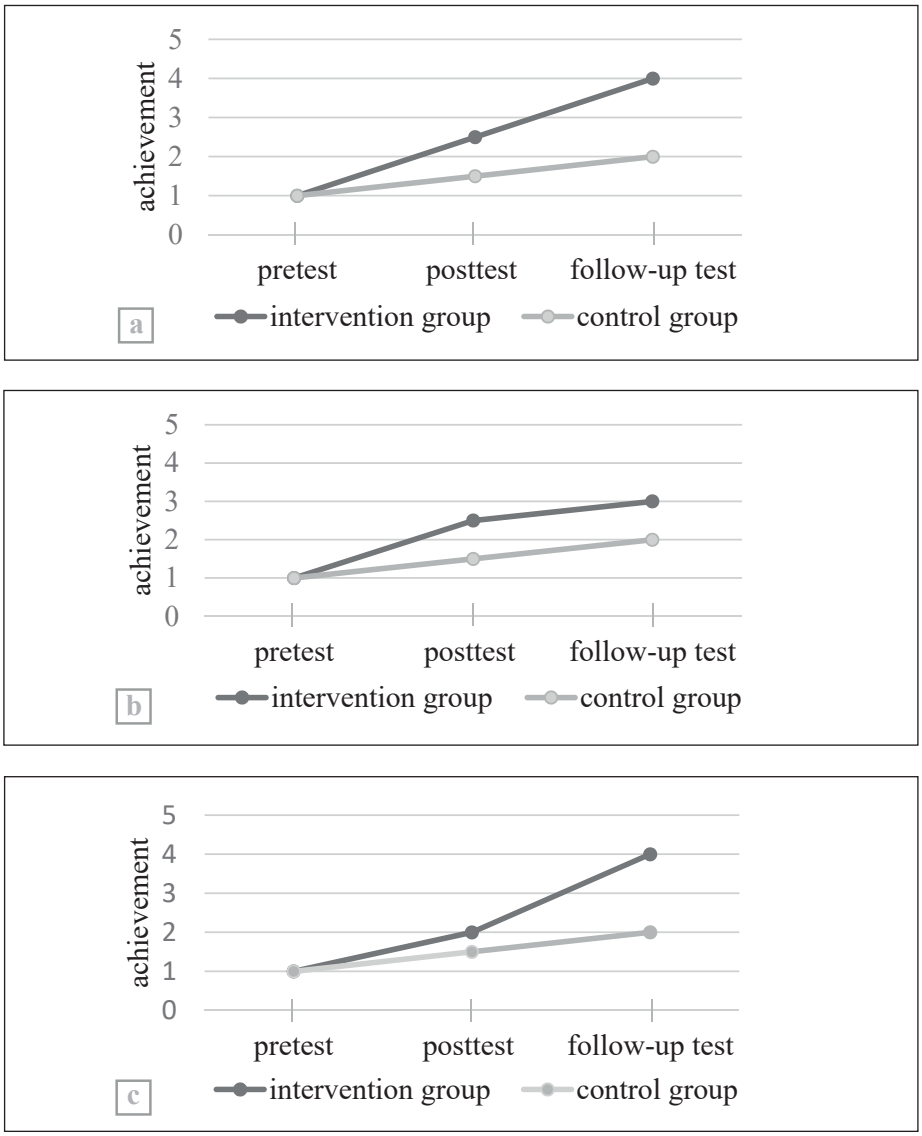

Fig. 4: Ways how sensitive items capture effects of a successful intervention: (a) ideal type of intervention, (b) successful intervention, and (c) with delayed effect (adapted from Kauffeld, 2010).

only help preventing insensitivity. They cannot fully replace a deeper analysis of instructional sensitivity.

Sensitivity of instruments to change due to treatments is regularly discussed in various domains like psychology or medicine (e.g., Benoy et al., 2019; Hays & Hadorn, 1992). Similar to these domains, sensitivity to teaching is oftentimes neglected in educational intervention studies, compromising the validity of inferences drawn from test scores (cf. Burstein, 1989; D'Agostino et al., 2007). In education-

al intervention studies, researchers need to make sure that instruments used for measuring the outcome criteria are capable of detecting potential intervention effects. Hence, measures of instructional sensitivity establish links between student responses and the inferential target and thus serve as validity evidence (cf. AERA et al., 2014; Naumann et al., 2019b). More specifically, information on instructional sensitivity supports (a) the evidence model in evidence-centered design (e.g., Mislevy & Haertel, 2006) or (b) the instructional and inferential facets within Pellegrino, DiBello and Goldman's (2016) validity framework, respectively. Accordingly, without sufficient information on instructional sensitivity, there is no argument supporting a specific instrument's use for measuring the intervention's outcome criteria. Consequently, instructional sensitivity is a necessary prerequisite for the valid use and interpretation of test scores in educational effectiveness research, as well as in educational intervention studies.

Following the psychometric framework by Naumann and colleagues (2017), absolute measures of instructional sensitivity essentially address the reliability of item responses or test scores on the level of learning groups (e.g., classes or schools) with respect to differences between (a) the learning environments students are exposed to (i.e. their learning groups or intervention conditions) or (b) the different stages of learning (i.e. time points of measurement), respectively. However, researchers should not be guided solely by the degree of sensitivity when designing a test, as otherwise effect sizes may increase as a function of item sensitivity (see Naumann et al., 2019b). As a result, inferences on teaching or intervention effectiveness may become invalid if the resulting test is not representative for the underlying task universe. That is, researchers need to clarify which test (Grossman et al., 2014) or which configuration(s) of items (Naumann et al., 2019b) is representative for the desired construct and provides the desired level of instructional sensitivity.

The previous considerations notwithstanding, item selection is not trivial even when information on instructional sensitivity is available. Despite van der Linden's (1981) request for validating instructional sensitivity measures, valid use and interpretation of measures with respect to teaching is still unclear for most of the item sensitivity statistics presented by Polikoff (2010). At best, statistics try approximating influences of learning environments students are exposed to on item responses by using classroom-membership as grouping variable when estimating item parameters (e.g., Robitzsch, 2009). At least partly, the LMLIRT model overcomes this issue as Naumann and colleagues (2019b) were able to provide empirical evidence supporting LMLIRT differential sensitivity measures validity.

Still, it appears hard to define upfront which specific teaching aspect(s) a single item can detect and which not (see also Ing, 2018). Ideally, items represent learnable leaps from one level of sophistication to the next level of sophistication within a domain. As such, they should be sensitive to adequate teaching of content and skills. Yet, while there are strong requests on what tests and items should not be sensitive to (e.g., inherited ability or SES; Popham, 2007), there is no consensus on which specific teaching aspects instruments should be able to capture (Polikoff, 2010). In

our view, the answer to this question largely depends on the purpose(s) of the assessment and the desired test score interpretation. For example, a test that serves as a criterion for judging whether or not a specific facet of teaching quality is effective should be sensitive to the quality of teaching. In educational effectiveness studies that resort to natural variation within a population, tests oftentimes serve for multiple purposes at the same time. Then, operationalizing instructional sensitivity may become all the more complex the more purposes have to be fulfilled, as each intended test score interpretation requires fitting validity evidence in the form of a proof of sensitivity (cf. Kane, 2013). Nevertheless, in the case of educational intervention studies, the purpose of the assessment can usually be expected to be clearly defined. Accordingly, tests should at least be sensitive to those teaching/intervention characteristics whose effectiveness intervention studies are about to judge.

References

Adl-Amini, K., Decristan, J., Hondrich, A.L., & Hardy, I. (2014). Umsetzung von peer-gestütztem Lernen im naturwissenschaftlichen Sachunterricht der Grundschule [Implementation of peer-supported learning in scientific science teaching]. *Zeitschrift für Grundschulforschung, 7*, 74–87.

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington: AERA.

Airasian, P.W., & Madaus, G.F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*, 103–118. doi: https://doi.org/10.1111/j.1745-3984.1983.tb00193.x

Anderson, L.W. (2002). Curricular alignment: A re-examination. *Theory Into Practice, 41*, 255–260. doi: https://doi.org/10.1207/s15430421tip4104_9

Baker, E. L. (1994). Making performance assessment work: The road ahead. *Educational Leadership, 51,* 58–62. Retrieved from http://www.ascd.org/publications/educational-leadership/mar94/vol51/num06/Making-Performance-Assessment-Work@-The-Road-Ahead.aspx.

Benoy, C., Knitter, B., Schumann, I., Bader, K., Walter, M., & Gloster A. (2019). Treatment sensitivity: Its importance in the measurement of psychological flexibility. *Journal of Contextual Behavioral Science, 13*, 121–125. doi: https://doi.org/10.1016/j.jcbs.2019.07.005

Bos, W., Valtin, R., Hornberg, S., Arnold, K.-H., Faust, G., Fried, L., & Schwippert, K. (2008). Zusammenfassung und Schlussfolgerungen [Summary and conclusions]. In W. Bos, R. Valentin, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, L. Lankes, Schwippert, K. & Valtin, R. (Eds.), *IGLU-E 2006. Die Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* [IGLU-E 2006. A National and International Comparison of the Federal States of Germany.] (pp. 143–156). Münster: Waxmann.

Burstein, L. (1989, March). *Conceptual considerations in instructionally sensitive assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Chen, J. (2012). *Impact of instructional sensitivity on high-stakes achievement test items: A comparison of methods* (Unpublished doctoral dissertation). University of Kansas, Lawrence.

Cox, R.C., & Vargas, J.S. (1966, February). *A comparison of item-selection techniques for norm referenced and criterion referenced tests*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.

D'Agostino, J.V., Welsh, M.E., & Corson, N.M. (2007). Instructional sensitivity of a state standards-based assessment. *Educational Assessment, 12*, 1–22. doi: https://doi.org/10.1080/10627190709336945

Decristan, J., Hondrich, A. L., Büttner, G., Hertel, S., Klieme, E., Kunter, M., et al. (2015a). Impact of additional guidance in science education on primary students' conceptual understanding. *The Journal of Educational Research, 108*, 358–370. doi: https://doi.org/10.1080/00220671.2014.899957

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., et al. (2015b). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal, 52*, 1133–1159. doi: https://doi.org/10.3102/0002831215596412

Deutscher, V., & Winther, E. (2018). Instructional sensitivity in vocational education. *Learning and Instruction, 53*, 21–33. doi: https://doi.org/10.1016/j.learninstruc.2017.07.004

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

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking". *Journal of Educational Psychology, 98*, 307–326. doi: https://doi.org/10.1037/0022-0663.98.2.307

Hardy, I., Kleickmann, T., Koerber, S., Mayer, D., Möller, K., Pollmeier, J., Schwippert, K., & Sodian, B. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter [Modeling scientific competence in primary-school age]. *Zeitschrift für Pädagogik, 56*, 115–125.

Hartig, J., Frey, A., & Jude, N. (2012). Validität [Validity]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* [Testing Theory and Design of Questionnaires]. (2nd ed., pp. 143–171). Berlin: Springer. doi: https://doi.org/10.1007/978-3-642-20072-4_7

Hascher, T., & Schmitz, B. (2010). *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* [Educational intervention research. Theoretical basics and empirical action knowledge]. Weinheim: Juventa.

Hays, R.D., & Hadorn, D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research, 1*, 1–73. doi: https://doi.org/10.1007/BF00435438

Hondrich, A. L., Hertel, S., Adl-Amini, K., & Klieme, E. (2016). Implementing curriculum-embedded formative assessment in primary school science classrooms. *Assess-*

ment in Education: Principles, Policy & Practice, *23*, 353–376. doi: https://doi.org/10.1080/0969594X.2015.1049113

Ing, M. (2018). What about the "instruction" in instructional sensitivity? Raising a validity issue in research on instructional sensitivity. *Educational and Psychological Measurement, 78*, 635–652. doi: https://doi.org/10.1177/0013164417714846

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. doi: https://doi.org/10.1111/jedm.12000

Kauffeld, S. (2010). *Nachhaltige Weiterbildung. Betriebliche Seminare und Trainings entwickeln, Erfolge messen, Transfer sichern* [Sustainable training. Developing operational seminars and trainings, measuring success, ensuring transfer]. Berlin: Springer. doi: https://doi.org/10.1007/978-3-540-95954-0

Kleickmann, T., Hardy, I., Möller, K., Pollmeier, J., Tröbst, S., & Beinbrech, C. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter: Theoretische Konzeption und Testkonstruktion [Modeling scientific competence in primary-school age. Theoretical conception and test construction]. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 265–284. Retrieved from http://hdl.handle.net/11858/00-001M-0000-0024-F649-7.

Klieme, E. (2019). Unterrichtsqualität [Quality of instruction]. In M. Gläser-Zikuda, M. Harring & C. Rohlfs (Eds.), *Handbuch Schulpädagogik* [Handbook School Pedagogics]. (pp. 393–408). Münster: Waxmann.

Kosecoff, J.B., & Klein, S.P. (1974, April). *Instructional sensitivity statistics appropriate for objectives-based test items*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.

Linn, R.L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*, 179–189. doi: https://doi.org/10.1111/j.1745-3984.1983.tb00198.x

Linn, R.L., & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, 109–118. doi: https://doi.org/10.1111/j.1745-3984.1981.tb00846.x

Marsh, H.W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A.J.S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106–124. doi: https://doi.org/10.1080/00461520.2012.670488

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. doi: https://doi.org/10.1007/BF02296272

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Phoenix: Oryx Press.

Mislevy, R.J., & Haertel, G. (2006). Implications of evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25*, 6–20. doi: https://doi.org/10.1111/j.1745-3992.2006.00075.x

Möller, K., Jonen, A., Hardy, I., & Stern, E. (2002). Die Förderung von naturwissenschaftlichem Verständnis bei Grundschulkindern durch Strukturierung der Lernumgebung [Fostering scientific understanding of primary school children by structuring their learning environment]. In M. Prenzel & J. Doll (Eds.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwis-*

senschaftlicher und überfachlicher Kompetenzen [Quality of Education at School: Academic and Out-of-School Requirements of Mathematical, Scientific and Interdisciplinary Competencies]. (pp. 176–191). Weinheim: Beltz.

Musow, S., Naumann, A., Hochweber, J., & Hartig, J. (2019a, April). *Multilevel IRT as a validation strategy for expert judgements on instructional sensitivity.* Paper presented at the Annual Conference of the National Council on Measurement in Education, Toronto, CAN.

Musow, S., Naumann, A., Ruiz-Primo, M.A., Hartig, J., & Hochweber, J. (2019b). *Expert judgments – Is it an appropriate approach to evaluate instructional sensitivity?* Manuscript in preparation for publication.

Muthén, B.O., Huang, L., Jo, B., Khoo, S.-T., Goff, G.N., Novak, J.R., & Shih, J. C. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis, 17*, 371–403. doi: https://doi.org/10.3102/01623737017003371

Naumann, A., Hartig, J., & Hochweber, J. (2017). Absolute and relative measures of instructional sensitivity. *Journal of Educational and Behavioral Statistics, 42*, 678–705. doi: https://doi.org/10.3102/1076998617703649

Naumann, A., Hochweber, J., & Hartig, J. (2014). Modeling instructional sensitivity using a longitudinal multilevel differential item functioning approach. *Journal of Educational Measurement, 51*, 381-399. doi: https://doi.org/10.1111/jedm.12051

Naumann, A., Hochweber, J., & Klieme, E. (2016). A psychometric framework for the evaluation of instructional sensitivity. *Educational Assessment, 21*, 1–13. doi: https://doi.org/10.1080/10627197.2016.1167591

Naumann, A., Musow, S., Aichele, C., Hochweber, J., & Hartig, J. (2019a). Instruktionssensitivität von Tests und Items [Instructional sensitivity of tests and items]. *Zeitschrift für Erziehungswissenschaft, 22*, 181–202. doi: https://doi.org/10.1007/s11618-018-0832-0

Naumann, A., Rieser, S., Musow, S., Hochweber, J., & Hartig, J. (2019b). Sensitivity of test items to teaching quality. *Learning and Instruction, 60*, 41–53. doi: https://doi.org/10.1016/j.learninstruc.2018.11.002

Pellegrino, J.W. (2002). Knowing what students know. *Issues in Science & Technology, 19*, 48–52. doi: https://doi.org/10.17226/10019

Pellegrino, J.W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*, 59–81. doi: https://doi.org/10.1080/00461520.2016.1145550

Polikoff, M.S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*, 3–14. doi: https://doi.org/10.1111/j.1745-3992.2010.00189.x

Polikoff, M.S. (2016). Evaluating the instructional sensitivity of four states' student achievement tests. *Educational Assessment, 21*, 102–119. doi: https://doi.org/10.1080/10627197.2016.1166342

Popham, W.J. (2007). Instructional insensitivity of tests: accountability's dire drawback. *Phi Delta Kappan, 89*, 146–155. doi: https://doi.org/10.1177/003172170708900211

Popham, W.J., & Ryan, J.M. (2012, April). *Determining a high-stakes test's instructional sensitivity*. Paper presented at the Annual Conference of the National Council on Educational Measurement in Education, Vancouver, Canada.

R Core Team (2019). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from http://www.R-project.org.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests [Methodological challenges in the calibration of performance tests]. In D. Granzer, O. Köller & A. Bremerich-Vos (Eds.), *Bildungsstandards Deutsch und Mathematik* [Scholastic Standards German and Mathematics]. (pp. 42–106). Weinheim: Beltz.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion [Course book on test theory – test construction]*. Bern: Huber.

Ruiz-Primo, M.A., Li, M., Wills, K., Giamellaro, M., Lan, M.C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching, 49*, 691–712. doi: https://doi.org/10.1002/tea.21030

Schmidt, W.H., Porter, A.C., Schwille, J.R., Floden, R., & Freeman, D. (1983). Validity as a variable: Can the same certification test be valid for all students. In G.F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 133–151). Hingham: Kluwer-Nijhoff Publishing. doi: https://doi.org/10.1007/978-94-017-5364-7_6

Stan Development Team (2019). *RStan: The R interface to Stan. R package version 2.19.2*. http://mc-stan.org/.

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online, 2*, 21–33. Retrieved from https://www.dgps.de/fachgruppen/methoden/mpr-online/issue2/art2/steyer.pdf.

van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of criterion-referenced test items. *Review of Educational Research, 51*, 379–402. doi: https://doi.org/10.3102/00346543051003379

8. How Can Test-taking Motivation Be Theoretically Understood and Measured in Educational Intervention Research?

Michaela Katstaller & Gabriela Gniewosz

Abstract: In educational intervention studies, which are usually based on a pre-test-posttest design (plus follow-up under certain conditions), performance tests are used to investigate whether the intended change in students' performance is achieved. However, the results in such tests can only represent the students' actual performance or change in their actual performance, if the students are willing to give their best in the test. Consequently, the so-called "real" performance or rather the change in students' performance tends to be underestimated. It is particularly problematic for theoretical and practical implications, if the test results are the basis for stakeholders, for instance, to make decisions on curricular changes in schools and universities. Particularly in low-stakes tests, which are mostly used in educational intervention studies, it is suggested to profoundly investigate students' test-taking motivation, its direct and indirect influencing factors, its methodological approach, and its practical implications. The aim of this chapter is to improve the understanding of test-taking motivation in educational intervention research. Specifically, it firstly provides an overview of the construct of test-taking motivation from various theoretical perspectives. Secondly, proximal as well as distal factors that may be related to interindividual differences in test-taking motivation will be presented. Thirdly, methodologically relevant issues such as typical measurement material, the question of "change" and the role of experiments are addressed. Finally, practical suggestions for educational intervention studies are made.

Theoretical Framework of Test-Taking Motivation in Educational Intervention Research

Motivation is a theoretical concept that is utilized to explain human behaviour. It explains individuals' motives to react and fulfill their needs (Elliot, Dweck, & Yeager, 2017; Jacot, Raemdonck, & Frenay, 2015). Motivation-oriented terms such as achievement motivation, test-taking motivation, intrinsic and extrinsic motivation are not theoretically distinct from each other and are therefore often used simultaneously in the context of educational intervention research (Rheinberg, 2006).

According to Skinner, Kindermann, Connell and Wellborn (2009, p. 225) motivation is defined as the "psychological processes that underlie the energy (vigor, intensity, arousal), purpose (initiation, direction, channeling, choice), and durability (persistence, maintenance, endurance, sustenance) of human activity". A challenge in educational intervention studies is that, compared to high-stakes tests (e.g., in admission tests for a study), educational intervention studies are mainly low-stakes tests which initially have no direct consequences for the test-takers, unless they do not pass or score badly (Penk & Schipolowski, 2015). The students only complete a task successfully if they are willing to do so (Rheinberg, 2006) and only the task or the *test-situation itself* can motivate the students to do their best. Therefore, the distinction between domain- and situation-specific motivation (Elliot et al., 2017) seems to be particularly relevant for educational intervention studies: *Domain-specific motivation* refers to an individual's motivation within a certain domain such as a certain subject, while *situation-specific motivation* relates to the motivation in a specific situation such as test-taking situations in which students participate during their school or academic career. It is assumed that domain- and situation-specific motivation are interrelated with each other. This chapter particularly focuses on the students' willingness and engagement to give their best in a situation-specific test-taking situation. It is important to keep in mind that not every effort in a test-taking situation originates in test-taking motivation. If students lack knowledge in their test performance in educational intervention studies, for example, it can be partially compensated with their general intelligence. In contrast, a low level of test-taking motivation can hardly be compensated (Baumert & Demmrich, 2001; Wang, Haertel, & Walberg, 1993).

In contrast to other personality traits such as conscientiousness, for instance, the test-taking motivation of a person is a more dynamic characteristic across situations and measurement points and is affected – among other aspects such as a person's interest in the intervention's topic – by previous experiences. Wise and DeMars (2005) state that test-taking motivation refers to the engagement and the effort a person is willing to show to achieve the optimum result in a test-taking situation. They emphasize that a broader understanding of test-taking motivation has been established including motivational components such as effort, interest, utility, value, and importance. Baumert and Demmrich (2001, p. 441) define test-taking motivation as the "engagement to work on test tasks and to invest effort and perseverance in this project". Thus, test-taking motivation is an important issue in educational intervention research, when it comes to measuring the effectiveness of students' learning outcomes.

As motivation is a rather complex construct, there are several motivation theories that address different motivational aspects. Early theoretical approaches to motivation such as Wilhelm Wundt's or William James' understood motivation as "willing" or "volition" and used methods to explore humans' inner action like desire and natural qualities (Pintrich & Schunk, 2002, p. 21). Later theoretical approaches such as Skinners' theory of operant conditioning (1938) or the drive reduction

theory and human behavior (Hull, 1943) were mainly behaviorally oriented and motivation was considered to be a function of observable events or stimuli from the environment excluding thoughts and feelings (Pintrich & Schunk, 2002, p. 26). Current theoretical frameworks are process-oriented approaches that are able to positively contribute to students' learning outcomes such as the self-determination theory (e.g., Vansteenkiste, Lens, & Deci, 2006), the ARCS model (attention, relevance, confidence, and satisfaction, Keller, 2008), the social cognitive theory (e.g., Wentzel & Wigfield, 2009), and the expectancy-value theory (Eccles & Wigfield, 2002; Eccles, 2005; Wigfield & Eccles, 2000). The latter seems to be most suitable for investigating students' test-taking motivation in educational intervention research, as it is assumed that there is a relationship between the amount of effort put into a task and the performance that can be achieved from the effort. More specifically, the expectancy-value theory proposes that motivation comprises an individual's cognitions that drive actions, and depends on individual, social and contextual factors. Although this theoretical framework is mostly used to explain domain-specific motivation, it can also be appropriately applied to situation-specific test-taking motivation (see Figure 1; right hand side): Test-taking motivation is specified by the extent to which students attribute meaning and value to the intervention study (or test-situation within the intervention study) which they are expected to take part in with their greatest effort. More precisely, the *expectancy component* refers to the students' beliefs about how well they will perform on the forthcoming tasks ("Can I do this task?") and is closely related to Bandura's self-efficacy concept (Wigfield & Eccles, 2000). It is usually divided into *ability beliefs* and *expectancies for success*. However, the ability beliefs are understood to be the students' general beliefs about their competencies in a specific domain and therefore affect the test-taking motivation rather indirectly. Here it becomes clear that domain- and situation-specific motivation are interrelated, due to the fact that the items per se represent a specific content. In contrast, students' expectancies of success conceive students' perception of their likelihood of succeeding in a specific upcoming task and relate to students' test-taking motivation in a narrower sense (Wigfield & Eccles, 2000). The *value component* refers to the extent to which a student values a task ("Why should I do this task?") and comprises the attainment value (importance of doing well in a task), the interest value (interest in the task's content), the utility value (usefulness of a task for individual goals), and the relative cost (perceived negative aspects of engaging in the task). It is proposed that the four value subcomponents affect the students' test-taking motivation, which can act as a potential source of construct-irrelevant variance (Asseburg & Frey, 2013). High test anxiety, for instance, is a factor that might impair students' direct performance (Eum & Rice, 2011). It is characterized by cognitive indicators such as the comparison of self-performance with that of peers, considerations for the consequences of test failure, low confidence in performance, and excessive worries of being tested (Brown & Hirschfeld, 2008; Cheng, Klinger, Fox, & Doe, 2014; Liu, 2008; Putwain, Connors, Woods, & Nicholson, 2012). From this perspective, test anxiety in educational intervention studies may bias the even-

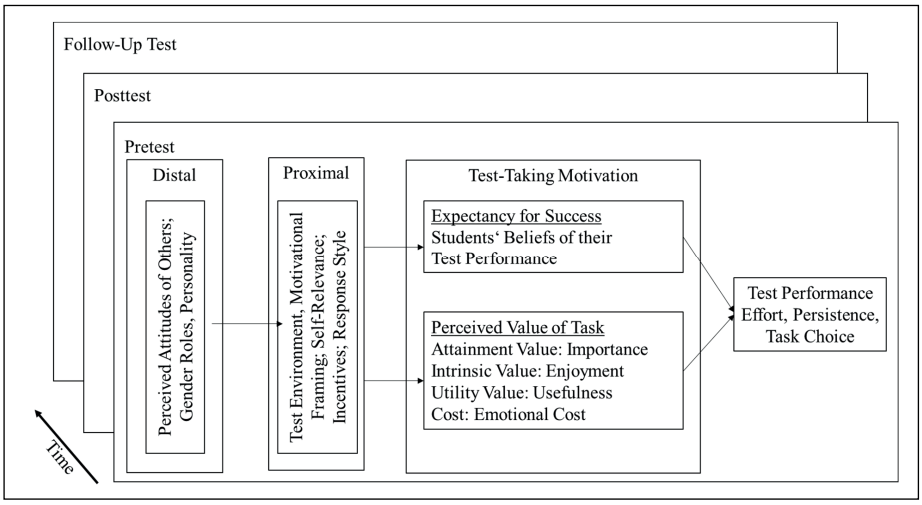

Fig. 1: The expectancy-value theory in the context of test-taking motivation in educational intervention research (adapted from Eccles & Wigfield, 2002; Wigfield & Eccles, 2000).

tual results even in low-stakes tests when the tests become objectively challenging for students (Elkhafaifi, 2005; Larsen-Freeman, 2001; Ellis, 1997; Meece, Glienke, & Burg, 2006).

Both components are necessary to predict not only students' test performance, but also the effort that a student actually puts into the test as well as the persistence a student is willing to invest to overcome more difficult (and frustrating) tasks. Although research findings indicate that students' expectancies better predict students' test performance, persistence and effort than students' values (Schunk, Pintrich, & Meece, 2010), the study of Penk and Schipolowski (2015) is the first study that focuses on both the expectancy- and value-components of the expectancy-value theory and test-taking motivation in a large-scale assessment study of German ninth-graders. It is recommended that both components are to be considered in prospective educational intervention studies in order to investigate changes in students' test-taking motivation and the effects on their test performance in the pretest, posttest, and follow-up test respectively.

Proximal and Distal Factors of Students' Test-Taking Motivation in Educational Intervention Research

Lack of examinees' test-taking motivation can have an impact on the reliability and validity (e.g., increase in construct-irrelevant variance) of the inferences that are drawn from test scores (DeMars, Bashkov, & Socha, 2013). Thus, proximal and distal factors of students' test-taking motivation must be taken into account when interpreting the achieved test results, especially in low-stakes tests, to avoid an under- or

overestimation of students' test performance (Cole, Bergin, & Whittaker, 2008; Eklöf, 2007; Lazowski & Hulleman, 2015; Schiel, 1996; Sundre & Kitsantas, 2004). Our focus is to synthesize findings of those proximal and distal factors that seem to be highly relevant for developing educational intervention programmes.

Proximal factors are considered to be factors that have a direct effect on the students' test-taking motivation or are related to their individual affective conditions. In general, the *consequences of test results* in high-stakes versus low-stakes tests on students' grades or their admission for a study or profession is an important factor that may influence the test-taking motivation of students. The disadvantage of low-stakes tests is that students tend to attribute less importance to tests that do not have negative consequences for them. Thus, they tend to show less willingness to exert themselves in low-stakes test-taking situations (Eklöf & Knekta, 2017). Interestingly, it is not only the question about the actual, but also about negatively perceived consequences while taking a test. If a student does not care about grading in a subject, for instance, he or she will hardly meet the need for conscientiousness and effort in a test-taking situation. Hence, negative perceptions include *subjective beliefs* about testing that are irrelevant to the students (Brown & Hirschfeld, 2008).

Additionally, the *test environment* is often underestimated in test-taking situations, but it seems to be highly relevant, especially in low-stakes tests, if students' performances vary throughout a whole test-taking situation (Penk & Richter, 2017). However, a test-situation may comprise different situational information such as given cues (e.g., physical stimuli or objects), classes (e.g., classified by a date, type) or situational emotions such as "stressful" or "boring" (Rauthmann, Sherman, & Funder, 2015). As test rooms usually have an unalterable format, it is proposed to ensure a friendly atmosphere by providing a good indoor climate and sufficient lighting (Petermann & Macha, 2005). Previous research shows that students have both negative and positive perceptions of the "character" of an assessment (Knekta & Sundström, 2019) due to situational aspects such as the friendliness of the test administrator(s) or individual motivational aspects such as their wellbeing. This may help participants avoid situationally negative feelings like stress, anger, pressure, and boredom (Putwain, 2009).

Research shows that the perceived benefit can be increased by giving motivationally supportive test instructions and by enhancing students' self-relevance of their test-results (Finn, 2015). *Motivationally supportive instructions* aim at explaining the usefulness of taking part in the test by clarifying the purpose, the goals and the required contribution of the test-takers at the very beginning as well as outlining the anticipated consequences of the test results for their school or educational institution. *Self-relevance* may be enhanced, for instance, by issuing a certificate to confirm a student's test performance. Providing feedback on students' test results is another approach to improve students' self-relevance because it helps them get a clue about their current skills in the tested content area. Although it seems to be reasonable to suppose that feedback increases the perceived benefits and subsequently positively affects students' willingness to try their best, there are only a few studies

that show empirical evidence for this assumption (Baumert & Demmrich, 2001; Finn, 2015; Wise & DeMars, 2005, 2010). Additionally, the format of the feedback such as task-specific feedback versus overall feedback or individual score versus the achieved ranking position in relation to a certain group as well as the extent to which the given feedback is considered to be fair and transparent has an effect on students' test-taking motivation and their test performance. Thus, it is sometimes advisable not to give formative feedback to avoid any confounding effects on the students' test performance.

Furthermore, offering *incentives* such as free meals, gift cards or a discount for books can motivate students to sign up for a (voluntary) test. More relevant for students' test-taking motivation are performance-based incentives for the purpose of motivating students to exert themselves (Baumert & Demmrich, 2001; Finn, 2015). Typical incentives are prizes to be won, public recognition, financial rewards for an outstanding performance on the overall test result or for each correctly answered task. Despite of the fact that such incentives can be quite cost-intensive, some studies hint that examinees score somewhat better when the test has a strong personal incentive for them (DeMars, Bashov, & Socha, 2013; Sundre & Kitsantas, 2004; Terry, Mills, & Sollosy, 2008).

Moreover, the *characteristics of the test and/or its tasks* have shown to affect the students' test-taking motivation. While both highly and lowly motivated students perform well in moderately difficult items, only highly motivated students score well on difficult tasks. Thus, the more difficult the items are, the more students' test-taking motivation decreases as more time is needed to work on the tasks (Asseburg & Frey, 2013; DeMars, 2000; Wise & DeMars, 2005, 2010). Consequently, research findings indicate that assessment tasks should be designed with 50% moderate task difficulty (Astleitner, 2006; Pintrich & Schunk, 2002). Additionally, tasks with more text instruction are associated with lower levels of test-taking motivation than multiple-choice items (DeMars, 2000).

Barry, Horst, Finney, Brown and Kopp (2010) have found that the amount of mental taxation to correctly answer a cognitively-demanding task may lead to low expectancies, low value associated with the task (see also Eccles et al., 1983), and may result in low test-taking motivation.

Distal factors are those factors that are indirectly related to students' test-taking motivation, usually via proximal aspects. Not only attitudes held by the test-takers, but also *attitudes expressed by test administrators*, teachers and other students may indirectly influence students' willingness or ability to show a high level of effort and – therefore – test-taking motivation (Lau, Jones, Andersson, & Markle, 2009; Putwain et al., 2012). Therefore, it is useful to look for administrators who have a positive attitude towards the participants, the organizing institution as well as the intervention programme. According to Lau and colleagues (2009), certain behaviour patterns can positively influence motivation before or during the test situation. Depending on the test-taking situation or test design, it is recommended to (a) warmly welcome participants before the test starts, (b) have them introduce

themselves, (c) convey a positive attitude in tone and behaviour, (d) show respect towards the participants, and (e) encourage participants to continue working hard if declining efforts are observed (if allowed).

Effects that are attributed to *gender roles* are frequently investigated in the context of test-taking motivation and test performance. Gender, or rather the gender reported by the students, is a proxy for a variety of differences between male and female test-takers. DeMars, Bashkov and Socha (2013) investigated gender differences in test-taking situations at university. The results show that male students are less likely to exert even the minimal effort to show up for an assigned testing session than female students. Not attending test sessions indicates extremely low levels of test-taking motivation. As gender may affect the psychometric properties of psychological measures such as test-taking motivation (Memetovic, Ratner, & Richardson, 2014), it is suggested to conduct tests of group-level measurement invariance before comparing scores between groups.

Another rather stable factor of the test-taking motivation is the *personality of the students* themselves. Although there is no extensive research on the link between personality measures and test-taking motivation, a few studies have confirmed that personality dimensions, especially with regard to the Big-Five personality traits, are related to test-taking motivation (Ackerman & Kanfer, 2009; Barry et al., 2010). Komarraju and Karau (2005), for instance, identified in their study that the Big-Five personality traits, especially conscientiousness, extraversion and openness, are the main sources of students' test-taking motivation (e.g., Colquitt, LePine, & Noe, 2000; Judge & Ilies, 2002; Kanfer, Ackerman, & Heggestad, 1996; Komarraju & Karau, 2005). Thus, the Big-Five personality traits seem to play an essential role in guiding the students' test-taking motivation in a certain direction (Ross, Rausch, & Canada, 2003). Participants who are conscientious, extroverted, and open are very likely to show stamina and persistence to (successfully) complete a test. However, it should be noted that these are small to medium (indirect) correlations between personality traits and test-taking motivation. In addition, it is not yet clear whether only individual traits (e.g., "only" conscientiousness) or the interaction of different traits (e.g., in terms of personality types) have a strong influence on students' test-taking motivation.

To sum up, a number of different direct and indirect factors appear to influence students' test-taking motivation in educational intervention research. Certainly, a number of other proximal and distal factors are also possible, so that the aspects presented here represent only the most important and noteworthy factors in educational intervention studies. However, a deeper understanding of these factors helps to better understand and interpret students' test performance. In recent years, several methods have been proposed to increase students' willingness to show their best, not only in high-stakes, but also in low-stakes tests. Thus, it is worthwhile to take a methodological look at the possibilities and limitations associated with measuring test-taking motivation.

Measuring Students' Test-Taking Motivation

If students are not motivated to do their best, they often do not achieve the maximum of their possible test performance. Knekta and Eklöf (2015) state that a lack of test-taking motivation often threats the validity of test scores. Consequently, students' low test-taking motivation may lead to construct-irrelevant variance in the test scores and, therefore, the pretests, posttests, and follow-up tests in educational intervention studies may not only measure the actual knowledge or competence of the test-takers, but rather the students' lack of test-taking motivation. This aspect is of particular interest in surveys (e.g., performance measurements) in educational intervention contexts, as these tests do not usually contain any direct consequences or feedback and are relatively unimportant to the participants themselves. For other stakeholders such as teachers, university lecturers and scientists, however, the conclusions drawn from the results in the pretests, posttests, and follow-up tests can be of central relevance, especially when the question of the "effectiveness" of educational intervention study comes into play (Sundre & Kitsantas, 2004; Wise & DeMars, 2005).

For measuring test-taking motivation, different methods such as self-reports, observations, interviews and item response time measures have been used (Wise & Smith, 2016). So far, the most widely used indicator of test-taking motivation is self-report measuring (DeMars et al., 2013). Self-reports typically capture a wider range of examinees' motivation, and thus, the relationship between scores on such measures and test performance has often been studied. The literature search for questionnaires on test-taking motivation shows that two widely used instruments are *the student option survey* (*SOS)* (Thelk, Sundre, Horst, & Finney, 2009) and *the online motivation questionnaire (OMQ)* elaborated by Boekaerts (2002). Although most of these and other measurement instruments (e.g., Baumert & Demmrich, 2001; Butler & Adams, 2007) are based on assumptions of the expectancy-value theory, none of these questionnaires contain all expectancy, value, utility, and cost aspects according to the theory. More precisely, some of them include items that relate to students' expectancies (e.g., O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005; Penk & Schipolowski, 2015), some questionnaires focus on the value aspect with items that refer to importance (e.g., Eklöf & Nyroos, 2013; Thelk et al., 2009), and only some of the available survey instruments operationalize the aspect of utility (O'Neil et al., 2005; Penk & Schipolowski, 2015). Another questionnaire to measure test-taking motivation is the *questionnaire of current motivation (QCM)* developed by Rheinberg, Vollmeyer and Burns (2001). With the use of this questionnaire, Freund, Kuhn and Holling 2011 investigated test-taking motivation of university and secondary school students who performed an abstract reasoning test. The expectancy component was captured by the subscale "probability of success", and the value component was measured by the subscales "challenge", "interest", and "anxiety". All test-taking motivation-related predictors explained 14% of the variance in the students' test performance.

In some studies (e.g., Baumert & Demmrich, 2001; Thelk et al. 2009) students' test-taking motivation was assessed *after* performing the achievement test. In the previously mentioned study, Freund and colleagues (2011) investigated test-taking motivation *before* the students' test completion. With regard to educational intervention studies, it is recommended to measure test-taking motivation measurements in the pre- and posttest (and follow-up test respectively) before, during and after the test-taking situation (Frey, Hartig, & Moosgrugger, 2009). This approach should ensure quantifying a person's motivational condition in a test-taking situation based on the interaction of person- and situation-centered characteristics. Additionally, this procedure allows analyzing both the effect of students' test-taking motivation on their test performance and the effect of students' perceived test performance on their test-taking motivation (Sanchez, Truxillo, & Bauer, 2000).

Even though self-reported questionnaires currently represent the central method for assessing test-taking motivation in educational intervention research, their validity is critically discussed due to their sensitivity to impression management tactics and self-deception (e.g., Donovan, Dwight, & Hurtz, 2003; Ortner & van de Vijver, 2015; Schmitt, Hofmann, Gschwendtner, Gerstenberg, & Zinkernagel, 2015; Viswesvaran & Ones, 1999). The use of behavioral indicators (e.g., observational instruments, response-time measures) is a possible alternative or complement to subjective tests. The idea of these more objective measures was initially proposed by James McKeen Cattell in 1890 with his set of mental tests. Based on this approach of personality assessment, it is assumed that heterogeneous data sources are needed to collect person-related data (e.g., motivation and other personality aspects) including self-report data (Q-data), life indicators (L-data) often derived from observer reports, and objective performance or tests (T-data) (Cattell, 1890; Cattell, 1946; Cattell & Kline, 1977). Compared to gathering subjective data, objective data are mostly collected in a highly standardized test-situation. The scores are not based on self-ratings with regard to the construct of interest, and the aims of the tests are masked or not apparently identifiable (Ortner & Proyer, 2018). However, not all constructs seem to be comparably addressed by these non-subjective methods (L- and T-data). For example, interpersonal behaviour and social variables (e.g., intrinsic motivation, extraversion) are reported to be difficult to assess by standardized objective tests (Pawlik, 2006). Additionally, the variety of objective tests in design and scoring does not allow comparing findings easily with the use of psychometric properties of one test to the other (Ortner & Schmitt, 2014).

Concluding Remarks for Prospective Educational Intervention Research

With the specific focus on educational intervention research, we have learned that test-taking motivation can be a key factor for improving students' learning out-

comes. Therefore, helping students develop optimal test-taking motivation is an important goal, both conceptually and empirically (Lazowski & Hullemann, 2015).

As already discussed in the theoretical overview, test-taking motivation can be viewed from different theoretical perspectives. Up to now, there is no comprehensive theoretical framework that brings together different research traditions and – in addition to the expectation and value components discussed here – includes additionally more learned, but "dispositional-like" and therefore stable motives relevant to explain individual differences in students' test-taking behaviour (e.g., seek success vs. fear failure, seek for control vs. loss of control; see Neyer & Asendorf, 2018; Pintrich & Schunk, 2002). A further under-specified aspect is the role different emotions of test-takers in test-taking situations apart from the value component "costs" play. Human emotions influence interests, effort and performance. Research shows that both positive and negative emotions matter to a range of academic outcomes (e.g., Izard, Stark, Trentacosta, & Schultz, 2008; Pekrun, Frenzel, Goetz, & Perry, 2007; Pekrun & Linnenbrink-Garcia, 2014). However, research has neglected when and why emotions are associated with students' test performance in educational intervention research.

The need for profound scientific research on test-taking motivation in educational intervention research is evident. It is recommended that prospective studies have both a theoretical and an applied perspective on the relationship between test-taking motivation and test achievement as well as test-taking motivation and other variables such as students' interest in the topic of the educational intervention study, students' test-taking emotions and students' response time effort that may be related to a change in students' performance between pretest, posttest and follow-up test. Differences between intervention and control group(s) such as gender, ethnical and social differences are also worthy of systemic investigation. Yet, empirical investigations on students' test-taking motivation have only explored parts of the Eccles and Wigfield's expectancy-value model yet (Eklöf, 2007; Eklöf & Knekta, 2017). For comprehensive theories such as the expectancy value theory, however, it is often not feasible to simultaneously investigate all components of the theoretical framework in a single educational intervention study. Students' test-taking motivation to do well in a test is most likely not only determined by the characteristics of the test-taking situation or the test(s) to be taken, but also by the students' goal orientations, their achievement history in school and at university as well as their personal reasons for performing in one way or another. Thus, many important aspects of the expectancy-value model and their relation to students' motivational dispositions in test-taking situations remain to be profoundly studied in educational intervention research (e.g., Covington, 2000; Deci, Koestner & Ryan, 1999; Eccles & Wigfield, 2002).

As far as practical implications to enhance students' test-taking motivation are concerned, the following aspects seem to be of particular relevance to increase, adjust or control the effects of test-taking motivation: Measures to enhance students' test-taking motivation should be taken into account in educational intervention

research studies because of the potential effects of students' test motivation on the validity of test scores. Firstly, test administrators who accompany the test-taking situations in a study may play an important role in students' test-taking motivation (Putwain et al., 2012; Wise & Smith, 2016). Thus, it is recommended to not only ask the test-takers, but also the test administrators about their motivational attitudes towards the intervention study. Secondly, any result indicating low test-taking motivation questions the potential impact of low test-taking motivation on students' test score validity. Thus, computer- or tablet-based tests should be used to monitor students' response time effort. Filtering students by their response time effort is a useful corrective action for subsequent analyses (Swerdzewski, Harmes, & Finney, 2009) that researchers and practitioners can take to address the problem of students' non-test-taking motivation in educational intervention studies (Wise, 2009). However, this requires a valid way of determining a threshold of what constitutes very low test-taking motivation. Thirdly, there is the idea of monetary rewards to students for improved test scores between two measurement points (pretest – posttest, posttest – follow-up test) on the class or course level. Fourthly, an advisable way to enhance the stakes of the test is to establish a positive and permissive test environment by avoiding an emphasis on competition between the intervention and the control group (Lau et al., 2009; Putwain et al., 2012). Being one of the last to finish the test is often perceived negatively by the students. To overcome this feeling of pressure, the test administrators could provide a recommended time frame for completion of the pretest, posttest or follow-up test. Fifthly, the purpose of low-stakes tests and the intended use of students' test results should always be clearly explained at the beginning of the intervention study (Sessoms & Finney, 2015) because it is assumed that the perception of control in a test-taking situation fosters students' test-taking motivation (Ryan & Deci, 2000) and raises students' positive attitudes towards a test (Urdan & Schoenfelder, 2006). Lastly, although there is little empirical evidence that feedback increases students' test taking motivation (Wise & Smith, 2016; see however Cole et al., 2008), it is recommended to enhance the perceived utility in taking part in an educational intervention study by giving feedback about the strengths and weaknesses and what specific content area they are advised to work on (Finn, 2015; Wise & DeMars, 2005, 2010). Additionally, it could be emphasized that participating in the study offers a great opportunity to practice the tested content.

To sum up, schools and universities should provide test-taking environments in educational intervention research in which these scientific and practical suggestions are directly implemented to foster students' test-taking motivation during educational intervention assessments and improve current methodological approaches to measure and model test-taking motivation. It does not only enhance the validity and interpretability of students' test performance, but also optimizes the testing conditions to profoundly evaluate the effectiveness of educational intervention studies.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and testtaker reactions. *Journal of Experimental Psychology: Applied, 15*, 163–181. doi: https://doi.org/10.1037/a0015719

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*, 92–104. Retrieved from https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2013_20130326/06_Asseburg.pdf.

Astleitner, H. (2006). *Aufgaben-Sets und Lernen. Instruktionspsychologische Grundlagen und Anwendungen* [Task-Sets and learning. Basics and applications based on instructional psychology]. Frankfurt: Lang.

Barry, C. L., Horst, J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 20*, 342–363. doi: https://doi.org/10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441–462. doi: https://doi.org/10.1007/BF03173192

Boekaerts, M. (2002). The online motivation questionnaire: A self-report instrument to assess students' context sensitivity. In P. R. Pintrich & M. L. Maehr (Eds.), *New Directions in measure and methods* (Vol. 12, pp. 77–120). Oxford: Elsevier.

Brown, G. T. L., & Hirschfeld, G. H. F. (2008). Students' conceptions of assessment: Links to outcomes. Assessment in Education: *Principles, Policy & Practice, 15*, 3–17. doi: https://doi.org/10.1080/09695940701876003

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement, 8*, 279–304. Retrieved from https://www.acer.org/files/butler.pdf.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15*, 373–381. doi: https://doi.org/10.1093/mind/os-XV.59.373

Cattell, R. B. (1946). *Description and measurement of personality.* New York: World Book.

Cattell, R. B., & Kline, P. (1977). *The scientific analysis of personality and motivation.* London: Academic Press.

Cheng, L., Klinger, D., Fox, J., & Doe, C. (2014). Motivation and test anxiety in test performance across three testing contexts: The CAEL, CET, and GEPT. *TESOL Quarterly, 48*, 300–330. doi: https://doi.org/10.1002/tesq.105

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*, 609–624. doi: https://doi.org/10.1016/j.cedpsych.2007.10.002

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85*, 678–707. doi: https://doi.org/10.1037/0021-9010.85.5.678

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology, 51*, 171–200. doi: https://doi.org/10.1146/annurev.psych.51.1.171

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*, 627–668. doi: https://doi.org/10.1037/0033-2909.125.6.627

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*, 55–77. doi: https://doi.org/10.1207/s15324818ame1301_3

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment, 8*, 69–82. Retrieved from https://eric.ed.gov/?id=EJ1062839.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the Randomized Response Technique. *Human Performance, 16*, 81–106. doi: https://doi.org/10.1207/S15327043HUP1601_4

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., Midgley, C. (1983). Expectancies, values, and achievement behaviours. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco: Freeman.

Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 105–121). New York: Guilford Press.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132. doi: https://doi.org/10.1146/annurev.psych.53.100901.135153

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*, 311–326. doi: https://doi.org/10.1080/15305050701438074

Eklöf, H., & Knekta, E (2017). Using large-scale educational data to test motivation theories: A synthesis of findings from Swedish studies on test-taking motivation. *International Journal of Quantitative Research in Education, 4*, 52–71. doi: https://doi.org/10.1504/IJQRE.2017.086499

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: Perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education, 28*, 497–510. doi: https://doi.org/10.1007/s10212-012-0125-6

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *Modern Language Journal, 89*, 206–222. doi: https://doi.org/10.1111/j.1540-4781.2005.00275.x

Elliot, A. J., Dweck, C. S., & Yeager, D. S. (Eds.). (2017). *Handbook of competence and motivation* (2nd ed.). New York: Guilford.

Ellis, R. (1997). *Second language acquisition.* Oxford: University Press.

Eum, K., & Rice, K. G. (2011). Test anxiety, perfectionism, goal orientation, and academic performance. *Anxiety, Stress, & Coping, 24*, 167–178. doi: https://doi.org/10.1080/10615806.2010.488723

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series, 2*, 1–17. doi: https://doi.org/10.1002/ets2.12067

Freund, P. A., Kuhn, J. T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences, 51*, 629–634. doi: https://doi.org/10.1016/j.paid.2011.05.033

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests [Effects of adaptive testing on motivation by the example of the Frankfurter Adaptive Concentration Achievement Test]. *Diagnostica, 55*, 20–28. doi: https://doi.org/10.1026/0012-1924.55.1.20

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior* theory (The Century psychology series). New York: Appleton-Century-Crofts.

Izard C., Stark, K., Trentacosta, C., & Schultz, D. (2008). Beyond emotion regulation: Emotion utilization and adaptive functioning. *Child Development Perspectives, 2*, 156–163. doi: https://doi.org/10.1111/j.1750-8606.2008.00058.x

Jacot, A., Raemdonck, I., & Frenay, M. (2015). A review of motivational constructs in learning and training transfer. *Zeitschrift für Erziehungswissenschaft, 18*, 201–219. doi: https://doi.org/10.1007/s11618-014-0599-x

Judge, T. A., & Ilies, R. (2002). Relationship of personality to performance motivation: A meta- analytic review. *Journal of Applied Psychology, 87*, 797–807. doi: https://doi.org/10.1037/0021-9010.87.4.797

Kanfer, R., Ackerman, P. L., & Heggestad, E. D. (1996). Motivational skills & self-regulation for learning: A trait perspective. *Learning and Individual Differences, 8*, 185–209. doi: https://doi.org/10.1016/S1041-6080(96)90014-X

Keller, J. M. (2008). First principles of motivation to learn and e3-learning. *Distance Education, 29*, 175–185. doi: https://doi.org/10.1080/01587910802154970

Knekta, E., & Eklöf, H. (2015). Modeling the test-taking motivation construct through investigation of psychometric properties of an expectancy-value-based questionnaire. *Journal of Psychoeducational Assessment, 33*, 662–673. doi: https://doi.org/10.1177/0734282914551956

Knekta, E., & Sundström, A. (2019). 'It was, perhaps, the most important one' students' perceptions of national tests in terms of test-taking motivation. *Assessment in Education: Principles, Policy & Practice, 26*, 202–221. doi: https://doi.org/10.1080/0969594X.2017.1323725

Komarraju, M., & Karau, S. J. (2005). The relationship between the Big Five personality traits and academic motivation. *Personality and Individual Differences, 39*, 557–567. doi: https://doi.org/10.1016/j.paid.2005.02.013

Larsen-Freeman, D. (2001). Individual cognitive/affective learner contributions and differential success in second language acquisition. In M. P. Breen (Ed.), *Learner contributions to language learning* (pp. 12–24). Harlow: Pearson.

Lau, A. R., Jones A. T., Anderson R. D., & Markle R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education, 58*, 196–217. doi: https://doi.org/10.1353/jge.0.0045

Lazowski, R. A., & Hulleman, C. S. (2015). Motivation interventions in education: A meta-analytic review. *Review of Educational Research, 86*, 1–39. doi: https://doi.org/10.3102/0034654315617832

Liu, M. (2008). An exploration of Chinese EFl learners' unwillingness to communicate and foreign language anxiety. *Modern Language Journal, 92*, 71–86. doi: https://doi.org/10.1111/j.1540-4781.2008.00687.x

Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology, 44*, 351–373. doi: https://doi.org/10.1016/j.jsp.2006.04.004

Memetovic, J., Ratner, P. A., & Richardson, C. G. (2014). Gender-based measurement invariance of the substance use risk profile scale. *Addictive Behaviors, 39*, 690–694. doi: https://doi.org/10.1016/j.addbeh.2013.10.016

Neyer F. J., & Asendorf, J. B. (2018). *Psychologie der Persönlichkeit* [Psychology of the Personality] (6th ed.). Wiesbaden: Springer. doi: https://doi.org/10.1007/978-3-662-54942-1

O'Neil, H. F., Jr., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185–208. doi: https://doi.org/10.1207/s15326977ea1003_3

Ortner, T. M., & Proyer, R. T. (2018). Behavioral and performance measures of personality. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–6). Cham: Springer. doi: https://doi.org/10.1007/978-3-319-28099-8_1281-1

Ortner, T. M., & Schmitt, M. (2014). Advances and continuing challenges in objective personality testing. *European Journal of Psychological Assessment, 30*, 163–168. doi: https://doi.org/10.1027/1015-5759/a000213

Ortner, T. M., & Van de Vijver, F. J. R. (2015). Assessment beyond self-reports. In T. M. Ortner & F. J. R. Van de Vijver (Eds.), *Behavior-based assessment in psychology* (pp. 3–11). Göttingen: Hogrefe.

Pawlik, K. (2006). Objektive Tests in der Persönlichkeitsforschung [Objective tests in personality research]. In T. M. Ortner, R. T. Proyer, & K. D. Kubinger (Eds.), *Theorie und Praxis Objektiver Persönlichkeitstests* [Theory and Practice of Objective Personality Tests]. (pp. 16–23). Bern: Huber.

Pekrun, R., Frenzel, A. C., Goetz, T., & Perry, R. P. (2007). The control value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 13–36). San Diego: Elsevier. doi: https://doi.org/10.1016/B978-012372545-5/50003-4

Pekrun, R., & Linnenbrink-Garcia, L. (Eds.). (2014). *International handbook of emotions in education.* New York: Routledge. https://doi.org/10.4324/9780203148211

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in large-scale assessments. *Educational Assessment, Evaluation and Accountability, 29*, 55–79. doi: https://doi.org/10.1007/s11092-016-9248-7

Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences, 42*, 27–35. doi: https://doi.org/10.1016/j.lindif.2015.08.002

Petermann, F., & Macha, T. (2005). *Psychologische Tests für Kinderärzte [Psychological tests for paediatricians]*. Göttingen: Hogrefe.

Pintrich, P. R., & Schunk, D.H. (Eds.). (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Englewood Cliffs: Merrill-Prentice Hall.

Putwain, D. W. (2009). Assessment and examination stress in key stage 4. *British Educational Research Journal, 35*, 391–411. doi: https://doi.org/10.1080/01411920802044404

Putwain, D. W., Connors, L., Woods, K., & Nicholson, L. J. (2012). Stress and anxiety surrounding forthcoming standard assessment tests in English schoolchildren. *Pastoral Care in Education, 30*, 289–302. doi: https://doi.org/10.1080/02643944.2012.688063

Rauthmann, J. F., Sherman, R. A., & Funder, D. C. (2015). Principles of situation research: Towards a better understanding of psychological situations. *European Journal of Personality, 29*, 363–381. https://doi.org/10.1002/per.1994

Rheinberg, F. (2006). *Motivation* [Motivation]. (6th ed.). Stuttgart: Kohlhammer.

Rheinberg, F., Vollmeyer, R., & Burns, B.D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen [FAM: A questionnaire to measure actual motivation in learning and achievement situations]. *Diagnostica, 47*, 57–66. doi: https://doi.org/10.1026//0012-1924.47.2.57

Ross, S. R., Rausch, M. K., & Canada, K. E. (2003). Competition and cooperation in the five-factor model: Individual differences in achievement orientation. *The Journal of Psychology, 137*, 323–337. doi: https://doi.org/10.1080/00223980309600617

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78. doi: https://doi.org/10.1037/0021-9010.85.5.739

Sanchez, R. J., Truxillo, D. M., & Bauer, T. N. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *Journal of Applied Psychology, 85*, 739–750. doi: https://doi.org/10.1037/0021-9010.85.5.739

Schiel, J. (1996). *Student effort and performance on a measure of postsecondary educational development* (ACT Rep. No. 96-9). Iowa City: American College Testing Program. https://doi.org/10.1037/e427342008-001

Sessoms, J., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing, 15*, 1–33. doi: https://doi.org/10.1080/15305058.2015.1034866

Skinner, B.F. (1938). *The behavior of organisms: an experimental analysis.* New York: Appleton-Century.

Skinner, E. A., Kindermann, T. A., Connell, J. P., & Wellborn, J. G. (2009). Engagement and disaffection as organizational constructs in the dynamics of motivational development. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 223–245). New York: Routledge.

Schmitt, M., Hofmann, W., Gschwendtner, T., Gerstenberg, F., & Zinkernagel, A. (2015). Assessments beyond self-reports. In T. M. Ortner & F. J. R. van de Vijver (Eds.), *Behavior-based assessment in psychology. Going beyond self-report in the personality, affective, motivation, and social domains* (pp. 29–44). Boston: Hogrefe.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2010). *Motivation in education: Theory, research, and applications* (3rd ed.). London: Pearson Education.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*, 6–26. doi: https://doi.org/10.1016/S0361-476X(02)00063-2

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*, 162–188. doi: https://doi.org/10.1080/08957347.2011.555217

Terry, N., Mills, L., & Sollosy, M. (2008). Student grade motivation as a determinant of performance on the business major field ETS exam. *Journal of College Teaching & Learning, 5*, 27–32. doi: https://doi.org/10.19030/tlc.v5i7.1244

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters using the student opinion scale to make valid inferences about pupil performance. *The Journal of General Education, 58*, 129–151. doi: https://doi.org/10.1353/jge.0.0047

Urdan, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology, 44*, 331–349. doi: https://doi.org/10.1016/j.jsp.2006.04.003

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210. doi: https://doi.org/10.1177/00131649921969802

Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist, 41*, 19–31. doi: https://doi.org/10.1207/s15326985ep4101_4

Wang, M. C., Haertel, G., & Walberg, H. J. (1993). Synthesis of research: What helps students learn? *Educational Leadership, 51*, 74–79.

Wentzel, K. R., & Wigfield, A. (Eds.). (2009). *Handbook of motivation in school.* New York: Routledge. doi: https://doi.org/10.4324/9780203879498

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68–81. doi: https://doi.org/10.1006/ceps.1999.1015

Wise, S.L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*, 152–166. doi: https://doi.org/10.1353/jge.0.0042

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1–17. doi: https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment, 15*, 27–41. doi: https://doi.org/10.1080/10627191003673216

Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204–220). London: Routledge.