Rotterdam School of Management
Erasmus University

Izaak Dekker

Academic Thriving: Optimising Student Development with Evidence-Based Higher Education

RSM - a force for positive change

Academic Thriving

Optimising Student Development with Evidence-Based Higher Education

Academic Thriving
Optimising Student Development with Evidence-Based Higher Education

Academisch floreren
De ontwikkeling van studenten optimaliseren
met evidence-based hoger onderwijs

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

2 June 2022 at 15:30 hrs
by

Izaak Dekker
born in Amsterdam

Erasmus University Rotterdam

Doctoral Committee

Promotor:	Prof. dr. M.C. Schippers
Other members:	Prof. dr. M. Meeter Prof. dr. A.A.C.M. Smeets Prof. dr. M.L.L. Volman
Co-promotors:	Dr. E. Klatter Dr. E.J. Van Schooten

ISBN 978-90-5892-619-7

Cover Photo: Arif Riyanto

MIX
Papier van verantwoorde herkomst
FSC® C022586
www.fsc.org

Table of Contents

For Katinka

General Introduction

Different Perspectives on College Success

The degree to which students have access to higher education, and successfully obtain a degree (within a certain time), is a worldwide concern (OECD, 2019; UNESCO, 2017; Vossensteyn et al., 2015). Higher education is generally considered a public good which is, partly or even completely, publicly funded. Among other goods, higher education prepares students for professions and economic productivity, it can teach powerful deliberative capabilities, or challenge students' opinions and assumptions (Brighouse & McPherson, 2015).

In countries such as the United States or the United Kingdom, access to higher education is limited by high tuition fees and strict admission policies, but dropout rates are relatively low (OECD, 2019). Other countries, such as the Netherlands and Denmark, have low tuition fees and relatively loose admission policies, but also relatively lower study progress rates and higher dropout rates (Vossensteyn et al., 2015).

From the perspective of educational policy, the Dutch government and universities tried to optimize study progress and graduation rates through a range of different policy measures. Some of these measures penalized universities by cutting their budget when they did not meet targeted retention and graduation rates (Jongbloed et al., 2020). Other measures included limiting grants and introducing fees for students who studied longer than expected, and academic probation for students who did not pass a certain amount of courses in their first year. Although this approach might improve the urgency for administrators, professionals and students to increase study outcomes, it also risks negative externalities such as deteriorating mental health (Auerbach et al., 2018), higher dropout rates (Sneyers & De Witte, 2017; 2018), and lower intrinsic motivation among students (Deci et al., 1999).

From the perspective of the educational professionals, who design and offer higher education courses, study progress hinges around 'learning outcomes'. The courses of study that colleges offer are made up of materials, classes and exams that target specific learning outcomes or 'educational goods' (Brighouse et al., 2016). Depending on the domain and type of university, these outcomes can range from theoretical knowledge of mathematical equations, practical skills such as administering a vaccine, or more abstract outcomes such as critical thinking. Based on some form of evaluation, students receive a grade, pass courses, and eventually graduate with a diploma or dropout without one. Passing a course or graduating is an estimation for whatever learning outcomes the educational professionals wanted to get across.

From the perspective of the students, however, obtaining a degree as soon as possible in order to bolster the economy, or learning the outcomes that the professionals deem important, might not be their (sole) consciously internalized aims in life. Students can strive for (for example) general self-development, societal pursuits, or do not yet know what their aim will become. Furthermore, the domain of the university and professionals often will not be the only important domain during this phase of their lives. Other domains could equally demand attention, such as student employment, taking care of family members, fitting in socially, or 'simply' remaining healthy or sane. The different and potentially conflicting demands can lead students to underinvest time and energy into their study.

The different perspectives complicate researching college success as a scientific construct. One could, of course, limit the scope of a study to one perspective (e.g., only academic performance, learning outcomes, or mental health). Yet, this leaves the responsibility of integrating the different perspectives to politicians, school leaders, professionals and (to a lesser degree) students. Additionally, it obscures the different effects that a single policy can have on both academic performance as well as 'side-effects' (Zhao, 2017) such as well-being. It would, therefore, seem to be a worthwhile endeavour to conduct educational research that measures both study outcomes in terms of grades and progress (academic performance), as well as other outcomes that are relevant during this phase of students' lives.

Academic Thriving

The existing concepts that could be used to describe a definition of success that includes multiple perspectives, have their particular focus and associations. College success, study success, academic achievement, study progress, academic performance, or student well-being, are useful constructs, but too one-sided to cover the abovementioned intention. 'Student success' comes close. Kuh et al., (2005, p. xiv) defined it as "academic achievement; engagement in educationally purposeful activities; satisfaction; acquisition of desired knowledge, skills, and competencies; persistence; and attainment of educational objectives". Yet, in reality student success is measured as a combination of academic performance, persistence to graduation, and equity (Schreiner, 2010, Baldwin et al., 2011; Chang et al., 2019; Fingerson & Troutman, 2020). Schreiner (2010) proposed using the concept of 'thriving' to stand for a more holistic view on success: "Thriving college students not only are academically successful, they also experience a sense of community and a level of psychological well-being that contributes to their persistence to graduation and allows them to gain maximum benefit from being in college" (p. 4). Thriving stems from human flourishing or 'eudaimonia', a term used by Aristotle, well over two millennia ago. Aristotle observed that -according to him- all forms of live seemed to be goal- or purpose oriented (telos). The seed of a tree has an innate purpose, which is fulfilled when it has grown towards it fullest version. The new-born has a predefined purpose, written into its biological composition. To Aristotle these goals were matters of fact that helped explain our innate conditions for flourishing (eudaimonia). The biological goals that define human beings, which can be summarised as physical maturity and reproduction, are present in us, just as with most mammals. On top of that, Aristotle observed that humans have additional goals that define whether they are flourishing specifically as human beings. Humans are 'zoön politikon', they are political creatures, and in order to flourish they need to also fulfil their social and political goals. This integrated concept has remained central to many scholars who reflect on the aims of education (e.g., Brighouse et al., 2016; Kristjánsson, 2016; Wolbert et al., 2018).

Academic thriving, as a specified type of human flourishing, is limited in scope to the period during which students engage in postsecondary education. In this thesis, it specifically stands for the degree to which students obtain learning outcomes, pass grades, and manage to simultaneously balance the other life domains during their time in college; e.g., student employment, well-being, and health. Aristotle argued that eudaimonia should be seen as a continuous process of self-improvement rather than a level that is completed. Similarly, I think that researching what contributes to academic thriving can be an aspiration. Research into academic thriving can be a reasonable combination of academic outcome measures and measures from other life domains that are expected to be interrelated or conflicting.

Evidence-Based Education

Research into contributors to- and sources of- academic thriving can partly be seen as a contribution to evidence-based education (EBE). According to Davies (1999), EBE entails a combination of 1) the capacity and discipline of educators to pose answerable questions about education, know where to find evidence, weigh the power and relevance of the evidence to their educational needs and environments, 2) the power to establish sound evidence where existing evidence is lacking or uncertain. Studying academic thriving ideally means establishing sound evidence on the effects and effectiveness of educational interventions on academic outcomes as well as other relevant outcomes in interrelated life domains. However, studying academic thriving also accommodates some of the criticism that was raised against the narrow focus of EBE on certain types of outcome measures (e.g., Biesta, 2010; 2015). In order to position the studies from this thesis, the first chapter will critically assess the different types of arguments that were recently raised against EBE and its (over)reliance on experimental study designs. Based on this assessment, three types of educational research are proposed that could commensurate the criticism: study local factors, mechanisms and implementation fidelity in RCTs; use (and improve) the available educational longitudinal data; use more combined interventions and outcome measures. The second chapter of the thesis will build on this approach and apply the first type of research method to study an intervention that could contribute to academic thriving.

Goal Setting

Interventions that aim to contribute to academic thriving should be directed at the specific challenges that students face during their time in college. The transition to college most often occurs during a phase which typically includes many life-events and new responsibilities: moving out of your parents' house, paying your own taxes, combining work with studying, making new friends, being legally allowed to vote, drive a car, and consume alcohol (although not simultaneously). From an academic perspective it often involves a transition to a type of educational context that expects more self-regulated learning (Vosniadou, 2020) and offers more specialized learning content. Given this specific context, interventions that aid students with reflecting on priorities and with devising plans that will align their behavior and new habits to match their priorities, seem particularly relevant.

In 2010, Morisano et al. published a study about a 'package' intervention (based on a program developed by Peterson and Mar [2004]) designed to simultaneously influence several variables related to goal pursuit across different domains of life. In their small-scale trial with 85 Canadian university students, Morisano et al. (2010) found that students assigned to the goal-setting intervention obtained a significantly higher GPA and dropped out less than the students in the control condition who made a control assignment with intervention-quality face validity. In addition to these academic performance indicators, students in the treatment group also reported significantly higher 'affect', measured with items like 'are you more generally satisfied with life?' The focus of the intervention on setting personal goals targeted at various domains of life, and the combination of different types of outcome measures, renders this an intervention and a study that contribute to academic thriving. Since that first trial, several other studies researched the potential benefits of this type of package goal-setting interventions on different and larger samples of university students (Dobronyi et al., 2019; Schippers et al., 2015; 2020; Travers et al., 2015).

Schippers et al. (2015) studied the effects of the intervention on a cohort of business school students with a time-lagged quasi experimental design. Students in the cohort that received the intervention obtained 22% more study credits and dropped out 20% less. They found specific benefits for males and students from ethnic minorities (improvements of up to 50%), which contributed to closing the achievement gap that is often reported. Using additional content analyses, Schippers et al. (2020) found that participating in the intervention on average related to increased performance, regardless of whether students picked an academic goal in their top 3 of most important goals. They also found that the quality and quantity of the plans related to more obtained credits. However, when Dobronyi et al. (2019) used a large-scale experimental design to measure effects of the goal-setting intervention developed by Peterson and Mar and used by Schippers et al. (2015; 2020), they found no evidence of any effect on obtained credits or dropout.

The relatively limited number of studies thus far contributed to goal-setting theory raise at least four important questions. The Morisano et al. (2010) RCT was small-scale, Schippers et al. (2015) used a large sample but quasi-experimental design. The findings of Dobronyi et al. (2019) question whether the initial findings of the small-scale RCT from Morisano et al. (2010) and the large-scale quasi experiment from Schippers et al. (2015) are replicable and scalable. More rigorous effect studies are required in order to develop a sound evidence-base for policymakers and practitioners.

By applying the principles of replication with variation (Locke, 2015), replication studies could additionally search for moderators or mediators that help explain how and why the intervention can work under which conditions.

A third question, is whether positive effects on affect as well as performance can be found. Morisano et al (2010) studied these outcome measures, but Schippers et al. (2015; 2020) and Dobronyi et al. (2019) do not report any 'affect' related outcomes. Schippers (2017) and Schippers and Ziegler (2019), however, do predict that this type of goal-setting intervention improves well-being in addition to performance.

The fourth question that yet remains unanswered, is whether the goal setting intervention can produce positive effects across different domains of higher education. The samples thus far are composed of predominantly business school and economics students. The fifth question relates to the follow-up on the goal-setting intervention. Schippers et al. (2015; 2020) and Dobronyi et al. (2019) report using respectively goal-diaries and academic reminders as a follow-up to the initial intervention, but also report contrasting results about their added value. Are these reminders required? What would constitute the optimal form of follow-up, if any?

Student Employment

Dutch universities of applied sciences, more so than research universities, place an emphasis on preparing students for a specific profession. The curriculum integrates several internships during which students accrue actual experience in their future vocation. From the perspective of academic thriving, this potentially leads to an interesting dynamic, because the goals and demands from the domains of the university and the jobs that students might have during their study could enhance each other or become a source of goal-conflict. Many students who engaged in the goal-setting intervention indeed report goals related to becoming such a professional as well as goals related to academic achievement. How the benefits and demands of different types of student employment affect study progress throughout the course of college is a topic that needs more research (Tight, 2021). The field of teacher education in The Netherlands is of particular interest from this perspective because pre-service teachers are frequently offered a teaching position before graduation. The fourth chapter reports the results from a longitudinal study into the effects of different types of student employment on study progress.

Outline of the Dissertation

This dissertation is structured into four chapters that aim to explore what could contribute to both academic outcomes as well as other relevant outcomes in life domains of students, in other words: 'academic thriving'.

The first chapter explores if and how evidence-based education can contribute to academic thriving. Many critics from within educational science recently criticized the epistemic, economic and normative foundation of evidence-based education and its preference for randomised controlled trials. This chapter weighs the arguments against EBE and proposes three types of research that could take the criticism into account while still furthering the cause of EBE: 1) RCTs which carefully monitor context and implementation 2) longitudinal studies that use the available educational data 3 studies with multidisciplinary interventions and combined outcome measures.

The second chapter reports the results of a large-scale field experiment into the effects of a goal-setting intervention on the academic performance, well-being, self-regulated learning, grit, resilience and engagement of first-year students in teacher and business education.

It is followed by a conceptual third chapter that aims to bring findings from different strands of literature together in order to develop an multidisciplinary follow-up intervention that targets academic outcomes as well as mental health.

The fourth and final chapter reports the results of a longitudinal study into the effects of different types of student employment on study progress. It sheds light on the interrelation between finding the right job and performing well in college.

The dissertation ends with a discussion which reflects on both the theoretical contributions of the different chapters, as on their practical use and aftermath.

Chapter 1

Evidence-Based Education: Objections and Future Directions

Over the past two decades, educational policy has favoured evidence-based educational programs and interventions. This led to a rapid increase in the number of large-scale experimental evaluations in education. Although Evidence-Based Education (EBE) and its preference for experimental studies is favoured by recent policies, it also met with resistance from educational researchers. Additionally, it seems that the tenets of EBE are only slowly influencing educational practice. This essay critically reviews the main objections against EBE and its preference for randomised controlled trials (RCT). The objections call for several adjustments to the current EBE and RCT practices, but do not justify abandoning EBE. Three future directions are proposed which could make higher education more evidence-based whilst taking the objections into account: 1) study local factors, mechanisms and implementation fidelity in RCTs, 2) utilize and improve the available longitudinal performance data and 3) use more integrated interventions and outcome measures.

1 Introduction

During high school, a teacher once told my class that research had proven that spreading study time ('spacing') turned out to be more effective than cramming right before the exam (e.g., Dunlosky et al., 2013). Although it was twenty years ago, I remember it quite vividly because it was the only occasion during my education in which a teacher explicitly used educational science to motivate her instruction. I expected higher education, the cradle of science, to be different. But from the perspective of educational sciences, my university courses in philosophy turned out to be rather arcane. Most professors gave uninterrupted 1-2 hour lectures. At best, these 'chalk and talk' sessions were highly interesting and inspiring, but more often they were a very ineffective way to spend the (only) eight programmed hours per week of college. Rarely were class activities (listening) aligned with the activities that were required for the exam (writing an essay) as educational scientists (e.g., Biggs, 1996) proposed they should. One might think that these practices belong to a dusty past. But I studied philosophy from 2006 until 2010 in a country and at a university which consecutively ranked among the international best (e.g., Times Higher Education World University Ranking)[1]. Unfortunately this example is not just anecdotal. Although cognitive psychologists identified effective teaching strategies (e.g., Chandler & Sweller, 1991) and course design methods (e.g., Van Merriënboer et al., 2003), higher education professionals are often unaware of these evidence-based practices (Henderson & Dancy, 2009). Professionals who are aware of them, rarely apply them in practice (Ebert-May et al., 2011; Henderson et al., 2011; Froyd et al., 2013; Stes & Van Petegem, 2011), or customise them by removing critical features (Dancy et al., 2016).

In 2002, Slavin remarked that "The scientific revolution that utterly transformed medicine, agriculture, transportation, technology, and other fields early in the 20th century almost completely

[1] It is telling that global university rankings are nearly always based on research output. Implicitly, the best university seems to stand synonymous for the university with the highest research output and number of citations.

bypassed the field of education" (p. 16). Many interventions are being "tried out" without proper evaluation or scientific basis. Evidence in the field of education is predominantly respected when it supports educational or ideological fashion. Why is education, and particularly higher education, so rarely based on scientific knowledge? Davies (1999) stressed that educational researchers should provide more sound evidence and that educationalists should ask for evidence, know where to find it and weigh it. Slavin (2002), Cook (2002; 2007), and others, advocated the need for more experimental research in particular. These experimental studies should even be complemented with unbiased replication studies, which are still extremely rare (0.13%) within educational research (Makel & Plucker, 2014). The dawn of the 21th century coincided with a rapid increase in funding and support of evidence-based education and educational reform (Slavin, 2020). However, EBE also stirred up a rich variety of critique from the academic field. Scholars criticized the status of RCTs and generalizations based on them (Deaton & Cartwright, 2018; Morrison, 2021). Other's questioned the cost-effectiveness of educational RCTs, or whether EBE overemphasized interventions that can be studied with RCTs (Cowen, 2019). A third strain of critique targeted the broader EBE paradigm and its moral implications for the teaching profession (Biesta, 2007; 2010; Wrigley, 2018). The sheer volume of criticism might be enough for practitioners to jump off the EBE 'bandwagon'. Many indeed opted for a seemingly middle-ground position of 'Evidence *Informed* Education'(EIE)[2]. But quantity is irrelevant when it comes to arguments, in order to weigh them we need to question how valid they are, how problematic they are for EBE and whether they are commensurable with EBE in some way. The current essay contributes to the EBE and educational research literature by critically weighing the critiques against the EBE movement.

[2] Confusingly, the definition of evidence informed education can mean practice that is influenced by robust research evidence or using evidence in addition to practical knowledge and judgment. Both definitions are hard to distinguish from the definition of evidence based education by Davies (1999).

Finally, in line with the call from Newton et al. (2020) for pragmatic evidence-based higher education, three types of research approaches are suggested which could make (higher) education more evidence-based whilst taking the most recent critique and insights into account: 1) experimental research which studies local contexts, mechanisms and implementation fidelity, 2) 'playing to our strengths' by more extensive use of administrative data and more appropriate statistical approaches, 3) using more integrated interventions and combined outcome measures that take both 'side-effects' and a more comprehensive definition of educational goods into account.

2 The Rise of Evidence-Based Education

In a lecture on 'Teaching as a research-based profession' in 1996 (published in 2000), Hargreaves compared the educational profession to the medical profession. Based on his comparison he proposed that it would improve education if, similar to medical science, practitioners could and would make more use of evidence. In an article that meant to define 'evidence-based education' Davies (1999) later stated that:

> "educational activity is often inadequately evaluated by means of carefully designed and executed controlled quasi-experiments, surveys, before-and-after studies, high-observational studies, ethnographic studies which look at outcomes as well as processes, or conversation and discourse analytic studies link micro structures and actions to macro level issues. Moreover, research and evaluation studies that do exist are seldom searched for systematically, retrieved and read, critically appraised for quality, validity and relevance, and organised and graded for power of evidence. This is the task of evidence-based education." (p. 109)

He went on to define the task of EBE movement as the combination of 1) the capacity and discipline of educators to pose answerable questions about education, know where to find evidence, read it and grade the power of evidence and determine its relevance to their educational needs and environments, 2) the power to establish sound evidence where existing evidence is lacking or uncertain. Davies mentions that this requires both insights from social sciences as well as humanistic perspectives (among others).

Slavin (2002) defended a similar view but specifically addressed the need for large-scale experimental evaluations in order to be able to answer questions about effectiveness. Large-scale experimental evaluations of educational interventions are complicated and costly to execute. School leaders often decide not to allocate funds that could be put into their direct responsibility (providing education) to expensive evaluations. Without specific funding, expertise and incentives to organise evaluations, they were altogether a rare phenomenon in the educational field of the twentieth century. Large government programs in the United States (the 'No child left behind' act in 2002 and the 'Every Student Succeeds' act in 2015) and the United Kingdom (the 'What works network' in 2013) managed to break through this evidence-impasse by providing the funding and incentives that are hard or even impossible to organise for single schools and school leaders. Public organizations such as What Works Clearinghouse (WWC) and the Education Endowment Foundation (EEF) additionally offered overviews and reviews to school leaders and professionals who want to know which programs have been evaluated as 'effective'. Other European countries now seem to follow the Anglo-Saxon suit. The Dutch government, for example, recently reserved 8.5 billion euro in extra funding for schools to catch up deficiencies caused by the pandemic. In a similar fashion to US and UK policy, the Dutch government now requires that the money should be spend on interventions from the EEF's registry of proven interventions.

3 Objections to Evidence Based Education

The pleas of Hargreaves (1996), Davies (1999), and Slavin (2002) for more evidence-based education stirred a rich variety of critique from within the educational research community. Although EBE stands for both improving the capacity of educators to make use of evidence and the call for researchers to provide sound evidence where this is lacking, most criticism of EBE was specifically targeted at its preference for randomised controlled trials. Perhaps this is due to the dominance of RCTs in the medical science which EBE emulates and Slavin's (2002) particular emphasis on the necessity for experimental research to answer 'what works' questions.

Cook (2002; 2007) summarized the objections to performing RCTs into 1) philosophical objections about random assignment and causality, 2) practical arguments against mounting experiments (e.g., the focused inequity in school resources that randomizing generates), 3) undesirable trade-offs (external versus internal validity), 4) the objection that schools will not use experimental results and 5) objections that favour other types of study designs. Since Cook presented his 'typology', many new objections and new insights regarding EBE and RCTs in education were published. Some build on arguments within the existing categories, other ontological, economic and normative objections seem to belong to altogether new categories (e.g., Biesta, 2007; Cowen, 2019). Some of these objections have been addressed by proponents of EBE. Slavin (2017; 2020) and Slavin et al. (2021) discussed a selection of the objections against the emphasis on RCTs and EBE: 1) Generalizability: can you really infer from 'it worked here' that it might work somewhere else?, 2) Experiments fail to account for differences in subgroups, 3) Does EB reform privilege experimental studies to the detriment of other types of educational research? Responses from proponents of EBE tackled these questions but left others unanswered. As Newman (2017) and Newton et al. (2020) observed: when different 'camps' are not confronting each other's arguments about the tenets of EBE, this might divide the field of educational science into isolated domains. Critics might have created a 'straw man' to characterise researchers and policy makers aligned with the EBE movement. To some EBE became synonymous with exclusively vouching for quantitative RCTs (e.g., Wrigley [2018] who cites Bennett as stating that we should *only* use educational programs that are effective according to RCTs) and a mere technical view of the teaching profession (e.g., Biesta, 2010). Researchers aligned with the EBE movement, on the other hand, have not always thoroughly dealt with the criticism against EBE's preference for RCTs and the wider potential repercussions of EBE for the teaching profession. The debate runs the risk of losing its intellectual use when the opposing sides divide into separate streams of scholarship. Or even worse, practitioners can select arguments from either side on why they should ignore educational research. It would be more beneficial if we weighed objections against EBE and incorporated feedback in a pragmatic manner. This essay classifies and examines three different types of criticisms against EBE. Objections are categorised as 'epistemic'[3] when they target methodological questions or assumptions and consequences at the level of philosophy of science (when do we know what causes something, for example). Economic objections target the economic feasibility or repercussions of the EBE paradigm. Finally, normative objections are moral by nature and object to the purpose (or lack thereof) of EBE.

3.1 Epistemic objections

As with most methodological issues in general, it is important to note that all methods have advantages and disadvantages. EBE does not a priori select or prescribe one research design over another (Davies, 1999). Hargreaves, Davies, and Slavin all explicitly described how different types of research perform crucial roles in the furthering of EBE. The right design is the one that matches the research question optimally. However, whenever one wants to know if an educational intervention affects a given outcome measure[4], experimental research is usually most suited to confirm a hypothesis about this question with the highest probability. Survey studies and qualitative studies can deepen our understanding of a problem. Design studies are well suited for developing an intervention, qualitative studies can suggest hypotheses about why something might (not) work or for whom. At the end of the cycle, a collection of synthesized experiments (which ideally use

[3] 'Episteme' derives from the Greek word of knowledge. Epistemology is a strand of philosophy of science that studies (what constitutes) knowledge, facts and the foundations for scientific claims.

[4] EBE is often associated with the "what works" question. As will be discussed later in this essay, this should be read as an abbreviation of "which intervention seems to improve outcome measure X". The selected outcome measure should always explicitly be defined and justified.

both quantitative and qualitative data), provides us with the least uncertainty about what might have 'caused' a significant improvement. Experiments therefore play an important role. However, their part is only the final part of a larger cycle or 'ecosystem' of research. Whenever someone wants to know "what works?", the urge might be to skip straight to the end, which can lead to the impression that RCTs and meta-analyses are the *only* type of research that matters to the EBE movement. This notion is incorrect: none of the scholars that publicly introduced the EBE movement (Hargreaves, Davies, Slavin) made this claim[5]. But it has led to heated discussions about the specific epistemic limitations of the RCT study design.

From the epistemic and methodological perspective, several scholars recently discussed how RCTs are misunderstood and overestimated within the EBE context (Deaton & Cartwright, 2018; Joyce & Cartwright, 2020; Morrison, 2021). These issues can be summarized as misunderstandings about a) randomization: it does not guarantee unbiasedness, b) the estimation of the average treatment effect: differences in variance are often not taken into account and averages are less reliable when the distribution is asymmetric, c) sample size balance and precision: with large amounts of potential external causes balance is nearly impossible, d) external validity, and e) causality itself.

Deaton and Cartwright (2018) succinctly described how RCTs can only give us unbiased estimates when randomization does not generate a random imbalance and covariates or confounders are not correlated with the treatment. When the sample is a convenience sample, which is most often the case, its estimate should not be generalized to the broader population or other populations (scaling up) or individuals (drilling down). Joyce and Cartwright (2020) add that external validity in education is highly problematic because educational contexts have great

[5] Davies (1999) urged the need of all types of research, both social sciences and humanistic interpretative science. Hargreaves (1996) urged for a combination of the best available evidence and professional judgement. Slavin (2021) stated that EBE does not prefer one type of research design above the other.

influence on how treatments work. Educational researchers should therefore theorise and study why and how something might work in a specific context. This means studying potential support factors, derailers, and the local structures that afford necessary causal pathways in addition to average treatment effects.

The epistemic arguments point out the limitations of RCTs and urge for improved RCTs and the use of additional types of study designs. However, neither is incompatible with the EBE maxim that urges educators to use the best available evidence. In his recent treatise against the dominance of RCTs, Morrison grudgingly admits that "*pace* Churchill, the RCT is the worst form of design except for all the others" (2021, p. 211). In other words: there is potentially much wrong with RCTs, but even more with other designs as a method of inferring causal relationships. Contributions such as Joyce and Cartwright (2020) raise the standard for the educational sciences and EBE, and urge both scholars, practitioners and policymakers, to be more knowledgeable about the type of research that could ideally answer contextual questions. From this perspective, RCTs should be improved and be complimented by other types of research, but still play a vital role. Calling them the 'gold standard' is too simplistic and leads to misunderstandings, but they are useful for many effectiveness questions as long as they are conducted rigorously and interpreted correctly.

There are more radical epistemic (and ontological) objections against EBE. Biesta (2007; 2010) argued that education is an 'open and semiotic system'. What causes learning is influenced by many variables that cannot be controlled and depends on interpretations by learners. We can therefore not determine 'causes' in a deterministic manner.

Does this objection pose a real threat to EBE? The objection based on the open and semiotic aspects of education applies to social sciences in general. Social scientists take this into account and therefore make probabilistic claims and use probabilistic methods. Contemporary social scientists in general do not claim to discover laws about social behaviour or cognition that are true with absolute certainty. *P*-values are one example of how the -inevitable- degree of uncertainty is taken into account. When performing lab experiments, the question remains whether a mechanism will also work in the field. When evaluating a field experiment, social scientists are aware that many confounding variables could impact results (e.g., Duflo & Banerjee, 2017). The design of an RCT usually counters this as best as possible. The combination of lab and field experiments brings us as close we can get to 'proving' causal relationships, but conclusions thus derived are never final. This uncertainty is completely compatible with EBE's maxim of using 'the best available evidence'. It is important, though, that the degree of uncertainty is never forgotten when evidence is weighed. The 'semiotic' (interpretation dependent) nature of many educational interventions is also typical of social sciences in general. It makes it valuable to not study behaviour alone but also study cognitive and affective factors and processes. Through the past decades, several scholars therefore pleaded for studying mechanisms as well as effects, in line with the epistemological requirements of critical realism. Some indeed developed theories that predict and measure the interactions between interpretations and behaviour (e.g., Oyserman & Destin, 2010; Locke, 2015).

Another set of Biesta's objections targets the epistemology that EBE assumes. In his articles, Biesta proposes using Dewey's epistemology to ground educational science. Instead of using a representational model of knowledge (spectator view) we should use Dewey's transformational model which assumes that reality is constantly changing. The transformational epistemology asserts that it is only possible to determine in hindsight *what worked* but never *what works*, because of the changing nature of reality and because the experimental methods of science change or distort the very reality that they aim to measure.

Summarizing the foundational epistemology and ontology for evidence-based education as the 'spectator view' of logical positivism is too simplistic and ignores the work done by philosophers of science such as Searle (e.g., 1999) and many others. EBE is usually grounded in critical or scientific realism which entails that (ontologically) the world can exist independent of the mind (or science) and that (epistemologically) theories about this world can be approximately true. Dewey's epistemology is notoriously problematic because it erroneously reduces the existence of all theoretical constructs (among which causality) to operational relations (Bulle, 2018). Even if we, for the sake of the argument, followed this fragile epistemology, it would still be compatible with the scientific endeavour to learn from experiences and experiments (e.g., with the design of an RCT). Inferring *what will work* from *what worked* can never be done with absolute certainty, but what has or hasn't worked in the past will often provide the best available evidence for either a theoretical model of causality or 'operational relations'. Surely Biesta does not suggest ignoring evidence about what worked (towards a relevant purpose) in the past when we prepare and choose educational interventions. This would limit even the use of the professional judgment that Biesta propagates, in as far as this is based on previous experiences. Social sciences are nearly always probabilistic. The best available evidence should be combined with professional judgement and deliberation about the desirable ends (Newton et al., 2020). Even these more radical epistemic objections are therefore compatible with evidence-based education.

3.2 Economic Objections

Performing and replicating large-scale experimental evaluations is complicated and expensive (Morrison, 2021). Unencouragingly, Lortie-Forgues and Inglis (2019) recently found that many rigorous large-scale field experiments produce uninformative results. The interesting question that they raised was 'what causes 40% (not the majority) of rigorous large-scale field experiments to provide uninformative results?' They suggested three options: A) the theory on which the programs are based is unreliable, B) these educational programs are ineffective because they have been poorly designed or implemented, C) the studies are underpowered because the outcome measures they use contain more 'noise' than we previously assumed. Explanation A is similar to the underlying cause of the wider 'replication crisis' in psychology (Maxwell et al., 2015), replication studies with large enough sample sizes would eventually 'solve' the problem by filtering out theories that are based on single and often small-scale studies. Explanation B could be solved if monitoring implementation fidelity became more common within the field. Explanation C also requires larger sample sizes and more awareness among researchers about the differences between real world outcomes such as school tests and outcome measures that are specifically catered to the research purpose. In summary, these are arguments that underscore the important added value of large-scale evaluations, replications, and additionally require researchers to carefully monitor implementation fidelity. It does not, however, incentivise school leaders to fund a large-scale evaluation. Not a lot of school leaders feel for investing in something that is likely to show that the efforts of you and your colleagues did not lead to significant (small, if at all) effects. This is a systemic problem that requires government policy which includes reserving sufficient research funding to accompany educational innovation.

Cowen (2019) raised an interesting objection against the predominance of RCTs that evidence based policy has caused. He observes that EBE allows policy makers to target interventions that teachers have to apply instead of policies which they are accountable for themselves. EBE favours teacher level interventions over structural change of the educational system given that this is easier (or even possible) to measure with an RCT. The effects of a program for learning languages is easier to evaluate with a large scale-RCT than the effects of a structural overhaul of the educational system. This 'bias' does have a function or upside. Structural overhauls of the educational system come at great costs (both financial and mental) and peril, this in itself should be an argument to be relatively more conservative when it comes to structural reorganizations. It is likely also more expensive to experimentally study structural changes in education. Similarly to the previous objection, this could partly be countered by (inter)governmental regulation of research funding that accounts for this 'bias'. On the other hand, Cowen (2019) points out that it could be solved if EBE would draw from the full range of available research techniques when it comes to studying potential benefits to structural changes to educational systems. This is compatible with the EBE maxim to use the best available evidence.

Another way in which economic objections about the costs of large-scale evaluations can be taken into account as well as possible, is by properly weighing the effects that are found. Greenberg and Abenavoli (2017) and Kraft (2020) recently offered insightful suggestions on how our interpretation of experimental evidence should be improved. Traditionally, many RCTs in educational research used outcome measures developed specifically to measure the expected effects (most often in the form of a survey), and measured the effects of targeted instead of universal interventions with standardized effect sizes (Cohen's *d* in particular). Specifically designed outcome measures used shortly after the intervention generally lead to larger effect sizes, which inflate expectations of the effects on actual practical outcome measures such as standardized tests and long term effects. Studying targeted interventions means using a more homogeneous sample which by definition leads to smaller variance and thus larger effect sizes. Cohen's *d* does not take relative risks into account and therefore 'overvalues' small-scale trials with low variance. Greenberg and Abenavoli (2017) present a clear example in their paper: a trial with an intervention that resulted in 0.9 % heart attacks (n = 104) in the treatment group, compared to 1.7 % in the placebo group (n = 189) indicated a relative risk reduction of 53%, which was so large they stopped the trial out of fear of mistreating the placebo group. But the standardized mean effect size (Cohen's *d*) was only 0.03. Most educational interventions are universal interventions, their comparative effects have often been undervalued compared to targeted interventions and unrealistic expectations of standardized effect sizes. Kraft (2020) suggests using a different interpretation of effect sizes that takes the design of the study (large-scale, heterogeneous sample, 'real' outcome measures, etc.), costs per pupil and scalability of the intervention into account. This should aid us in making sense of large-scale RCT outcomes and subsequently helps define what we should interpret as successful educational innovations.

A final interesting objection to how RCTs are currently used in EBE was raised by Zhao (2017). He argued that educational researchers too often fail to take 'side-effects' into account in their trials. If we narrowly focus on one learning outcome, we might fail to notice trade-offs. Emulating medical science, as EBE purports to do, should include using a wider range of relevant outcome measures in RCTs to monitor side-effects. Zhao claims that even some of the most contested subjects in educational research might be 'appeased' if we acknowledged the trade-offs of different interventions. Using direct instruction as a didactic teaching strategy leads to higher learning outcomes, but fails to convince critics who value the potential 'costs' to creativity or professional flexibility too much. Mounting more RCTs that show the positive effect of direct instruction on learning outcomes will probably not convince critics who value the other types of outcomes. Experiments that take learning outcomes as well as its impact on creativity and curiosity into account (and report this) will be more constructive to the debate (Zhao, 2017). Finding out what the potential side-effects are requires researchers to improve their study designs (e.g., to exploratively search for potential side-effects qualitatively, track long term effects, also measure wellbeing etc.) and requires the educational domain to define which educational goods are most relevant. This second requirement will return as part of a normative objection in the next paragraph.

3.3 Normative objections

A third category of objections against RCTs and the EBE movement in general is normative by nature. This means that the objections are targeted at the aims of EBE, the paradigm which it stands for, or the moral implications that it has. It is functional to distinguish objections to RCTs as a research design from objections to EBE in general. While epistemic and economic arguments primarily addressed arguments against the predominance of RCTs, normative arguments have mainly stressed the broader EBE paradigm. In a range of articles and books, Biesta (e.g., 2007; 2010) argued that EBE is misguided because education is not effect-driven but value-driven, it is an inherently normative profession. Learning should always be directed at some educational good. Biesta divides educational goods in three categories: qualification, socialisation and subjectification.[6] According to Biesta EBE is misguided because it places too much emphasis on qualification and too little on subjectification, and because EBE will inherently value outcomes that can be measured.

There are two things to consider here. Are the goals of EBE misguided? And are there educational goods that cannot be measured? Regarding the first question, every researcher should be transparent about outcome measures. Every society and school should likewise test transparent learning goals and outcomes with every single examination that is undertaken. Outcome measures such as reading and math ability scores on standardised tests are prevalent because there is an overwhelming democratic consensus about their value. The more idiosyncratic and subjective goals become, being a good citizen, or being a good person even, the less democratic consensus can be found and the less they belong in public education. As soon as a social or personal educational good is agreed upon, researchers can study it as an academic performance measure. In elementary schools and secondary schools in most western countries, the educational goods are partly defined by democratic governments and defined by schools in order to compete for (parents and) kids. In post-tertiary education goals are largely decided by the teaching staff and potentially by representatives of a vocational field. If a vocational school targeted at hotel management considers 'hospitality' an important educational good, they can teach and assess it. If an art-school demands their students to create authentic masterpieces that depict their personal subjectivity they can reward this with grades or other marks. Grades, and study progress can be studied as academic performance. In itself, academic performance can stand for any type of goal. The problem of the educational researcher is therefore similar to the problem of the teacher or curriculum designer. The argument of Biesta (2010) and others (e.g., Wrigley, 2018; Akkerman et al., 2021) is an

[6] These three educational goods are not exhaustive. Brighouse et al. (2016) suggest a wider categorization of educational goods and add distributive and independent values.

addition to the educational debate because it draws attention to the importance of outcome measures both in education and educational research. Their position becomes incompatible with EBE once they argue that there are educational goods about which there is public consensus, that you can teach to students, but cannot evaluate.

4 Conclusion

While reviewing higher education practices, Newton et al. (2020) describe how, even today, ineffective teaching practices and subjective student evaluations persist. The adversary of EBE is not non-experimental educational evidence, but practice based on no evidence at all, or a wrong application or interpretation of evidence. The arguments that were discussed above call for a nuanced view on the usefulness of different types of research designs and disciplines, but no argument validly warrants ignoring *the best available* evidence. There are many problems to consider when interpreting outcomes from RCTs (e.g., they create only a probabilistic equivalence between the groups being contrasted, and then only at pre-test, and many of the ways used to increase internal validity can reduce external validity). Yet, in most instances experimental studies offer the least unreliable estimators or effectiveness. Despite the widespread acknowledgement of their relative superiority, RCTs are still too rare in educational research (Cook, 2007). The recently growing evidence base from experimental studies can improve the influence of educational research on educational practice. Especially if they are conducted according to high standards of rigor.

One risk that should be avoided though, is catering to a need for extremely brief answers to simplified questions: "what works?" Articles, reviews and books that summarize research findings about what works into oversimplified promises fall short of delivering on their promises. As the philosopher Hilary Putnam supposedly put it: "a philosophy that can be put in a nutshell, belongs in one." Dumbing down and summarizing too much stimulates wrong interpretations of evidence. Educational researchers that aspire to contribute to evidence-based education have a responsibility to conduct rigorous research that takes both epistemic, economic and normative objections into account. Educational professionals, in turn, have a responsibility to be curious about the best available evidence.

5 Future Directions

Newton et al. (2020) offered a useful model for 'pragmatic' evidence-based education for practitioners. The final part of this essay will build upon their model by suggesting three directions for researchers interested in furthering the usefulness of evidence-based education. These three directions for future educational research are based on the earlier discussed objections to EBE. They do not exclude other types of research, but could be of specific use for the particular problems that were raised against the use of RCTs in EBE.

5.1 Context-Centred Experiments

RCTs and especially large-scale field experiments fulfil an important 'deciding' role in the ecosystem of educational research. However, in order to realise this potential they need to meet high standards of rigor (Morrison, 2021). Among other 'standard' conditions they should be based on theory, have sufficient power, use baseline measures in addition to randomisation, and use clear protocols. In addition to these regular standards, educational researchers conducting experiments should strive to meet three additional sets of standards that make experiments more useful to educational practice.

The first thing to consider is the context in which the experiment is conducted (Deaton & Cartwright, 2020). This means studying support factors, derailers, and the local structures that afford causal necessary pathways. Qualitative case-studies, or qualitative evaluations of these factors can be of great added value to field experiments. This allows us to not only learn if something worked in a specific context, but why it worked differently in several contexts.

Second, studying not just academic outcomes, but also the mechanisms that explain how interventions work, will contribute to building theories that are relevant to practice. Theories that explain a whole causal step-wise process can be applied more reliably and transparently. Interventions with a clear mechanism allow both researchers and teachers to take a look 'under the hood' whenever any application of this theory is not producing the expected effects. 'Replication with variation', studying both the outcome as well as the mechanisms, is a suitable way to do this (Locke, 2015).

Third, it could be of great added value to make implementation an integral part of the research design (Moir, 2018). Adopting new programs necessitates change. Professionals involved need to be ready for this change. To understand effectiveness, both the intervention and its implementation should be evaluated. Implementation science has been employed in clinical, health, and community settings, but is relatively new within education (Lyon et al., 2018).

All these standards surely do not make it easier, or less expensive, to conduct large-scale educational experiments. They should therefore preferably be used when a causal issue is important but either lacks evidence or when the evidence is contradictory (Cook, 2007). These high demands shall not always be met, just like with every other research design. But as a standard to aspire to, they show how educational experiments can become even more useful.

5.2 Play to the Strengths of the Educational Domain

Many critics have rightly mentioned that EBE is hard or even impossible because the educational domain is special or at least different from domains such as medicine or agriculture (Morrison, 2021). This debate will probably not be settled anytime soon. Some aspects of the educational domain do indeed make it complicated to study effectiveness. Yet, there are elements of the educational domain that offer benefits to educationalists interested in EBE. Schools, colleges and universities keep track of grades, the status, and many other student and course variables. There is an abundance of longitudinal performance data already available to most schools, colleges, and universities. Grading itself is not free from bias and noise, but with the appropriate statistical methods (i.e., multilevel growth modelling), this offers the potential of studying predictors of differences in performance over time. In many instances researchers use self-report questions about grades or performance where this can readily be supplied with administrative data. Although grades are important, they do not represent the only educational goods. Most schools, colleges and universities evaluate their lessons, curriculum and teachers. These types of student evaluations can be targeted at anything, and have an enormous potential research value. Potentially, because they rarely stand up to scholarly standards (Newton et al., 2020). They are seldom used for scientific study, and they are rarely designed with the scientific rigor that the students who are expected to fill them in have to one day adhere to (the psychometric qualities are often not even studied or transparent, or whenever qualitative, not coded up to scientific standards). EBE should not just be known for using or advocating experimental studies, it should be known for a more scientific approach to educational data as well.

5.3 Integrated Interventions and Outcome Measures

Outcome measures are of fundamental importance to EBE. Two critiques against current research practices could bolster the further development of EBE. Zhao (2017) proposed studying potential side-effects. This entails monitoring the potential trade-offs of an intervention, exploratively monitoring unexpected experiences, and tracking long term effects whenever possible. Biesta (2007) argued that learning should always be directed. Instead of asking "what works" and implying that the educational good is self-explanatory, educational researchers should ask which educational goods are at stake. A subtle but important nuance. It means reflecting on and taking responsibility for transparently chosen outcome measures (Akkerman et al., 2021). In practice this means that educational researchers should critically reflect on what type of interventions should be designed to further which types of educational goods. Educational outcomes interact with other life domains and vice versa. Multidisciplinary approaches could offer integrated interventions that target both educational outcomes as well as other life domains of students, and use outcome measures that reflect multiple educational goods at stake.

6 This Dissertation

This chapter is meant as a critical reflection on the type of research that I hope to contribute to as an aspiring educational researcher. As such I hope this can prove to be a form of reflexivity towards the purpose and consequences of the studies in this dissertation.

The second chapter reports of a large-scale field experiment that measured the effects of a reflective type of goal-setting intervention on both study outcomes and well-being (recommendation 3) of first-year students. Besides measuring the treatment effect, we also monitored implementation fidelity (recommendation 1) and tested whether self-regulated learning, resilience, grit or engagement mediated the treatment effects. The findings show a significant positive treatment effect on course credits and dropout but no evidence of effects on well-being or mediating psychological variables.

The third chapter is a review that proposes a method of follow-up on the goal-setting intervention. It combines findings from AI-research, clinical psychology and educational science to propose using an AI-enhanced chatbot that could deliver personalised follow-up targeted at study skills as well as mental health (recommendation 3).

The fourth chapter reports a longitudinal study which combines data of the study progress of teacher education students over a 4 year span at 25 repeated measures with information about the types and amount of work these students performed besides their study. This study found that paid work in education relates to more study progress while unpaid work in education and paid work outside of education do not. By using the available longitudinal data and appropriate statistical methods it plays to the strengths of the educational domain (recommendation 2). Studying the potential trade-off between the educational good of study progress that is required for an important societal need, and meaningful employment, adheres to the third recommendation made in this essay.

Chapter 2

Reflective Goal-Setting Improves Academic Performance in Teacher and Business Education: A Large-Scale Field Experiment

Students often have trouble adjusting to higher education, which affects their performance and well-being. Scholars have suggested applying reflective goal-setting interventions, and reported positive effects of this intervention on academic performance. However, one study found no effects, which highlights the need for understanding the underlying mechanisms that can explain when the intervention works and why. This study assessed these mechanisms through a rigorous effect test, using a randomised controlled trial with repeated measures throughout the first year of college. We measured the effects of a reflective goal-setting on self-regulated learning, resilience, grit, engagement, wellbeing, and academic performance at three points in time among first-year teacher and business education students (n = 1,134). The treatment group earned significantly more course credits and had a 15% lower relative risk of dropping out compared to the control group. Contrary to the findings of previous studies, these effects were independent of gender or ethnicity. Self-regulated learning, resilience, grit, or engagement did not mediate the effects. Differences in implementation fidelity could explain the varying effect-sizes in previous studies.

1 Introduction

More than a quarter of all students leave western higher education without obtaining the degree for which they enrolled (OECD, 2019). The majority of the dropouts happen in the first year (Willcoxson, 2010), and ample evidence exists that this might be due to students having trouble adjusting to higher education (e.g., Respondek et al., 2020). Difficulty in adjusting to a university and its specific features can lead to stress, poor mental well-being (Bayram & Bilgel, 2008; Morosanu et al., 2010), and academic underachievement, manifested in low grades, reduced course credits, and high dropout rates (Kuh et al., 2007; Reis & McCoach, 2000).

Several rigorous experimental studies have reported that targeted interventions can improve the performance of at-risk students (e.g., Sherman et al., 2013; Walton & Cohen, 2011; Walton et al., 2015). However, universal interventions that target a broad student population are rarely tested with controlled experimental designs.

Morisano et al. (2010) trialed a goal-setting intervention that was low-cost, and potentially scalable, on a small sample. They reported that the intervention, in which students reflected on their desired futures, prioritised goals in line with goal-setting theory (Locke & Latham, 2002), and developed strategies in an essay, helped improve both GPA and student retention. Dobronyi et al. (2019) and Schippers et al. (2015; 2020) tested the effects of similar goal-setting interventions on larger samples. The studies by Schippers et al. (2015; 2020) used a quasi-experimental design on multiple European business school student groups (n = 3,144 and 2,928, respectively). In the 2015 study, the intervention enhanced retention rates and course credits by 22%, and the performance of male students and ethnic minorities improved the most (Schippers et al., 2015). The latter study reported that participation was related to improved academic performance, regardless of the chosen goal (academic, social, etc.) (Schippers et al., 2020). Dobronyi et al. (2019)'s large field experiment with first-year students from a Canadian university (n = 1,356) compared the academic performance of a control group, an intervention group, and a group who received the intervention and a brief mindset intervention at the start of the year. Contrary to Morisano et al. (2010) and Schippers et al. (2015; 2020), they found no treatment effect. This might imply that the effects of this goal-setting intervention are not replicable, or that certain moderators not included in the previously mentioned studies account for the equivocal results.

Research indicated the existence of three different types of factors that could shed light on the mechanism behind the intervention. First, Schippers et al. (2015) suggested that gender and ethnicity moderated the effects, with the intervention being more effective for male students and ethnic minorities (demographics). Second, Schippers et al. (2020) found that the number of words the students wrote correlated with the intervention's effect, suggesting that the extent and earnestness of student participation, as well as their understanding of the purpose, might influence results (implementation fidelity). Third, psychological constructs could mediate the effect of goal-setting on performance, given that goal-setting aims to direct thoughts and behaviors (self-regulation, engagement, grit, and resilience) that subsequently lead to performance. Measuring the impact of goal-setting on both performance and psychological constructs simultaneously could make it possible to test whether the psychological constructs mediate the effect of goal-setting on performance. In his article on goal-setting theory, Locke (2015) wrote that further development of the theory called for "replication with variation" (p. 410). Replication with variation entails searching for moderators and mediators to inductively expand the theory's generality across different conditions. Testing the aforementioned types of potential moderators and mediators can expand goal-setting theory in education, and help explain when and why this type of goal-setting interventions are effective.

Additionally, within higher education, goal-setting interventions have been tested almost exclusively in business and economics courses. To generalize the results to different higher education domains and verify whether the intervention is domain-specific or not, samples should also include other types of university students.

Accordingly, we measured four types of moderating and mediating effects to perform a replication with variation. We tested the potential treatment with a rigorous and well-powered design. To situate the results and implications, we divided the literature review into three sections: (1) an overview of goal-setting theory and the intervention's effects on academic performance in higher education, (2) why and how we expect several psychological constructs to mediate the treatment effects on performance and well-being, and (3) the role of implementation fidelity in experimental studies and replications.

2 Literature Review

2.1 Goal-setting theory and interventions

Scholars have extensively studied the goal-setting theory and its behavioral effects in organizational contexts, sports, and healthcare (Locke & Latham, 2002; Epton et al., 2017). Goal-setting intervention studies began with establishing specific and ambitious goals in low-complexity contexts, such as setting targets for optimizing truck loads, e.g., trying to increase trees that can be loaded onto a truck. An increasing number of studies are modifying and applying goal-setting interventions to the first-year higher education environment, which is a highly complex context, given that the tasks, environment, and the high expected self-regulation are new for first-year students. Experimental goal-setting studies within this context have not yet been included in the goal-setting meta-analyses of Mento et al. (1987), Kleingeld et al. (2011), and Epton et al. (2017). Appendix A Table A.1 offers an overview of all experimental studies examining the effect of goal-setting interventions on academic performance in higher education.

The literature offers three different experimentally tested types of goal-setting in the first year of higher education. The first type asks students to set goals for the grades or the number of course credits that students set out to achieve (Clark et al., 2019; Van Lent, 2019; Van Lent & Soeverijn, 2020). For example, van Lent and Soevereijn (2020) performed a field experiment with 1,092 Dutch economics students and instructed a random subset of mentors to encourage students to set grade goals. Half of these mentors were further instructed to motivate students to raise their grade goal. Students in the grade-goal group performed significantly better, but those who were pushed to raise their grades performed significantly worse. Van Lent (2019) also conducted a field experiment with 2,100 Dutch economy students, asking half of them to set grade goals or other goals in a short survey. These students did not perform better than the control group on their exams. Similarly, in their field experiment with 1,967 American microeconomics students, Clark et al. (2019) reported an insignificant increase in the performance of those who set grade goals. The evidence thus far indicates that 'grade goal-setting' produces little to no positive effect on academic performance.

The second type of goal-setting intervention targets the specific tasks one wants to complete. Clark et al. (2019) conducted another field experiment with 2,004 American students enrolled in a microeconomics course. The students randomly allocated to the treatment group were encouraged to set task goals (e.g., number of online practice exams they would complete before their final examination), while those in the control group received no goal-setting encouragement. Students in the treatment group reported significantly higher task completion levels and scored marginally higher on performance. Despite the modestly positive results, a placebo effect risk is possible, as the control group did not receive a control intervention.

The third category asks students to write about their personal life or 'growth' goals, and write about how they will execute their plans (Dobronyi et al., 2019; Latham & Brown, 2006; Morisano et al., 2010; Schippers et al., 2015; Schippers et al., 2020; Travers et al., 2015). In a small-scale trial conducted on academically struggling students from a Canadian university ($n = 85$), Morisano et al. (2010) tested a version combining expressive writing exercises (Pennebaker & Chung, 2011) with mental contrasting (Oettingen et al., 2010), implementation intentions (Gollwitzer, 1999), and goal-setting theory. The treatment group obtained a significantly higher GPA than the control group. Schippers et al. (2015; 2020) and Dobronyi et al. (2019) used a version that involved similar exercises but included negative scenarios (e.g., what will happen if you do not change your habits?).

Although these different versions offer slightly different experiences, they draw on similar mechanisms. The different steps in the Morisano et al. (2010) version are comparable to the different stages of the reflective goal-setting model developed by Travers (as described in e.g., Travers et al., 2015). For ease of reading this type of goal-setting will here be categorized as forms of 'reflective goal-setting' interventions.

Grade, task, and reflective types of goal-setting interventions in higher education share a common ground in goal-setting theory, but they differ in how directed and extensive they are. Reflective goal-setting interventions seem a promising candidate for replication with variation because results thus far indicate both the largest potential effect as well as contradictory results. As Locke (2015) argued, employing the right moderators or mediators can expand goal-setting theory by improving our understanding of when it works and why. The chosen moderators, which may even be population dependent, may have caused varying effects in previous studies. Furthermore, these studies only included small samples of struggling students and large samples of business or economics students. Schippers et al. (2015) reported a moderating effect for gender and ethnicity: males and students from ethnic minorities benefited more. Therefore, we formulated the following hypotheses:

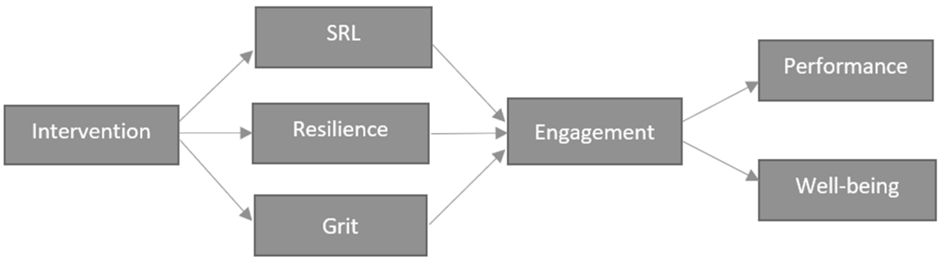
Hypothesis 1. Students in both business and teacher education, who have received a reflective goal-setting intervention at the start of their study, will obtain more course credits and dropout less than their peers in the control condition.

Hypothesis 2. Gender and ethnicity (higher effects for males and ethnic minorities) will moderate the intervention's effect on course credits and dropout rates.

2.2 Potential Mediators: Self-regulated Learning, Resilience, Grit, and Engagement

The recent diversification in the application of goal setting in the educational context has already led to proposed alterations and additions to the goal-setting theory that must be experimentally tested. For instance, Schippers et al. (2020) reported that only one out of five students that participated in the intervention prioritised an academic goal. Nevertheless, the intervention improved their academic performance, regardless of the subject of their goals. This finding differs from goal-setting theory that argues that task specificity is an essential criterion for success. Travers et al. (2015) studied 92 English university students and found that when students wrote about proximal intermediate goals, this induced an immediate increase in effort, a form of self-regulatory behavior. The increase in effort was sustained through persistence and self-efficacy, and many reported that this had led to an upward spiral of subsequent engagement. This mechanism overlaps with several of Schippers' (2017) propositions. Given that a particular intervention may increase students' goal-oriented behaviors, sense of purpose, and explication of their desired futures, Schippers (2017) suggested a focus on improving students' resilience and self-regulatory strategies, as these could lead to higher engagement, academic performance, and well-being (Figure 1).

Figure 1 Mediating mechanisms between goal-setting intervention and outcomes



Note. SRL or self-regulated learning is a multidimensional and modular construct (Pintrich & De Groot, 1990). For this study we used the modules of effort regulation, self-efficacy, intrinsic goal orientation, metacognition, and attention. Adapted from Schippers (2017).

In education, self-regulatory behavior is commonly defined as self-regulated learning (SRL), a multi-dimensional construct that includes "the cognitive, metacognitive, behavioral, motivational, and emotional/affective aspects of learning" (Panadero, 2017, p. 1). In their meta-analysis of SRL's effects on students and professionals, Sitzmann and Ely (2011, p. 422) noted that "one commonality across all the theories is that goal-setting triggers self-regulation." In practice, SRL manifests itself in higher levels of academic initiative, such as active class participation, fewer absences, and less misbehavior in class (Hoyle & Sherrill, 2006; Oyserman et al., 2006). These practical implications are why we expect SRL to improve engagement and academic performance (Pintrich & De Groot, 1990; Sitzmann & Ely, 2011).

Setting goals and anticipating how one should act in trying situations, is expected to improve resilience, defined as the capacity to deal with adversity (Connor & Davidson, 2003). Resilience supports both academic performance and well-being (Johnson et al., 2015; Martin et al., 2015), and could mediate a goal-setting intervention's influence on academic performance and well-being (see Figure 1).

Grit, related to SRL, engagement, and resilience, could also potentially explain why students, who have formulated their goals, persevere and perform well. Duckworth et al. (2007), who coined the term, defined it as "perseverance and passion for long-term goals" (p. 1087); it can also predict academic performance and engagement (Duckworth et al., 2007; Bowman et al., 2015; Hodge et al., 2017).

Engagement, characterized by dedication, vigor, and absorption, is "a persistent and pervasive affective–cognitive state that is not focused on any particular object, event, individual, or behavior" (Schaufeli & Bakker, 2004, p. 295). Dedication is "a sense of significance, enthusiasm, inspiration, pride, and challenge," and to work with vigor means to have "high levels of energy and mental resilience [...], the willingness to invest effort in one's work, and persistence also in the face of difficulties" (p. 295). Absorption refers to a state in which one loses track of the time by being highly concentrated and immersed in an activity. Travers et al. (2015) found that students who engaged in a reflective goal-setting intervention had higher vigor, dedication, and absorption levels. Overall, engagement relates to observed learning activities and course grades, and may be a mediating factor between SRL and academic performance (Bakker et al., 2014). Accordingly, reflective goal-setting could potentially improve SRL, resilience, grit, and engagement. If engagement is affected, this could, in turn, lead to improvements in performance and well-being (Schippers, 2017).

2.3 Well-being

Student well-being has become an issue of concern in academia (Auerbach et al., 2018). Policymakers and scientists argue that many measures that aim at improving academic performance do so at the cost of students' well-being. However, reflective goal-setting interventions aim to improve both academic performance and well-being because they challenge students to set academic, social, and health-related goals (Schippers, 2017; Schippers & Ziegler, 2019). Having the right priorities and strategies should help students engage in activities that allow them to pursue their goals in a healthy way. In a meta-analysis Klug and Maier (2015) synthesized that successful goal pursuit is significantly related to well-being (r = .43). We expect the engagement as a consequence of setting goals and persevered striving (through SRL, resilience, and grit) to lead to increased well-being. In line with Schippers (2017) and based on our expectations of a reflective goal-setting intervention's mechanisms, we propose the following hypotheses (following Figure 1's conceptual model).

Hypothesis 3. Students in the treatment condition will report higher levels of SRL (effort regulation, self-efficacy, intrinsic goal orientation, metacognition, and attention), resilience, grit, engagement, and well-being than their peers in the control condition.

Hypothesis 4. Gender (higher effect for males) and ethnicity (higher effect for ethnic minorities) will moderate the intervention's effect on SRL, resilience, grit, and engagement in both business and teacher education students.

Hypothesis 5. SRL, grit, resilience, and engagement will mediate the intervention's effect on course credits, dropout rates, and well-being.

2.4 Implementation Fidelity

Implementation fidelity, or the degree to which an intervention is delivered as intended, is critical for successfully translating evidence-based interventions into practice. The inconclusive results of previous studies could be a result of the differences in intervention implementation. Durlak and DuPre (2008) found that careful implementation can result in larger effect sizes. Following Dane and Schneider's (1998), and Carroll et al.'s (2007) models, Horowitz et al. (2018) applied their findings to the field of educational psychology and summarized the fidelity concerns into six categories: program differentiation, dosage, adherence, quality of delivery, student responsiveness, and fidelity of receipt.

Program differentiation is the degree to which the tested intervention can be differentiated from the regular program. Using similar interventions with different names might pollute the potential effects—this is a particular risk for certain elements in goal-setting interventions, considering the theory has been around for decades (Locke & Latham, 2002). *Dosage* refers to 'how much' of the intervention was completed. This could be estimated with completion rates, time spent on the intervention, or output variables, such as the number of written words. Students are explicitly encouraged in the intervention to 'keep on writing' and Schippers et al. (2020) found that the number of written words was related to an increase in academic performance, even when controlling for the number of stages students completed and the quality of their goal achievement plans. *Adherence* refers to whether the treatment's parts were followed in the correct sequence. *Quality of delivery* is successful when participants experience the main points as easy to process, true, and emerging naturally. *Student responsiveness* involves students' responses to the adherence and quality of delivery. *Fidelity of receipt* refers to the degree to which students internalize the main points that the intervention aims to communicate. These dimensions require attention, as they provide conditional information expected to influence the results of an experimental study (Durlak, 2015; Durlak & DuPre, 2008).

3 Methods and Materials

3.1 Research Design

We conducted a student-level field experiment at the beginning of the 2018-2019 academic year to test hypotheses 1-5. The intervention was a Dutch translation of the goal-setting intervention from Morisano et al. (2010), and was tested on Dutch students using the think-aloud method. Minor changes were made to increase understandability. This version was translated back to English and then corrected by one of the authors of the original version. The students were randomly assigned to a treatment or control group, and were told not to communicate with other students about the assignments. External surveillants treated the interventions similar to an examination and monitored whether the students did not communicate during the assignments. The participants in the control group received control assignments that looked nearly identical to those of the intervention group, but contained questions about the past instead of the future. The two parts of the intervention or control assignments were sent to the students by e-mail, who completed them individually in computer rooms at the university. Part 1 was made in the first week of college; part 2 was scheduled 3–7 days later. Students had three hours to complete part one and three hours to complete part two; the median time spent was 36 minutes on the first and 51 minutes on the second. To measure the effect of receiving the intervention on SRL, grit, resilience, and engagement we used a baseline survey and two repeated measures after the intervention (T0, T1, T2). We conducted T0 survey at the start of the year and 1–3 days before the intervention, T1 survey two weeks before the end of the first semester, and T2 survey two weeks before the end of the second. We measured the effects of receiving the intervention on academic performance in accumulated course credits and study status (dropping out of the course of study or not) at T1 (+ 2 weeks) and T2 (+ 2 weeks) with the help of administrative data.

3.2 Participants

The sample consisted of first-year students enrolled in 13 courses of study[7] from two faculties within a large Dutch university of applied sciences, located in an urban environment. As part of our selection procedure, we compared the existing program to all parts of the reflective goal-setting

[7] The Dutch higher education system differs from the Anglo-American system in that students have to enroll for a specific 'course of study' (comparable to choosing a major) that consists of a standard curriculum with few or no electives in the first year. Dropping out in this context means abandoning a complete course of study with all of the courses that it contains. Under the current Dutch law, students are not allowed to re-enroll for a course of study at the same university if they fail to successfully obtain a threshold amount of 42 course credits in the first year and all the required course credits of the first year (60) within two years.

intervention to determine program differentiation. None of the courses had used any parts of the intervention.

Seven percent of the student population in applied sciences universities in The Netherlands followed a preparatory scientific track (students from this track constitute the majority of students in the studies by Schippers et al., 2015; 2020), 43% followed a general academic track in high school, 31% had a vocational education background, and 19% used an admission test or an eligible international degree (The Netherlands Association of Universities of Applied Sciences, 2020). We controlled for this sample characteristic in our analysis, because it differs from the samples of previous studies (Dobronyi et al., 2019; Schippers et al. 2015; 2020) and because previous education here is strongly related to central exam scores (similar to SAT scores) (Van der Zande et al., 2018).

The sample was taken from teacher education and business studies faculties. The engineering and medical faculties were also invited, but they declined to participate. All of the courses of study within the two participating faculties were invited. Within the business faculty, 2 out of the 5 courses participated with all of their 302 first-year students. In the teacher education faculty, 11 out of 13 courses participated with a total of 832 first-year students. Table 1 shows an overview of the participant characteristics.

Table 1. Sample characteristics of the freshmen per faculty and condition

	Business		Education		Treatment		Control	
	N	%	*N*	%	*N*	%	*N*	%
Participants	302		832		571		563	
Male	208	69	333	40	268	47	276	49
Ethnic min.	73	24	275	33	177	31	175	31
Voc. backgr.	85	28	225	27	154	27	158	28

The internal review board of the researchers' affiliated university approved the experiment before its execution. All the participants signed informed consent forms beforehand. The procedure in the data management plan ensured the use of pseudonyms before datasets were merged, and anonymous and save storage afterward. Directly after the experiment, all the participants were debriefed and received a book about classroom management (teacher education) or a business journal (business education). The debriefing included information about the design of the experiment and the two assignments. Students who had received the control condition were offered the reflective goal-setting intervention after the experiment.

In total, 1,073 (95%) students started the assignments and 942 (81%) students finished both parts of the treatment or control assignment. The participation rates of the treatment and control group did not differ significantly. According to the teaching staff of the participating courses, 81% completion rate for two assignments is similar to normal participation in course assignments during the first weeks of college. We took the participation rate into account in our assessment of the treatment fidelity. Out of the total of 1,134 students, 1,060 completed every item of the T0 survey and 504 finished the T1 survey online. To secure enough response for the third survey, we distributed the T2 survey in paper format during the classes (653 responses). To assess whether missing responses had potentially led to a non-response bias, we performed several non-response analyses. Specifically, we used a multilevel logistic regression analysis to test whether participation in one of the surveys significantly correlated with being part of the treatment group or relevant control variables (gender, ethnicity, and previous education). Assignment to treatment group, gender, or previous education did not significantly correlate with responding to one of the three surveys. Ethnic minority did significantly correlate with non-response, although the difference was relatively small. The strength of the correlation between being an ethnic minority and finishing the survey was $r = (1{,}133)\ .19$, $p < .001$ for survey T0, $r = (1{,}133)\ .10$, $p < .05$ for survey T1, and $r = (1{,}133)\ .08$, $p < .05$ for survey T2.

After screening, we removed 104 cases in the T0 survey, 21 cases in T1 and 23 cases in T2 (those who had the same answer to all questions or did not clearly write their identification number in the analog T2 survey [8 cases]). The final dataset contained 1,134 cases with demographic data, course credits, and dropout status, of whom 956 had T0 survey scores, 483 had T1 scores, and 630 had T2 scores. As we used repeated measures to measure the effects on the selected psychological variables, we could apply full information estimation in MLwiN (Rasbash et al., 2020), leading to a sample of at least 1,045 in the repeated measures estimation.

We calculated power with the G*Power 3 program (Faul et al., 2007). Power analyses for testing hypotheses 1-5 can be found in the Appendix (Figure B.1-B.3). We corrected the sample size for multilevel structure (13 clusters with an average n of 87) according to Hox et al. (2018, p. 223) with the following formula:

Effective $n = n\ /(1 + [\text{mean cluster size} - 1]\text{rho})$

In all instances the sample was still large enough to find at least a small effect ($f = 0.15$) with a power level of at least 0.90.

At the end of the year, we selected 20 students randomly from the treatment group to partake in qualitative focus groups for evaluation purposes; 14 attended. We asked them to evaluate the two parts and describe if they had learned anything and had applied what they had learned beyond the intervention. All courses of study, except pre-service economics teachers, were represented in this group. Eight of the participating students were female, four were ethnic minorities, and seven had a vocational education background.

3.4 Data analysis

Testing randomization. We conducted independent sample t-tests and χ^2 tests to verify randomization success. This involved ensuring no significant difference between the control and treatment groups before the intervention (T0) in dependent variables (SRL, grit, resilience, engagement, and well-being), demographics, and high school GPA (previous performance is a strong predictor of future performance). As Levene's test indicated unequal variances for metacognition (F = [949] 4.37, p = .04) and resilience (F = [949] 5.86, p = .02), we adjusted the degrees of freedom accordingly (Table 2). T0 baseline survey scores showed no significant variable differences between the treatment and control groups (Table 2), indicating successful randomization.

Table 2. Baseline balance checks with administrative and survey data

	Control Sample mean (*SD*)	Difference with treatment group *(SE)*	χ^2 or t-value (*df*)	*p*-value	*N*
Male*	.49 (.50)	.02 (.02)	.582 (1)	.45	1,134
Ethnic min. background*	.30 (.46)	.01 (.02)	.010 (1)	.92	1,134
Vocational background*	.28 (.45)	-.01 (.05)	.01 (1)	.94	1,134
GPA High School[8]	6.50 (.44)	-.48 (.24)	-1.56 (70)	.12	701
T0 Effort regulation	3.73 (.52)	.05 (.03)	1,474 (958)	.14	960
T0 Self-efficacy	3.92 (.56)	.01 (.03)	-.14 (96)	.89	958
T0 Intrinsic g. orient.	4.21 (.50)	.05 (.02)	1.43 (95)	.15	956
T0 Metacognition	3.42 (.62)	.03 (.03)	.67 (947.23)	.50	952
T0 Attention	3.46 (.67)	.05 (.03)	1,057 (947)	.29	949
T0 Resilience	3.93 (.48)	.00 (.03)	.010 (948.93)	.99	956
T0 Grit	3.65 (.52)	.05 (.03)	1,370 (958)	.17	960
T0 Engagement	3.32 (.66)	.01 (.03)	.34 (954)	.73	956
T0 Well-being	4.55 (.73)	-.04 (.03)	-.75 (954)	.46	956

*= *tested by means of* χ^2 *since variable is dichotomous. df* = degrees of freedom

[8] GPA in Dutch high schools is measured on a 10-point scale, 6 is the threshold for passing. Students with a Dutch tertiary vocational education degree are admissible to a university of applied sciences without needing a GPA score.

Estimating Treatment Effects on Course Credits. As the sample consists of natural groups (13 courses and 2 faculties), we conducted multilevel regression analyses when the intra-class correlations of the study program or faculty appeared to be significantly larger than zero. We tested the models' fit improvements by means of the difference in deviance (-2*loglikelihood) between nested models. This difference has a chi-square distribution with the difference in the number of parameters estimated in both nested models as degrees of freedom. Effect sizes were calculated as the proportions of explained variance, both for total variance and each of the variances in the random part of the model. Using an RCT as the study design, the condition's (intervention) effects on T1 or T2 reflects the effects of receiving the reflective goal-setting intervention. For the dependent variable 'course credits' (Hypothesis 1), we used multilevel regression analyses with two variance levels at T1 and only one level at T2. We present the random parts in Appendix Table B.5.0 and B.6.0. For a histogram of the credits in the treatment and control group at T1 and T2 see Appendix Figure B.4-B.5. We analysed the effects of the intervention with and without controlling for gender, ethnicity and previous education. We tested whether faculty (hypothesis), gender, or previous education (hypothesis 2) moderated the intervention's effect (Appendix B.5.2 and B.6.2).

Estimating Treatment Effects on Dropout Rates. As dropping out of a study program is a binary variable (1 = dropout, 0 = retained), we used logistic regression analyses for this dependent variable and verified whether a multilevel logistic regression was needed. We obtained the starting values for this analysis using first-order marginal quasi-likelihood and the final model fit with second-order predictive quasi-likelihood (Rasbash et al., 2020). Adding the course level to a logistic regression model did not significantly improve the model fit, $\chi^2 = (1)$ .18, p = n.s. It can be inferred that the faculty level is not needed either, because courses are nested in the faculties. Therefore, we conducted a binary logistic regression in SPSS to measure the treatment's effect on dropout rates, with and without controlling for gender, ethnicity, and previous education (hypothesis 2). We used Nagelkerke's r-square to estimate the proportion of explained variance per model, and the difference in Nagelkerke's r-square for the fit improvement between nested models. Also, we calculated the absolute risk reduction, relative risk reduction and "number needed to treat" as an indication of the independent variables' effects (Schechtman, 2002).

Estimating Treatment Effects on Psychological Variables. The intervention's effect on the psychological variables (hypothesis 3) was estimated with multilevel repeated measures regression analyses in MLwiN (Rasbash et al., 2020). In the analyses, we estimated the treatment's effect on the trend in the repeated measures as the interaction between trend and intervention. We included random intercepts for faculty (n = 2) and course (n = 13) whenever this led to a significant model fit improvement (meaning the intraclass correlation or 'rho' is significantly larger than zero).

We also estimated the hypothesized moderation effects of gender, previous education, and ethnicity (hypothesis 4) through these repeated measures regression analyses. The independent variables in the models included intervention, gender, previous education, ethnicity, trend, and the interaction between trend and intervention as fixed effects. When no direct effect was found, we could also exclude a mediated effect (Hypothesis 5). After fitting the repeated measures models, we performed the same ordinary multilevel regression analyses we used for Hypothesis 1 for every psychological construct separately to verify if this process resulted in different outcomes.

Monitoring fidelity. We recorded and transcribed the two focus group conversations, and followed a particular protocol to ensure that we evaluated all parts of the intervention, student experiences, and the degree to which they had internalized the main points. Specifically, we used axial coding to form categories from the answers, and asked the students, through an email member check, whether they agreed with the derived summary and answer categories.

Measures. The selected university used the European Credit Transfer and Accumulation System (ECTS). Within a year, students are expected, when successful, to obtain 60 ECTS course credits that stand for 1,680 study hours (1 credit equals 28 study hours). In their first year, students need to obtain a minimum of 42 out of 60 ECTS to continue studying. Thus, we measured academic performance by tracking the participants' obtained ECTS credits and dropout rates supplied by the university administration.

The following standardized scales measured SRL (effort regulation, self-efficacy, intrinsic goal orientation, metacognition, and attention), resilience, grit, engagement, and general psychological well-being (PGWB). The modular subscales for effort regulation, self-efficacy, intrinsic goal orientation, metacognition, and attention stem from the Motivated Strategies for Learning Questionnaire (MSLQ) (Duncan & McKeachie, 2005; Pintrich et al., 1993). Both subscale selection and the Dutch translation were based on a study that had tested the instruments on Dutch professional higher education students (De Bruijn-Smolders, 2017). We measured resilience with a translated 10-item Connor-Davidson Resilience Scale (Campbell-Sills & Stein, 2007), grit with a translated 10-item GRIT-S scale (Duckworth

& Quinn, 2009), and well-being with a translated six-item PGWB scale (Grossi et al., 2006). Schaufeli et al.'s (2006) nine-item UWES scale served to measure student engagement.

Most subjective and psychological well-being scales include items closely related to having a goal or purpose (Klug & Maier, 2015). This could cloud conceptual clarity and make the correlation between goal pursuit and subjective well-being spurious. The short PGWB scale covers six health-related quality of life domains and none of the items overlap with setting or having a goal: anxiety, depressed mood, positive well-being, self-control, general health, and vitality. Therefore, using this scale allows for a more valid testing of goal setting's effect on well-being.

Six months before the experiment, we pre-tested all the scales on a small sample of students from a different cohort with the think-aloud method (Ryan et al., 2012) and made minor language adjustments to replace complicated words and ambiguous formulations.

We measured dosage fidelity by tracking the completion rates and the number of words that students wrote (Table 1). Three items at the end of the intervention and control group tested student responsiveness to the intervention on a five-point Likert scale, ranging from disagree to agree: serious participation, if they learned something, and if the intervention shaped their thoughts about their future. We also qualitatively assessed both student responsiveness and receipt fidelity at the end of the year with two focus groups (n = 14, intervention only).

Psychometrics. We performed Confirmatory Factor Analyses (CFA) with the Mplus program (Muthén & Muthén, 1998–2006) on the questionnaire items to verify effort regulation, self-efficacy, intrinsic goal orientation, metacognition, attention, resilience, grit, engagement, and well-being scales' validity. We calculated the covariance structures using weighted least squares with means and variances (WLSMV), because the scores are categorical (Likert scales). For each measurement moment, we conducted a separate CFA. After the initial CFA, we used modification indices and factor loadings to identify problematic items. As the variables were summed per used scale in the repeated measures' multilevel regression analyses, the models for each of the three measurement moments must contain the same items. Based on the modification indices, only two items had to be removed. Table 3 shows the results of the CFAs before and after this removal from all repeated measures. Table 4 depicts the reliability of the scales at every repeated measurement and after item removal. Their Cronbach's alpha reliabilities range from moderate (.65) to robust (.86) (Taber, 2018). All scales have alphas above .70, except for effort regulation and intrinsic goal orientation that are slightly under.[9] Tables 5, 6, and 7 present the intercorrelations between the latent traits in the CFAs.

[9] The authors of the final validated MSLQ version reported similar (.69 - .74) alpha coefficient's for these subscales (Duncan & McKeachie, 2005).

Table 3. Reliability of the item sums per construct at T0, T1 and T2 (after removal of 2 items)

Scale	n	Cronbach's α	N -items	Range c-i-t-c	items removed
T0 Self efficacy	958	.75	5	.43 - .62	-
T1 Self efficacy	499	.75	5	.41 - .65	-
T2 Self efficacy	617	.75	5	.41 - .55	-
T0 Intrinsic g.o.	956	.70	5	.35 - .56	-
T1 Intrinsic g.o.	497	.73	5	.37 - .59	-
T2 Intrinsic g.o.	624	.68	5	.40 - .49	-
T0 Metacognition	952	.77	7	.43 - .58	-
T1 Metacognition	497	.75	7	.28 - .57	-
T2 Metacognition	607	.77	7	.41 - .53	-
T0 Attention	947	.78	6	.40 - .65	-
T1 Attention	496	.79	6	.38 - .68	-
T2 Attention	641	.78	6	.45 – .60	-
T0 Effort regulation	953	.65	5	.30 - .53	1
T1 Effort regulation	500	.67	5	.31 - .58	1
T2 Effort regulation	654	.66	5	.35 - .55	1
T0 Resilience	944	.82	10	.36 - .58	-
T1 Resilience	481	.86	10	.41 - .63	-
T2 Resilience	611	.81	10	.30 - .56	-
T0 Grit	937	.78	10	.26 - .56	-
T1 Grit	494	.75	10	.25 - .55	-
T2 Grit	592	.72	10	.24 - .53	-
T0 Engagement	951	.83	8	.32 - .70	1
T1 Engagement	485	.85	8	.46 - .72	1
T2 Engagement	617	.80	8	.37 - .66	1
T0 Well-being	956	.79	6	.49 - .64	-
T1 Well-being	483	.85	6	.56 - .71	-
T2 Well-being	614	.86	6	.52 - .73	-

c-i-t-c= corrected item total correlation

Table 4. Results CFA (WLSMV)

	T0	T1	T2
χ^2	5,388.69	4,359.32	5,496.47
df	1,793	1,793	1,793
p	.000	.000	.000
RMSEA (90% CI)	.05 (.04-.05)	.05 (.05-.06)	.06 (.05-.06)
CFI	.89	.86	.81
TLI	.89	.85	.80

Note. CFA performed with 62 items (after removal of 2 items). For an extended table with the results before removal of 2 items see Table B.1 (in the appendix). Sample sizes: T0 $n = 960$; T1 $n = 505$; T2 $n = 666$.

Table 5. Correlation matrix based on CFA T0 (correlations between constructs without error)

Variables	1	2	3	4	5	6	7	8
1 Grit	-							
2 Self-effic.	.44***							
3 Intrinsic g.	.64***	.50***						
4 Metacogn.	.76***	.43***	.67***					
5 Attention	.66***	.46***	.52***	.59***				
6 Effort reg.	.81***	.45***	.83***	.79***	.55***			
7 Resilience	.68***	.60***	.45***	.47***	.42***	.55***		
8 Engagem.	.48***	.34***	.51***	.54***	.43***	.58***	.39***	
9 PGWB	.36***	.41***	.17***	.32***	.28***	.24***	.51***	.29***

Significance: $*p < .05$; $**p < .01$; $***p < .001$

Table 6. Correlation Matrix Based on CFA T1 (correlations between constructs without error)

Variables	1	2	3	4	5	6	7	8
1 Grit	-							
2 Self-effic.	.59***							
3 Intrinsic g.	.65***	.52***						
4 Metacogn.	.66***	.45***	.49***					
5 Attention	.58***	.49***	.42***	.48***				
6 Effort reg.	.76***	.40***	.80***	.83***	.48***			
7 Resilience	.69***	.68***	.46***	.43***	.33***	.42***		
8 Engagem.	.50***	.39***	.74***	.55***	.45***	.67***	.34***	
9 Well-being	.36***	.38***	.12***	.32***	.36***	.19***	.48***	.34***

Significance: **p* < .05; ***p* < .01; ****p* < .001

Table 7. Correlation Matrix Based on CFA T2 (correlations between constructs without error)

Variables	1	2	3	4	5	6	7	8
1 Grit	-							
2 Self-effic.	.43***							
3 Intrinsic g.	.56***	.34***						
4 Metacogn.	.67***	.40***	.48***					
5 Attention	.63***	.43***	.44***	.48***				
6 Effort reg.	.71***	.28***	.84***	.74***	.46***			
7 Resilience	.63***	.57***	.20***	.37***	.38***	.20***		
8 Engagem.	.51***	.37***	.82***	.51***	.47***	.63***	.24***	
9 Well-being	.35***	.43***	.04	.22***	.27***	.13**	.51***	.26***

Significance: **p* < .05; ***p* < .01; ****p* < .001

We used several fit indices to evaluate the model fit. As the χ^2 statistic is highly sensitive to sample size and tests exact fit, which is too strict a criterion for the social sciences, we also used the Comparative Fit Index (CFI), Tucker Lewis index (TLI), and Root Mean Square Error of Approximation (RMSEA). Generally, a model is considered fair when CFI and TLI $\geq$ .90, and good when CFI and TLI $\geq$ .95 (Hu & Bentler, 1999). In addition, RMSEA values (upper estimate of the 90% confidence interval) of $\leq$ .05 are considered a close (good) fit, between .05 and .08 a fair fit, between .08 and .10 a mediocre fit, and > .10 a poor fit (Hu & Bentler, 1999). The χ^2 of the three models indicate no exact fit and all the RMSEA values of the models indicate a good or fair fit, but the CFI and TLI range between .80 and .89, which is slightly below the fair fit value. All items load significantly on the factor they are supposed to measure, and we also did not find perfect correlations between factors. Therefore, the overall validity of the instruments seems reasonable, the different constructs show good discriminant validity, and the reliabilities are moderate to robust.

3.5 Implementation Fidelity

We assessed implementation fidelity using Horowitz et al.'s (2018) six categories. Regarding the dosage fidelity, 536 students (94%) finished part 1 of the intervention and 470 (82%) finished both parts. We ensured that every student completed parts 1 and 2 in the right sequence by closing the access to part 1 before sending part 2 to the students (adherence). We were able to cover quality of delivery, because the intervention was delivered online and the conditions were controlled in surveilled computer classrooms. The items that measured responsiveness indicate that 69.9% of the participants in the treatment condition, who completely finished both parts, agreed that they took the assignments seriously. One in five (20.1%) neither agreed nor disagreed, and 9.2% disagreed. The degree to which the students took the assignment seriously correlated significantly with the number of written words (r = [917] .36, $p < .001$). In the focus groups, two students reported they did not take the assignment seriously because "it was part of an experiment" and "because I don't like writing so much." A few students reported the intervention had influenced their behavior, and three of them noted its influence in other domains as well. One student said the intervention had helped him combat both his planning and financial issues right at the start of his studies. Another student noted remembering writing down a social and academic goal: “the intervention made me realize that I should stop my loner behavior and try to fit in socially [...] the academic goal made me ask for help sooner whenever I got stuck.”

Half of the students in the focus group, seven of 14, initially did not remember taking part in the intervention, as other researchers reported (Walton & Cohen, 2011). However, some remembered it later during a conversation: “It was right at the start of the study, it was a chaotic period, and I’ve forgotten nearly everything that happened.” Some of these students later admitted that it brought them more focus at the start of their study. When we discussed potential intervention improvements, all the students in the focus group agreed that a more personalised follow-up would help them internalize and utilize the intervention throughout the year. As one student put it: “One’s teacher or coach should recall the intervention one period later. [...] What about your goals now?” Asked about email reminders, the students reported that they already received too many emails and it would be an extra burden. These results indicate moderate implementation quality and we expect to find a (suboptimal) effect of the intervention.

4 Results

Students received an average of 17.24 course credits in the first semester. Those in the treatment group, on average, earned 1.04 course credits more than their peers in the control group during the first semester, which is a significant difference (Table 8, models 1 and 2). This advantage becomes slightly larger and remains significant when we first control for previous education, ethnicity, and gender (Table 8, models 3 and 4). To test whether the intervention works better for subgroups, as determined by Schippers et al. (2015), we added the interaction effects between the intervention and vocational background, ethnic minority, and male respectively, to a model. However, none of these moderator effects proved a significant improvement (Appendix Table B.2.1 and B.2.2). This result suggests that the intervention did not work differently for males, ethnic minorities, or those with a vocational education background.

Table 8. Multilevel regression analyses of treatment effects on course credits after one semester

Effect	Parameter	Course credits			
		Model 1	Model 2	Model 3	Model 4
		Fixed effects			
Intercept	γ_{00}	17.24 (0.94)	16.73 (0.97)	20.86 (0.96)	20.33 (0.99)
Intervention (= 1)	γ_{01}		1.04* (0.53)		1.09* (0.50)
Vocational background (= 1)	γ_{02}			-3.59*** (0.60)	-3.60*** (0.59)
Ethnic minority backg. (= 1)	γ_{03}			-3.52*** (0.59)	-3.54*** (0.59)
Male (= 1)	γ_{04}			-3.21*** (0.55)	-3.20*** (0.55)
		Random effects			
Course variance	μ_{0j}	10.13 (4.46)	10.12 (4.45)	9.00 (3.98)	9.00 (4.00)
Student variance	e_{0ij}	77.04 (3.27)	76.77 (3.26)	70.03 (2.97)	69.73 (2.96)
Total variance	$\mu_{0j}+ e_{0ij}$	87.17	86.88	78.99	78.72
% expl. var. student level			0.35	9.10	0.42
% expl. var. course level			0.17	11.58	0.04
% expl. var. total			0.33	9.39	0.34
		Goodness of fit			
Deviance		8,102.86	8,098.92	7,995.29	7,990.59
Model of reference			Model 1	Model 1	Model 3
Chi-square fit improvement			$\chi^2_{(1)} = 3.94$	$\chi^2_{(3)} = 107.58$	$\chi^2_{(1)} = 4.70$
P-value			$p < .05$	$p < .001$	$p < .05$

Note. Standard errors are presented in parentheses. All p values in this table are two-tailed. Student $n = 1{,}134$; courses of study $n = 13$; faculty $n = 2$. $^*p < .05$ $^{**}p < .01$ $^{***}p < .001$

At the end of the first year, the students earned an average of 42 course credits. Students assigned to the treatment group earned 2.70 credits more ($p < .05$) than their peers in the control group. After controlling for previous education, ethnicity, and gender (Table 9, models 3 and 4), the difference between the treatment and control groups decreases to 2.53 credits but remains significant ($p < .05$). As with the course credits in T1, there are no significant interaction effects: the intervention seems equally beneficial for all sub-groups and independent of gender, education, or ethnicity (Appendix B.3.1 and B.3.2). After controlling for these three factors, the intervention explained 0.34% of the variation in credits in T1 and 0.35% in T2, which equals .11 standard deviations (based on the *SD* of the control group). The intervention, on average, cost students less than two hours, while 2.53 study credits equal 70.84 study hours. Kraft (2020) proposed taking scalability and costs into account when interpreting effect sizes from experimental studies as small, medium, or large. Given that the intervention can be sent to any number of students and requires little time of the teaching staff or university funding, it can be considered low-cost and scalable. According to Kraft (2020), an effect of .11 standard deviation "should be considered large and impressive when they arise from large-scale field experiments that are pre-registered and examine broad achievement measures" (p. 248).

Table 9. Multilevel regression analyses of treatment effects on course credits after one year

Effect	Parameter	Course credits			
		Model 1	Model 2	Model 3	Model 4
		Fixed Effects			
Intercept	γ_0	42.01 (0.67)	40.65 (0.95)	50.52 (1.09)	49.21 (1.27)
Intervention (= 1)	γ_1		2.70* (1.34)		2.53* (1.28)
Vocational backg. (= 1)	γ_2			-9.96*** (1.50)	-9.95*** (1.49)
Ethnic minority b. (= 1)	γ_3			-7.00*** (1.46)	-7.01*** (1.46)
Male (= 1)	γ_4			-7.56*** (1.30)	-7.50*** (1.30)
		Random Effects			
Student variance	e_{0i}	508.26 (21.35)	506.44 (21.27)	463.86 (19.50)	462.26 (19.41)
% expl. var. student (= total) level			0.36	8.74	0.35
		Goodness of fit			
Deviance		10,284.11	10,280.02	10,180.50	10,176.53
Model of reference			Model 1	Model 1	Model 3
Chi-square fit improvement			$\chi^2_{(1)} = 4.08$	$\chi^2_{(3)} = 103.66$	$\chi^2_{(1)} = 3.97$
P-value			$p < .05$	$p < .001$	$p < .05$

Note. Standard errors are presented in parentheses. Student n = 1,134; course of study n = 13; faculty n = 2. Adding course level did not lead to a significant fit improvement, but did lead to a slightly higher treatment effect of 2.72 without controlling for gender, ethnicity and previous education, and 2.68 course credits with these control variables ($p < .05$ in both cases). *$p < .05$ **$p < .01$ ***$p < .001$

With respect to dropout rates, the results were similar: 39% of all students in the control group dropped out of their study program during the first year, compared to 33% in the treatment group (Appendix, Table B.4). The logistic regression shows that the intervention significantly predicts dropout rates ($p = .036$). The standardized effect size (Nagelkerke's r^2) is .01 (Table B.4, Appendix). However, standardized effect sizes undervalue the impact of universal interventions because they are sensitive to base rates and underappreciate the natural heterogeneity in large samples (Geenberg & Abenavoli, 2017). Greenberg and Abenavoli advice interpreting the practical impact of a universal educational intervention on a variable such as dropout with the concept of 'risk reduction' used in medical trials. The absolute risk reduction of the intervention is 5.98% (95% CI [0.39, 11.57]), meaning that if 100 students received the intervention, about 6 would be prevented from dropping out. The relative risk reduction of the intervention is 15.17% (95% CI [1, 27.31]), which stands for the reduced risk of bad outcomes relative to the control group. The number needed to treat is 16.72 (95% CI [8.60, 256.90]), meaning that on average, 17 students need to receive the intervention for 1 student to benefit from its effect. After controlling for previous education, ethnicity, and gender, the intervention's effect is still significant ($p = .042$) and the three covariates together are highly significant ($p < .000$). Therefore, Hypothesis 1 is accepted, while Hypothesis 2 is rejected.

Hypothesis 3 predicted a treatment effect on SRL, resilience, grit, engagement, and well-being. Contrary to expectations, we found no evidence of direct significant treatment effects on effort regulation, self-efficacy, intrinsic goal orientation, metacognition, attention, grit, resilience, engagement or well-being (Appendix Table B.5.1-B.13.1). Therefore, hypothesis 3 is rejected. Both students in the treatment and control group showed a significant decline in well-being, engagement, and SRL at the end of the two terms of the first year. This decline is typical for the first year and end of term (e.g., Corpus et al., 2020; Wang et al., 2014).

Although a moderator effect without a direct effect is unlikely, it is still possible. We continued to test whether significant treatment effects could be found with gender, domain, and ethnic minority as moderators (hypothesis 4). Teacher education students in the treatment group reported significantly higher intrinsic goal orientation (B.7.3), grit (B.11.3), and engagement (B.12.3) than their peers in the control group. Students from an ethnic minority in the treatment group showed a significantly lower drop in well-being than those in the control group (B.13.4). The effect sizes of these moderator effects, however, were negligible (below 0.00). Therefore we found no sufficient evidence to support Hypothesis 4.

Hypothesis 5 supposed that the selected SRL modules, grit, resilience, and engagement would mediate the treatment effect on performance and well-being. However, we did not find a direct effect of the intervention on well-being, hence no mediation could occur, rejecting Hypothesis 5.

5 Discussion

As universities are looking for scalable and low-cost "universal" interventions that could aid a broad population, a reflective goal-setting intervention could provide a solution. However, the evidence about its effectiveness is divided, mechanisms that could explain why and when it works are still underexplored, and the domains in which it is tested are relatively limited. Offering the reflective goal-setting intervention in this study yielded a significant positive effect on course credits and dropout. Although the standardized effect size is small (0.11 standard deviation), it can be considered large because of the low costs per pupil, the scalability of the intervention, and the study design to measure the effects (Kraft, 2020). The intervention, on average, cost students less than two hours, while its gains equaled 70.84 study hours and an absolute dropout risk reduction of 5.98%. In contrast to earlier results (Schippers et al., 2015), the effect was independent of domain, gender, ethnicity, or educational background. Additionally, contrary to expectations, the treatment group did not differ significantly in SRL, resilience, grit, and engagement; these constructs do not appear to be mediators between the intervention and academic outcomes.

Our findings expand the literature on reflective goal-setting and life-crafting's effects on academic performance in several ways. First, we found a potential explanation for the conflicting findings on their effectiveness. Previous studies did not monitor implementation fidelity or did only partially. To our knowledge, this was the first goal-setting intervention study to assess implementation fidelity as part of the design. Owing to its moderate fidelity, we expected the effect of the intervention to have been suppressed. The degree to which the intervention has been successfully implemented could potentially explain the differences in effect sizes. For instance, in terms of student responsiveness, 70.1% reported taking the intervention seriously. Among the reasons for not engaging seriously were a lack of communication and being part of an experiment. These issues are particular to the design of large-scale experiments and could also explain smaller effect sizes. A second example is the intervention's dosage fidelity. Prior research showed the number of written words to be a significant predictor of academic performance (Schippers et al., 2015; 2020). Students in our study wrote an average of 1,134 (*SD* = 671) words, or nearly three times less than the average of around 3,000 words in Morisano et al. (2010) and Schippers et al. (2020).[10] Writing more can be an indicator of more extensive reflections and more specific goal achievement plans. Thus, part of the intervention's effect could potentially be attributed to dosage fidelity. Future studies can build on this approach to ensure that implementation fidelity is closely monitored and taken into account in a meta-analysis. Practitioners could monitor this variable as a potential condition for success.

Second, we found no proof that the intervention improved the self-regulated learning (SRL) modules, grit, resilience, engagement, or general psychological well-being. There was no evidence that these constructs mediate the treatment effect, contrary to Schippers' (2017) expectations, nor did the intervention lead to significant benefits to well-being, as suggested by Schippers and Ziegler (2019). Strikingly, we found no intervention effects on SRL, or effort regulation, given all the previous findings on this effect in other contexts (Locke & Latham, 2002). This result might suggests that the first year of higher education is substantially different from the contexts in which goal-setting interventions have been tested so far, that reflective goal-setting has a different effect compared to regular goal-setting, or that we simply did not find effects that might have been there nonetheless.

[10] Dobronyi et al. (2019) did not report the number of words.

Third, we expanded the intervention to a new domain. Specifically, reflective goal-setting interventions have mainly been studied with students studying business or economics, and we showed that their effects can also be reproduced in the context of teacher education.

Finally, we found significant positive treatment effects on course credits both after a semester and at the year-end. The effect on dropout was only significant after a year and not after one semester. This result most likely means that the treatment improved course credits, which then allowed the students to continue their enrollment. First-year students who do not obtain a certain threshold of course credits in The Netherlands are forced to drop out by regulation. As the treatment effect on obtained course credits grew proportionately, the intervention had a durable benefit that improved over time. This finding is in line with Walton (2014) as well as Schippers and Ziegler (2019), who argued that a well-timed intervention at the start of one's studies can create a positive recursive spiral or stop a negative spiral. It might well be that the intervention aided students to organise and prioritise their studies during a crucial period. Indeed, students in the focus group had mentioned that participating in the intervention had helped them organise their studies, finances, and social lives.

5.1 Limitations and Future Directions

On account of the rigorous double-blind controlled experimental design, the students and teachers received limited information about the intervention and none about its expected benefits. This situation might have lowered participation rates: 81% of all enrolled students finished both parts of the intervention or control assignment. Including students who finished both parts would probably lead to a larger effect size and more precise estimation of the intervention's effect, but measuring the effects of *offering* the intervention, instead of *participating* in it, offers a more realistic estimation of effectiveness in a field setting. In the focus-group interviews, students mentioned that the limited information and experimental status had made them skeptical. They reported that integrating the intervention in the regular curriculum and having a mentor follow-up during the regular coaching sessions would increase the positive effect. Some students remarked about too little follow-up, except for the emails that they perceived bothersome. Future studies could look into innovative and personalised ways of organizing follow-ups, perhaps using a chatbot-coach as a personal reminder for their goals (Dekker et al., 2020), for such interventions to yield a larger effect.

In line with the principles of replication with variation (Locke, 2015), this study examined grit, engagement, resilience, and several modules of SRL as mediators for the intervention's effect to expand the related literature's generality. Given that these constructs did not prove a part of the core mechanisms in this context, future studies could also explore the mediating or moderating effects of other potential constructs, such as procrastination, or other variables that do not require self-reported measures, such as time spent on study and attendance. Further information on mediating constructs can aid the effective directed implementation in the right conditions and contexts.

Although we carefully considered all the aspects of implementation fidelity, we still cannot compare the results to other studies as they did not report on these aspects. This study appears to be the first to examine implementation fidelity. Future research should include transparent measures on its aspects to compare and weigh its impact.

Finally, we found that gender, previous education, and ethnicity were strong predictors of academic performance and dropping out in the first year of college. Studying interventions that could potentially mitigate these negative effects, both in the first year and later years, remains a relevant topic.

6 Conclusion

The teacher and business education students who received a reflective goal-setting intervention at the beginning of their study obtained significantly more course credits and dropped out significantly less than those who received a control assignment. The treatment effects were independent of gender, ethnicity, or previous education. Grit, resilience, engagement, or SRL did not mediate the direct effects. The intervention did not significantly influence the students' general psychological well-being, and its implementation fidelity was moderate, suggesting that the latter may have suppressed the treatment's effects. These findings indicate that reflective goal-setting has a significant effect on academic performance, even when it is implemented at a moderate level. As the intervention took students less than two hours to complete and their gains equaled 70.84 study hours (2.53 course credits) and an absolute risk reduction of 5.98% of dropping out (relative risk reduction = 15.17%), this is good news for educators seeking to improve academic performance. A marginal addition of credits may make a significant difference for low-performing students. Carefully implementing this low-cost and scalable intervention can ensure that more students benefit from the intervention's positive effects.

Chapter 3

Optimising Students' Mental Health and Academic Performance: AI-Enhanced Life Crafting

One in three university students experiences mental health problems during their study. A similar number leaves higher education without obtaining the degree for which they enrolled. Research suggests that both of these problems could be caused by students' lack of control and purpose while they are adjusting to tertiary education. Currently, universities are not designed to cater to all the personal needs and mental health problems of their students. Within the literature aimed at preventing mental health problems among students (e.g., depression), digital forms of therapy are suggested as scalable solutions. Integrative psychological artificial intelligence (AI) in the form of a chatbot shows potential. Simultaneously, within the literature aimed at improving academic performance, an online life-crafting intervention in which students write about their future goals has shown to improve academic performance. Because the life-crafting intervention is delivered through the curriculum and does not bear the stigma that is associated with therapy, it can reach more students. However, life-crafting lacks the means for follow-up or the interactivity that online AI-guided therapy offers. This narrative review integrates the current literature on mental-health chatbot interventions with research about a life-crafting intervention. When a chatbot asks students to prioritise both academic as well as social and health-related goals and provides personalised follow-up coaching, this can prevent interrelated academic and mental health problems.

1 Introduction

One in three students leaves higher education without attaining the higher education degree for which they enrolled (Organisation for Economic Co-operation and Development [OECD], 2010; 2013; 2019). Research suggests that students are having trouble adjusting to tertiary education, leading them to underperform academically (Perry, 1991). For example, students are said to have problems with integrating academically and socially (Tinto, 1998; 1999) and with managing their learning processes (e.g., goal setting, planning, monitoring, and time management; Robbins et al., 2004; Richardson et al., 2012). Not only does the first year of college come with a relatively high risk of not succeeding academically, it also coincides with a higher risk of mental health-related issues and subsequently low levels of well-being (Choi, 2018; Auerbach et al., 2018; Bruffaerts et al., 2018; Hunt & Eisenberg, 2010). Mental health and well-being are related and contribute to the decrease of students' academic performance (in the current study defined as student retention, grade point average and obtained credits; Bruffaerts et al., 2018; Foster, Saunders, & Stang, 1995). College students with mental health problems are twice as likely to drop out (Kessler et al., 1995; Hartley, 2010), and depression and suicidal thoughts relate to a lower GPA (De Luca et al., 2016; Mortier et al., 2015). Therefore, mental health and academic performance are interrelated.

Underlying both mental health and academic performance is a broader conception of 'eudaimonic' well-being as self-realisation and meaning (Ryan & Deci, 2001; Waterman, 1993). Research suggests that undergraduate students often have difficulty with finding meaning (Steger et al., 2008) or a clear sense of purpose or direction in life (Schippers & Ziegler, 2019). However, having self-concordant goals (i.e., goals that align with one's values and passions), relates to higher academic performance (Sheldon & Houser-Marko, 2001), higher subjective well-being (Sheldon, 2002), and lower symptoms of depression (Sheldon & Kasser, 1998).

From this point of view, Schippers and Ziegler (2019) suggested using life-crafting interventions in order to help students reflect on their salient personal goals and improve their academic performance and well-being. Life crafting is a combination of techniques that allows people to (1) find their values and passions using expressive writing; (2) contrast desired habits and domains of life with the current state using mental contrasting; (3) use goal setting to prioritise ambitions and guide effort; and (4) effectuate their plans using implementation intentions. Thus, it helps people to become more specific about their goals and goal achievement plans (GAP). Together the exercises lead to a process of life crafting, defined as:

> A process in which people actively reflect on their present and future life, set goals for important areas of life—social, career, and leisure time—and, if required, make concrete plans and undertake actions to change these areas in a way that is more congruent with their values and wishes. (Schippers & Ziegler, 2019, p. 3).

The potential impact of life-crafting interventions appears to be promising, particularly because it is online and, therefore, scalable. However, it also has three shortcomings. First, the current intervention technique does not ask follow-up questions. When students write brief answers to the life-crafting questions, the online questionnaire is not programmed to encourage students to explicate their thoughts and write more. The second shortcoming concerns the methods for follow-up. Students who participated in the life-crafting exercises suggested that the intervention would improve if the intervention were to include personal guidance after the initial phase. The email reminders used so far were not interactive or personalised. Thirdly, the current programme does not differentiate for individual needs. For students without problems or with minor problems, the life-crafting program might be enough to boost their academic performance and well-being. However, others might require more follow-up and interaction, or might need coaching on mental health problems that interfere with their academic performance. Coaches and psychologists could facilitate personalised follow-up and interaction, but it would be time-consuming and costly. Most higher education institutions do not have the capacity to offer this kind of support. Therefore, there is a need for other scalable solutions, that offer a personalised and interactive program and contribute to early recognition of problems with academic performance or well-being, in order to prevent more severe problems.

A contemporary solution that is gaining momentum in the mental health-care sector is a mental-health chatbot (Abd-alrazaq et al., 2019; Provoost et al., 2017; Vaidyam et al., 2019). A chatbot is a computer program designed to simulate human conversation and is able to create the illusion of intelligent conversation (Abdul-Kader & Woods, 2015; Warwick & Shah, 2014) (for a review, see Fulmer, 2019). In a university setting, chatbots are predominantly used to provide cognitive behavioural therapy (Fitzpatrick et al., 2017; Fulmer et al., 2018; for an overview see Lattie et al., 2019). Other potential positive effects (e.g., on academic performance or well-being) have not yet been studied. Although in general chatbots show promising results (Provoost et al., 2017; Lattie et al., 2019), they are focused on offering therapy, and individuals may not use a health care service due to fears of stigma (Clement et al., 2015; Stewart et al., 2019). To illustrate, less than half of the college students who report suffering from one or more mental disorders seek treatment for those problems (Auerbach et al., 2018; Stewart et al., 2019; Zivin et al., 2009). Furthermore, the majority of students will probably not require cognitive behavioural therapy but would benefit from individualised coaching to overcome the problems they face during the transition to tertiary education. Therefore, in this narrative review, we propose to combine the two lines of research and to deliver a life-crafting intervention through an interactive chatbot. The chatbot can stimulate students to elaborate their answers to the life-crafting intervention, offer interactive and personalised follow-up, and also mental health coaching if needed.

Several studies (e.g., Tinto, 1975; 1998; 1999) indicate that both the transition to tertiary education as well as processes underlying student attrition never occur in isolation; they are the result of a longitudinal process of interrelated individual and environmental factors. We, therefore, advocate a holistic approach that stimulates students to steer their academic work, their social life, and health in the right direction. This proposed life-crafting method offers a positive approach aimed at improvement instead of a more narrower problem-centered approach toward remediation of mental health problems in student populations (Schippers & Ziegler, 2019). Therefore, the intervention can be targeted at all first-year students instead of a group of identified at-risk students, which lowers the threshold to engage with the intervention and it avoids stigma.

Below, we first provide more background information about the mental health and well-being of students and how this relates to academic performance. Subsequently, to provide a rationale for combining a life-crafting intervention with a mental health chatbot, we will first outline what a life-crafting intervention looks like, and then focus on describing in more detail current internet-based mental health care and especially mental health-care chatbots. After that, we describe how we propose to integrate life crafting into an AI-enhanced mental health chatbot. Finally, we present a conceptual model and guidelines for future research to examine the effectiveness of the proposed intervention.

2 Mental Health, Well-being, and Academic Performance

Generally speaking, mental health problems have a high prevalence among students in higher education. One in three college students reports one or more mental health problems (Auerbach et al., 2018; Bruffaerts et al., 2018; Hunt & Eisenberg, 2010). According to a recent study, including students attending 19 colleges across eight countries (n = 13,984), depression disorders are most common, followed by generalised anxiety disorders (Auerbach et al., 2018). At this moment, worldwide, roughly 70% of high school graduates attend college (Auerbach et al., 2018; Bruffaerts et al., 2018). The college years are a peak period for the onset of many common mental disorders, particularly mood, anxiety, and substance use disorders (De Girolamo et al., 2012; Ibrahim et al., 2013).

Part of these problems can be explained by 'study stress' and academic underperformance. Having to study and perform under pressure in college is found to correlate with anxiety and lower well-being (Centre for Education Statistics and Evaluation, 2015; Cant, 2018). Procrastinating and underperforming in college have been found to predict depression, low self-esteem, and anxiety (Van Eerde & Klingsieck, 2018; Saddler & Sacks, 1993). Simultaneously, mental health-related issues influence academic performance (Bruffaerts et al., 2018; Hartley, 2010; Kessler et al., 1995; Kim & Seo, 2015; Steel et al., 2001). There is, as such, an interrelatedness between academic performance and mental health issues. To understand this interrelatedness, and propose solutions that do not improve one at the cost of the other, we should clarify two different underlying conceptions of well-being.

The symptoms of mental health issues are mostly coined in terms of negative affect, i.e., feelings of pain, stress, depletion. The absence of negative affect, in combination with positive affect (feelings of pleasure and joy), determines one's subjective (or 'hedonic') well-being (Kahneman, 1999). In itself, this hedonic perspective on well-being can be a bad indicator of healthy living, given that it can lead to a focus on symptoms only or shortcuts (Ryff & Singer, 2008). A lifestyle aimed solely at hedonic well-being is more likely to be detrimental to well-being in the long run (Huppert et al., 2004; Baumeister et al., 2013). More specifically, pursuing hedonic well-being can conflict with academic and career success, given that studying or working is not always fun and can require hard and arduous work.

Contrary to the hedonic view on well-being, the 'eudaimonic' view on well-being, states that well-being is attained when people live according to their most deeply felt values and are holistically engaged (Waterman, 1993). Both types of well-being are overlapping, yet distinct, and correlate moderately (Compton et al., 1993). Ryan and Deci (2001) argue that obtaining the basic needs (competence, relatedness, and autonomy) improves both hedonic as well as eudaimonic well-being. Lacking one or more of these needs, on the other hand, decreases both types of well-being.

When students attend college, they make the transition from late adolescence to emerging adulthood. Emerging adulthood (ages 18-29 years) is a developmentally crucial period that can be defined by shifts in autonomy (e.g., leaving the home, being expected to organise self-study), relational instability, and shifts in expected competence (Auerbach et al., 2018; Bruffaerts et al., 2018; Burris et al., 2009; Evans et al.,, 2009). This could explain why this period, and the first year of university, in particular, involves such a high rate of dropout and academic underperformance. Interventions that aid students in their shifts in autonomy, relatedness, and competence could thus be of particular value at the start of the study.

3 Life Crafting

As a method of improving both the academic performance of students and their well-being, Schippers and Ziegler proposed using a 'life-crafting' intervention. The online life-crafting intervention consists of several integrated components. These components build on a range of empirically tested mechanisms that aid its participants to reflect on the present and future life, set goals and make plans and undertake actions in a way that is congruent with their values (Schippers & Ziegler, 2019).

The first stage of the intervention guides participants through the process of finding a self-concordant passion or purpose. This phase is not merely aimed at understanding what one likes or enjoys (hedonic well-being), but about finding out what one values as relevant and meaningful. Similar to the Japanese concept of 'Ikigai'; a reason for being (Sone et al., 2008), and eudaimonic well-being, which includes meaning and self-realisation (Ryan & Deci, 2001). This purpose is self-concordant when it is both intrinsically as well as extrinsically worth pursuing (Sheldon & Houser-Marko, 2001; Sheldon 2002). The exercises stimulate participants to choose goals that the person truly holds to be important. This improves the chance that one's (goal pursuing) actions are in accordance with one's values.

Secondly, the planning phase involves ranking goals and mental contrasting (Oettingen, 2000; 2012). This phase helps students to formulate how their desired future differs from their current state. Participants contrast their imagined best possible outcome that is related to the goal with an inner obstacle. This technique is applied to competencies, habits, social life, career, and health. Questions direct the students to describe what competencies and habits they already possess and which desired and needed competencies and habits they lack. Merely thinking about an ideal future can lead to positive affect but decreases the chances that a person takes action in order to realise the desired future (Oettingen & Sevincer, 2018). Contrasting the ideal future with the current state, on the other hand, leads to more effort and positive outcomes (Oettingen, 2012; Oettingen et al., 2010). Knowing which habits one would like to change, improves the chances of actual behavioural change (Holland et al., 2006; Graybiel & Smith, 2014). With the use of a goal attainment plan (GAP), participants can bridge this gap (Schippers & Ziegler, 2019). The same questions are then applied both on their social life, their career/study, and their health.

Thirdly, participants use the goal-setting technique to formulate and prioritise their most important goals. They are encouraged to balance and prioritise social, career, and health-related goals. By doing so, they are stimulated to develop harmonious passion instead of obsessive work passion, which is related to conflicts between different domains of life (Curran et al., 2015). When writing their goals, they are asked to formulate ambitious but specific and attainable goals. This is a technique which is developed by Locke and Latham (2002). Goal setting directs energy to the goal-related actions and improves self-regulated learning and motivation. Prior research has shown that writing about passions and goals is related to increased academic performance (Morisano et al., 2010; Schippers et al., 2015; Schippers et al., 2020). Although it matters whether these are grade goals or task goals (Clark et al., 2016), and reflective goal setting has shown both positive (Morisano et al., 2010; Schippers et al., 2015; Schippers et al., 2020) as well as no results (Dobronyi et al., 2019).

Finally, as part of the life-crafting process, participants design implementation intentions they require to execute their plans. Implementation intentions are 'if-then' plans which aid the person in making goal-related choices in a clutch situation (Gollwitzer, 1993; 1999). These are especially beneficial when they face obstacles or distractions and have a strong effect on goal achievement (Gollwitzer & Sheeran, 2006). Allowing oneself to get distracted from studying (procrastination) is a particular risk for students and a predictor of depression (Saddler & Sacks, 1993), decreased well-being, personal health (Van Eerde & Klingsieck, 2018), and academic achievement (Kim & Seo, 2015; Steel et al., 2001). Imagine that someone wants to spend more time studying, but knows that their phone often distracts them from doing so for a longer period of time. The implementation intention could then be: 'when I am going to study, I will turn off my phone until I've spent at least four hours studying.'

When students have trouble adjusting to the demands and context of tertiary education, they risk finding out about study issues when it's too late. By the time the first exam results come in, it is hard to catch up, given that resits compete with the next exams that are scheduled (Schmidt et al., 2010). Self-efficacy and self-esteem moderately predict success, but the relationship works both ways (Lane et al., 2004; Honicke & Broadbent, 2016). In other words: past performance is also a predictor of self-efficacy and self-esteem. A weak or strong start thus reinforces the self-image and role of efficacy and esteem. When offered at the start of the study, the life-crafting intervention can kickstart self-regulated learning in time (Schippers & Ziegler, 2019).

Preventing these problems right on time, at the start of the study, could prevent a negative spiral. But apart from preventing these negative processes, this approach can also inspire a positive upward spiral. Walton (2014) reviewed an array of short, scalable psychological interventions that have large effects. He deems these wise because when they are offered to the right people at the right time, they can start a recursive process that reinforces itself. Reflective goal setting, according to participants who were followed over a longer period of time with a journal study (Travers et al., 2015) does just that, by bringing about engagement and experiences of flow. We thus propose that a life-crafting intervention right at the start of the study can start a recursive process. Life crafting shows great promise in terms of enhancing academic performance. Combining it with internet-based care could tackle three problems at the same time: the problems associated with adjusting to college life and self-discipline in studying, and mental health issues of students, as well as finding more meaning in life (Schippers & Ziegler, 2019). Below, we discuss findings related to internet-based care.

4 Internet-Based Care

Compared to online treatment, treating mental health issues with traditional face-to-face methods is costly. Internet-based or digital forms of mental care can have the advantage of being scalable and, therefore, cost-effective. Several recent meta-analyses show that internet-based care can be as effective as traditional face-to-face therapy in treating mental health problems (Andersson et al., 2014; Carlbring et al., 2018). Because of its positive effects and its broad potential benefits, the Royal College of Psychiatrists in the United Kingdom advised universities to increase the availability of evidence-based online interventions for students (Royal College of Psychiatrists, 2011). Australia even has an official e-mental health strategy since 2006 (Meurk et al., 2016).

Although meta-analyses seem to show that online and analog therapeutic interventions have similar effects, some forms of online therapy and coaching have better adherence rates than others. We know, for instance, that (mental) health apps are generally used for a short period of time (about two weeks) before being abandoned (Baumel et al., 2019). While it may be that within this period, the beneficial effects are being delivered, it may also be desirable that people make use of such solutions for a longer period of time. Diefenbach and Niess (2015) found that 42% of users stop self-improvement technologies before significant progress is made. Lattie et al. (2019) showed that trials that lasted for eight weeks showed the largest treatment effects in university student populations.

A recent meta-analysis aimed at online interventions in university contexts (Harrer et al., 2019) showed significant general effects of the interventions on stress (g[11] = .20), depression (g = .20) and anxiety reduction (g = .27), role functioning (g = .41) and eating disorders (g = .52). Only four studies out of the 48 included trials measured outcomes on well-being. These four studies all used different scales for well-being (PWB, Core-OM, WEMWBS, and MHC). One of these studies (Kvillemo et al., 2016) used expressive writing exercises as an active control, to measure the effect of a mindfulness intervention, while expressive writing is known to improve well-being (Pennebaker, 2004; Pennebaker et al., 1990). If the latter study is excluded for this reason, a general significant effect of g = .25 on well-being can be found. Harrer et al. (2019) urge future researchers to study which interventions work best for specific types of students. They expect this ‘differentiation’ to further improve the effectiveness of the interventions.

Lattie et al. (2019) did a meta-analysis on internet-based care for university students, which included two trials that involved a chatbot (Fitzpatrick et al., 2017; Fulmer et al., 2018). Both trials

[11] Hedges g was used as a common denominator in the meta-analysis of Harrer et al. (2019) because it adjusts for small sample size bias (Hedges & Olkin, 2014).

reported high retention rates and significant positive effects on anxiety and depression. Other potential positive effects (e.g., performance or well-being) have not yet been studied, and chatbots have so far only been used to deliver CBT in a university context. However, these results seem promising. An intervention integrated with a chatbot is scalable, easily accessible, and adherence rates seem to be better than those for traditional online care.

Although the mental health and academic performance of students at the start of tertiary education are related, the literature and interventions aimed at preventing the interrelated problems are mostly separated. The first one aims at treating or preventing anxiety, depression, and other mental health problems among first-year students with online, digital interventions (Harrer et al., 2019; Lattie et al., 2019). This research and debate take place at the crossroads of clinical psychology, psychiatry, and information technology. Within this line of research, it is argued that going to college coincides with a decisive developmental phase into emerging adulthood (Arnett, 2006). The start of tertiary education coincides with a peak in the occurrence of mental issues (Auerbach et al., 2018; Bruffaerts et al., 2018; Ibrahim et al., 2013). Online or digital treatment is (mainly) a more scalable and cost-efficient method to treat these difficulties (Ebert et al., 2018). The expected mechanism by which online therapy can help or aid is implied to be similar to the mechanisms that guide the effects of the 'analog' type of therapy (with a particular effective and often-used therapy: Cognitive Behavioral Therapy [Davies et al., 2014; Harrer et al., 2019]). A potential unique beneficial quality of online treatment is anonymity, which was found to be related to more self-disclosure (Lucas et al., 2014; Lucas et al., 2017). A downside seems to be higher attrition rates of participants (Baumel et al., 2019). Regrettably, students often do not feel inclined to volunteer to use these available treatments in time; only 20% of those who need it receive minimally adequate treatment (Auerbach et al., 2016), which is likely to result in worse clinical outcomes (Cheung et al., 2017). Research about the more durable and campus wide practical implementation of these treatments lacks in the current literature (Lattie et al., 2019). Chatbots that use AI and offer interactive therapy are at the forefront of the technological development within this field (Fitzpatrick et al., 2017;

Fulmer et al., 2018), with more of the advantages of online therapy, and with a more personalised approach. These are applications that combine the benefits of anonymity with 'rapport' (Lucas et al., 2017).

The second line of research is aimed at improving the academic performance and well-being among students with goal-setting interventions. The data so far shows that goal setting can improve effort and direct effort to the right priorities (Locke & Latham, 2002). Goal setting helps students to allocate their time wisely and improve their academic performance and retention. Within this line of research, life crafting aims beyond just educational goals and strategies (Schippers & Ziegler, 2019). These integrative interventions stimulate students to formulate any type of goal, be they academic-, social- or health-related goals. Formulating goals and strategies to obtain the goals improves academic performance, regardless of whether the students formulated academic goals (Schippers et al., 2020). They argue that a potential spill-over effect is in place: If one formulates goals and does well in pursuing these within one field of life, this translates to positive effects in other domains. A meta-analysis from Klug and Maier (2015) shows that goal pursuit defined as progress instead of attainment, indeed increases (subjective) well-being. Together with Schippers et al.'s (2020) findings, this supports the hypothesis that formulating and strategizing about goals can be beneficial to both academic performance and well-being simultaneously.

We argue that the first line of research lacks the benefits of a more inclusive 'positive' approach that is aimed at all students through the curriculum of their university. This approach can be found in the second line of research. However, the second line of research, in turn, lacks the interactiveness and follow-up that online CBT therapy and chatbot technology provide. By combining these lines by integrating a goal-setting intervention with a chatbot and online CBT, we expect to activate three core mechanisms (right on time, inclusive approach, differentiated follow-up) that allow the integration of mental health chatbot- and life-crafting interventions to be worth more than the cumulation of its parts. In the following, we will specify how these mechanisms work within a chatbot platform and show concrete examples.

5 Mental Health Chatbots

Chatbots, also known as conversational agents, have gradually established themselves as companions to a multitude of modern devices. In the 1960s of the last century, Joseph Weizenbaum at MIT developed ELIZA (Weizenbaum, 1966), an early natural language processing computer program that simulated conversation and that is generally perceived as being the starting point in the development of conversational agents (Henderson, 2007; Jacques et al., 2019). Figure 1 shows a sample of a conversation between a human and ELIZA. Weizenbaum wanted to show how superficial the communication was between a human and a machine, but was surprised to find out that many individuals (including his secretary) would become emotionally attached to the program. They would even forget that they were conversing with a computer, and Weizenbaum's secretary reportedly even asked him to leave the room from time to time in order to have a "real conversation" with the program (Bassett, 2019). The most famous script, DOCTOR, simulated a therapist that used the Rogerian way of conversing. Carl Rogers was a therapist who used non-directional questioning and often repeated back what a client said. The system would parrot phrases back, or ask to elaborate.

Figure 1. Conversation between a human and chatbot [Reprinted from Weizenbaum (1966)].

A typical conversation is the following:

Men are all alike.

IN WHAT WAY?

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE?

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

Since then, conversational systems have come a long way via intelligent assistants like Siri (Apple), Alexa (Amazon), and Cortana (Microsoft), social chatbots aimed at general conversation, and task-focused chatbots (Jacques et al., 2019; Park et al., 2018; Shum et al., 2018). Chatbots are spreading fast among websites and online services in functional areas such as customer service, marketing, entertainment, healthcare, and more. In order to improve the clarity of the discourse on chatbots, Braun and Matthes (2019) propose a framework via which chatbots can be categorized in terms of four characteristics beyond the functional application domain (see Table 1). Despite developments in speech recognition based on (a combination of) keywords, the development of *conversational skills* (e.g., actively keep a conversation going that feels natural) of AI has not improved in a similar pace (e.g., Jacques et al., 2019; Park et al., 2018).

Table 1. Chatbot classification framework (adapted from Braun and Matthes, 2019).

Characteristic	Elements	Description	
I/O	Voice	Speaking	The majority of current chatbots are text based.
	Text	Typing	
Timing	Synchronous	Real-time, direct interaction.	
	Asynchronous	Delayed interaction.	
Flow	Sequential	A specified order in which interaction is structured.	
	Dynamic	Information is processed in an arbitrary order.	
Platform	Messenger	Most current chatbots are connected to or build in a related functionality (like a website) and only a limited number are standalone.	
	Social media		
	Standalone		
Understanding	Notifications	Only sending messages.	
	Keywords	Automated word recognition.	
	Contextual	Include previous messages in the conversation thereby demonstrating understanding of context.	
	Personalized	Take information from external sources and/or previous conversations into account.	
	Autonomous	Independently communicate with humans and even other chatbots.	

Early chatbots depended on deterministic responses that are the result of a rule-based process, which results in chatbots that are perceived as less smart. The more commonly used machine learning techniques allow chatbots to go beyond fixed semantic responses. These techniques have the form of 'supervised learning', using large datasets to train the chatbot which answers are appropriate and which are not; 'unsupervised learning' using Markov-chain based models; and 'hybrid intelligence' which combines both (c.f., Radziwill & Benton, 2017). The result has the form of highly complex decision trees consisting of if-then statements. Though this may sound like a simple principle, the fact that there is no fixed semantic model underlying the communication (i.e., an open conversation can be about anything, using any phrasing) leads to highly complex decision trees or even networks of decision trees. Training an algorithm capable of providing appropriate responses is complex and takes a lot of time, effort, and large quantities of training material and processing power (Lambert, 2018). Mass availability of personalised and autonomous chatbots, therefore, is expected only in 5 to 10 years (Weidauer, 2018).

6 Design of a Mental Health-Oriented Chatbot for Education

The use of chatbots in education is still in its infancy. Though AI applications have been used to support learning for several decades, the overall application appears to be modest, but expectations regarding the future application and added value are high (Winkler & Söllner, 2018). A systematic review of 80 scientific papers on the use of chatbots in education (Winkler & Söllner, 2018) shows the main focus areas are health and well-being, language learning, providing feedback, and the support of metacognitive thinking, motivation, and self-efficacy. They found the usage of chatbot technology in support of learning to be influenced by individual student characteristics like personality traits, trust of and attitude toward technology, educational background, technological skills, and levels of self-efficacy and self-regulation. These findings match findings from the field of information systems research on technology acceptance (e.g., Taherdoost, 2018; Davis, 1989; Venkatesh & Davis, 2000).

The most prominent theories of technology acceptance include the Technology Acceptance Model (TAM; Davis, 1989; Venkatesh & Davis, 2000) and the Unified Theory of Acceptance and Use of Technology (UTAUT), which both are rooted in the Theory of Reasoned action (TRA; Ajzen, 1985) and the Theory of Planned Behaviour (TPB; Fishbein & Ajzen, 1975). Research in this area has revealed a multitude of factors that contribute to technology acceptance, of which key predictors include the perceived ease of use and perceived usefulness of an application (Davis, 1989), playfulness (Moon & Kim, 2001), perceived presentation attractiveness (Van der Heijden, 2004) and peer Influence (Chau & Hu, 2002). In the case of chatbots, perceived helpfulness has been found as an important predictor of user attitudes towards the use of technology (Zarouali et al., 2018). Technological applications in the area of education, personal development, and healthcare all share these characteristics underlying user acceptance.

The appeal of social chatbots in the area of mental health and well-being is large and primarily lies in their ability to make a social connection to users (Bickmore et al., 2005; Fitzpatrick et al., 2017; Shum et al., 2018). These chatbots show more promise than general mental health applications, through their potential to dynamically recognize emotion and to engage the user throughout conversations by showing appropriate responses (Shum et al., 2018). One of their main shortcomings, however, regards the so far inchoate ability to convincingly convey empathy to the user (Morris et al., 2018). In a clinical environment, for example, empathy has been identified as a key contributor towards better clinical outcomes as it lowers anxiety and distress, enhances satisfaction, and is directly related to higher patient enablement (Derksen et al., 2013). These effects are even more pronounced in the context of mental health interventions (Gateshill et al., 2011). Just as humans, nonhuman agents may struggle to express empathy (Morris et al., 2018). Still, research on mental health-oriented applications shows an overall user preference towards a chatbot compared to general non-conversational applications. Moreover, the use of non-conversational applications has been found to be abandoned after about two weeks by the majority of users (Baumel et al., 2019). By comparison, the adherence rate for a chatbot with a similar focus seems to be four times as long, as a chatbot can actively reach out and initiate communication with participants in a conversational way (Bickmore et al., 2005; Fulmer et al., 2018; Kamita et al., 2019). Expectations regarding the ability of chatbots to understand natural language and have meaningful natural conversations have not been met yet. However, as systems improve, the difference between humans and machine responses are expected to diminish (Jacques et al., 2019).

7 Integrating the Life-Crafting Intervention with the AI-Enhanced Mental Health Chatbot

Both life-crafting interventions and online mental health chatbot interventions have shown promising results when it comes to improving academic performance as well as mental health and subjective well-being. Integrating both can help in compensating for the downsides of each intervention. For instance, the life-crafting intervention is relatively static in its current form and could profit from the more interactional style from the chatbot. As mentioned before, a downside of the life-crafting intervention was that it did not respond to answers they gave or ask any follow-up questions whenever answers were brief. Writing more words corresponded with a larger effect of the treatment (Schippers et al., 2020), and stimulating students to write more, might make the intervention more effective. The life-crafting intervention starts in a browser and shows uniform texts, images, and videos that introduce uniform writing exercises (Schippers & Ziegler, 2019). Apart from demanding that students write at least one letter per question, there is no response to the brevity or content of what students write. Also, there is no differentiation in the intervention based on choices or texts from the students. All questions and follow up consisted of identical emails with goal setting diaries, which, according to students, did not feel personal and were soon experienced as spam.

The previously mentioned downsides of AI mental health chatbots are that students might be reluctant to volunteer for these interventions because of the stigma that is associated with mental health problems and because many students have trouble recognizing early symptoms of potentially serious mental health issues. Furthermore, these applications are mainly focused on alleviating mental health problems, and not on improving academic performance or eudaimonic well-being.

For these reasons, applying the chatbot to a more holistic approach (aimed not only at mental health problems but at life in general), in which the life-crafting intervention is integrated with an AI-enhanced mental health chatbot shows great promise. By combining a focus on life crafting, personal interactive coaching, and mental health, this approach is aimed at increasing general student academic performance and well-being, instead of merely focusing on potential problem areas. We suggest that all students receive this intervention at the beginning of their first year in tertiary education. That way, accessibility is large as all potential users will receive the intervention at the beginning of their first year. The opportunity to start using the chatbot at the start of the university studies has an added benefit toward early recognition and remediation of potential problems. The chatbot can play an important role in detecting (the development of) mental health problems as well as academic problems early on in the academic year. This way, we expect that the development of mental health problems can be prevented, or the student can receive additional online coaching on mental health issues by the chatbot early on, or the chatbot can refer the student to other sources of mental health coaching. Furthermore, the chatbot can also pro-actively seek contact with the student on the moments that the students' stress level is expected to be on a high. For example, in the three weeks before a test week, the chatbot may check in with the student, inform whether the student is doing well, what learning goals have priority for the student at the moment, and ask if the student might need some help. We propose that this holistic, positive program aimed at what is most important for students combined with more differentiation could further enhance the user experience and improve its subsequent effects. A chatbot can thus be used not only in a curative way but also to detect problems early on and to prevent mental health issues from arising (Bendig et al., 2019; Schippers & Ziegler, 2019). Furthermore, the life-crafting intervention integrated into the chatbot can enhance academic performance and increase well-being for all participating students.

Within the chatbot platform, it is possible to differentiate between the needs of different students and thus offer a more personalised intervention. This personalisation can be achieved in several ways. With regard to goal setting, self-regulated learning, and academic performance, students might report a wide range of issues. For example, some students might need help with the formulation or the prioritisation of goals. Others might need help with regard to planning, monitoring, and time management, or ask for advice on how to learn in a better manner, for example with respect to learning strategies. With the chatbot, the set of effective self-regulatory processes for academic performance in higher education (De Bruijn-Smolders et al., 2016), based on the framework of self-regulatory processes as proposed by Sitzmann and Ely (2011) will be addressed with complementing evidence-based interventions. For example, with regard to planning, monitoring, and time management, students can be offered guidelines such as to study each day, to study the most difficult part first, and to use a to-do list when studying, and to make three kinds of planning, that is, for the day, the week, and for the long-term (for example until the test week; Gettinger & Seibert, 2002; Hattie 2009; Plant et al., 2005). With respect to mental health, in line with the literature, we expect anxiety and depression to be most prevalent among the students (Auerbach et al., 2018). If students score high on the surveys on anxiety or depression, the chatbot will advise them to visit a student-advisor, and will advise them to follow cognitive behavioural therapy, online via the chatbot, or with an external professional. With respect to both, we want the chatbot to check in with the student at the right time (Walton, 2014); on the moments when we expect that students' stress peaks will be on the rise, for example in the weeks before important exams, the chatbot will check-in with the student and offer customized coaching. Table 2 shows examples of what these conversations between a student and the chatbot could look like.

Table 2. Artificial enhanced life crafting, sample conversations chatbot and student.

Life crafting intervention, four stages (Schippers and Ziegler, 2019)		
Stage 1: Finding Self-concordant passion or purpose (Sheldon and Houser-Marko, 2001)		
Chatbot: What do you like to do? Write about this for approximately 2 min...	**Student:** Windsurfing	**Chatbot:** Hello <<student>> I notice that you have answered this question with one word. To use this intervention optimally, for your academic performance and well-being, it is important that you elaborate more on this. Can you tell me more about why you like windsurfing? Write about this for approximately 2 min...
Stage 2: Mental contrasting (Oettingen, 2000, 2012) and formulating goal attainment plans (Schippers and Ziegler, 2019)		
Chatbot: What do you admire in others...	**Student:** I admire fashion designers. Take for example Marc Jacobs. In a documentary I was inspired by his enormous drive, high energy level and positivity. His work really seemed to align with his passion. I hope I will also find what my passion is, my goal in life, and which job may fit with my passion optimally...	**Chatbot:** Thank you for your detailed answer, <<student>>. With the following questions, I aim to help you with finding your passion, your goal in life, and with making steps to attain your goal.
Stage 3: Formulating and prioritizing students' most important social, career, and health-related goals (Curran et al., 2015)		
Chatbot: Now that you have your ideal future in mind, describe the goals that you can derive from this...	**Student:** Study more often (Academic goal), find a study-related job (Career-goal), Learn to deal with stress (Mental Health goal)...	**Chatbot:** Thank you <<student>>! Good to hear about your goals. I clearly see that you know what you want to achieve, which will be conducive for your academic success and well-being. However, unfortunately you will not be able to achieve all your goals at once. For this reason, can you rank them in order of importance?
Stage 4: Designing implementation intentions ["If-then plans which help the student to set the right priorities in a complex situation (Gollwitzer, 1993, 1999)"]		
Chatbot: Now you have specified your goals, make an if-then plan that will allow you to choose to do something that will help you to achieve your goal.	**Student:** I am not sure, can you give me an example?	**Chatbot:** Sure <<student>>. With respect to studying, an if-then plan might be: If I come home and am tempted to watch a drama series, then I will make sure that I first study for 2 h.
	Student: Ah, thank you <chatbot> When I am going to study, I turn off my phone until I've spent at least 4 h studying.	**Chatbot:** A solid if then-plan <<STUDENT>>, for studying more! Can you also formulate if-then plans for your other goals?
Measurement scales well-being (anxiety, depression, self-efficacy, test anxiety,...)		
Chatbot: This scale will help me get a better sense of your experience with anxiety OK? ...	Student fills out General Anxiety Disorder 7-item scale (GAD7) (Spitzer et al., 2006).	**Possible answers chatbot:** This screening is not meant to be a diagnosis.... **Option 1:** Your responses indicate no or very few signs of anxiety. Keep in mind that minor ups and downs are normal. **Option 2:** However, your responses indicate some signs of anxiety. Would you like me to offer you cognitive behavioral therapy? This has helped other students in the past. **Option 3:** However, your responses indicate symptoms of anxiety. I recommend visiting one of the university student councilors, click here <<HYPERLINK>> to make an appointment. If you need immediate support you can call this number free of charge <<CRISIS_NUMBER>>.
Follow up: Stimulating and monitoring students goal progression via goal diaries[a]		
Chatbot: OK <<STUDENT>>, let's review the goals you worked on this week. Are you ready?	**Student:** Yes	**Chatbot:** What was the most important goal you wanted to accomplish last week?
	Student: Studying more	**Chatbot:** And, did you successfully complete what you were trying to accomplish?
	Student: No	**Chatbot:** What was the main reason for you that you were not able to study more?
	Student: I had not foreseen that certain tasks would take so much time. I did not have enough time to finish my study tasks this week.	**Chatbot:** Sorry to hear that <<STUDENT>>, do I understand it well that you had problems planning in your study tasks this week, and with your time management?
	Student: Yes	**Chatbot:** I want you to know that I can offer some coaching on this point. Are you interested to hear more from me about possible strategies to manage your time better, for studying more?
	Student: Yes	**Chatbot:** Alright <<STUDENT>> Let's start (chatbot offers study interventions for better planning, monitoring, and time management) (e.g., see Gettinger and Seibert, 2002; Plant et al., 2005; Hattie, 2009).
Chatbot pro-actively seeks contact with student. To begin with, the chatbot checks in on a weekly basis. Later on, the chatbot will check in more or less frequently, dependent on student's needs.		
(1) Monitoring goal progression		
Chatbot:<<STUDENT>> I want to check-in with you. Have you been making progress to your goals this week? To start with your health-related goal: did you exercise two times last week?	**Student:** No	**Chatbot:** What has gotten in the way?
	Student: I could not motivate myself to go to the gym, after I had had dinner at home.	**Chatbot:** I see, what can you do to make more progress with respect to your goal: exercise two times per week? Let me rephrase your if-then plan: "If I want to go to the gym the next day, then I will pack my sports bag the night before." Was this if-then plan helpful with respect to your goal?
	Student: No, my if-then plan was not supportive for reaching my goal.	**Chatbot:** Can you think of an alternative if-then plan that will help you to accomplish your goal?
	Student: Yes. If I want to go to the gym, I will take my sports bag with me and go right after attending college.	**Chatbot:** Great! Well done <<STUDENT>>! Let's see if your new if-then plan helps you better to reach your goal.
(2) Emotional check-in		
Sample conversation chatbot and student		
Chatbot: Hello <<STUDENT>>. How are you?	**Student:** I am worried, my grades are not so high, I do not know if I will pass this year.	**Chatbot:** I am sorry to hear that, <<STUDENT>>. In case you want to talk with someone about your situation, please feel welcome to contact one of the student advisers of <<educational program>>. You can click here <<HYPERLINK>> to make an appointment right now.
(3) Check-in right in time (e.g., 3 weeks before an exam period)		
Chatbot: Hello <<STUDENT>>. Just checking in. How are you?	**Student:** I am very stressed. I am so busy preparing for my exams.	**Chatbot:** I am sorry to hear that, <<STUDENT>>. I can coach you on this, also I can offer you some strategies for dealing better with Stress. Sounds good?
Student seeks contact with chatbot (24/7)		
Sample conversation chatbot and student		
	Student: Hello <<chatbot>> I am feeling very lonely at the moment.	**Chatbot:** I am sorry to hear that <<STUDENT>>. Can you tell me more about why you are feeling lonely? ...

[a] *As a follow up of the life-crafting intervention, the student receives six goal diaries to fill out, two-monthly, for monitoring and stimulating goal-progression.*

Moreover, within the life-crafting intervention, differentiation could also be applied. For example, according to Powers et al. (2005), implementation intentions, which is a part of goal setting, can be detrimental to students who score high on perfectionism. Some parts of the life-crafting intervention or even the complete intervention could not be beneficial to this particular subgroup. Short personality scales could be used before the onset of the intervention, to diversify the content of the intervention or even the complete intervention. A chatbot could start with an intake in which the student answers a survey on personality and well-being that allows the chatbot to offer a tailor-made program.

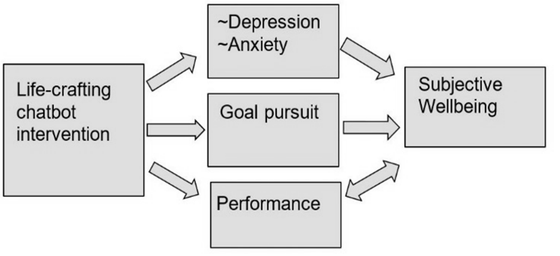
After the intake and a tailor made life-crafting intervention, the chatbot should remain readily available for regular cognitive behavioural therapy. But, as was also described by students who evaluated the life-crafting intervention, there should be a pro-active follow-up on the intervention. The chatbot will use the goals and strategies that the student has decided on to check-in on their progress. A chatbot can stimulate students to regularly reflect on, and remind them of, their goal progress with questions and personalised feedback. Schippers and Ziegler (2019) mention examples of questions that could be used for effective follow-up on the intervention: "Did I invest enough time into my goals? What could I do to improve this? Which smaller sub-goals could help me to achieve my objective? What obstacles do you face? What ways do you see to overcome them?" (p. 11-12). The chatbot can use cues in the answers of the students to offer the right type of strategies, for improved planning or combating procrastination for example.

8 Conceptual Model

Some researchers state that merely having a goal already improves well-being (e.g., Freund & Baltes, 2002; Klinger, 1977). Gollwitzer and Brandstätter (1997) distinguish different phases in goal pursuit: pre-decisional (deciding about preferences between different goals or wishes), pre-actional (the initiation of goal directed actions), actional (successfully performing actions that bring a goal about) and post-actional (evaluating results with the original intentions). Gollwitzer and Brandstätter state that it is to be expected that setting goals triggers pre-decisional and pre-actional goal pursuit. We predict that adding follow-up questioning and coaching via a chatbot can also improve the actional and post-actional part of goal pursuit. In other words, setting goals initiates goal pursuit, but the follow-up through coaching from a chatbot can also improve the later phases of the pursuit of goals. Prior research has shown that goal pursuit, when conceptualized as goal progress instead of goal attainment in turn has an average effect of $\rho = .45$ on subjective well-being (Klug & Maier, 2015).

We expect the low-threshold CBT therapy that the chatbot can offer based on intakes and scales that are included in the first part of the intervention to decrease anxiety and depression (Fitzpatrick et al., 2017; Fulmer et al., 2018). Including a large population of regular students in the treatment group might lead to results that differ from previous studies that only included students who volunteered to participate. Testing this is a necessary next step in the development of the literature. It is expected that goal progress influences SWB through an increase in positive affect, and the prevention of depression and anxiety improves SWB mainly through the negation of negative affect (see Figure 2). It is thus important to know how such a chatbot can be designed.

Figure 2. Conceptual model with expected mechanism of a life-crafting chatbot intervention



9 Design Principles for a Life-Crafting Chatbot

Extant literature and experience have shown that the use of experimental or novel technologies is always associated with risks of low adoption. As Lattie et al. (2019) observe, digital mental health interventions, in particular, tend to fail due to acceptability, usability and feasibility issues. While in the previous parts we discussed potential issues and limitations that oftentimes plague such implementations, we stress the importance of the design philosophy before zooming in on the different design aspects themselves. Overall, human-computer interaction (HCI), in the context of every application, is a complex and dynamic experience that ever-evolves (as software gets updated). Naturally, the goal-setting intervention underlying the present study, as well as the chatbot used as the agent of delivery, also evolve based on the feedback and results received with each intake of students. The design principles, however, guiding the blueprint and evolution of the intervention should be grounded in suitable paradigms of HCI. In our cases, these are the design rationale (what user requirements does the platform intend to address? What are the reasons behind its particular features or the ones it doesn't have? What are the trade-offs?) and usability engineering (*iterative development* based on usability specifications, *participatory design* by involving students in the development of the platform, *impact analysis* and overall cost-effectiveness evaluations) (Carroll, 1997). Following these two paradigms will allow us to address a number of issues related to the successful implementation of the intervention in a structured manner.

Current chatbot interventions in the university context can further improve their user-friendliness by 1) being more tailored to the intended users 2) addressing issues that are most important to the users 3) ensuring user privacy 4) offering a trustworthy experience and 5) offering aid in emergencies (Lattie et al., 2019). If user-friendliness is low, this will likely have a negative effect on the scalability, and durability of the intervention. Following a design rationale perspective, future research could address the first two concerns by identifying the specific needs of the target audience and their key issues that the intervention should be seeking to address. Following a usability engineering approach, we aim at fine-tuning and evolving the intervention in order to address its key shortcomings. This process involves focus groups and regular surveys over a prolonged period. To address the privacy and trust concerns of students, thorough regulation and transparency regarding the data management should be employed and effectively communicated to all participants.

The success of the intervention should be evaluated not only based on user satisfaction metrics but also by the overall user acceptance. The prolonged involvement of students with the chatbot is dependent on its user-friendliness. A chatbot is, by its nature, inherently more interactive and open then most used online interventions. Still, in the Fulmer et al. (2018) trial students did report that the chatbot biggest shortcomings included the chatbot not feeling natural (12/50), misunderstanding replies (11/50), not interactive enough (7/50) and impersonal (6/50). Extensive tests could make the chatbot more user-friendly.

If the chatbot is supposed to play a catalytic role in sustaining user-engagement throughout the intervention, key principles of HCI design need to be combined with key findings from the Technology Acceptance literature. As technology acceptance is concerned not with the unique experience and satisfaction but with the intention of users to change their ways and adopt a new technology in their routines, there needs to be focus on aspects of the design stimulating the key antecedents of acceptance, namely perceived usefulness/helpfulness, ease of use, and playfulness (Moon & Kim, 2001) as well as related antecedents of those such as technology readiness (optimism, innovativeness, discomfort, insecurity) (Parasuraman, 2000) or technostress (Ayyagari et al., 2011). Developing such an integrated chatbot, with the use of modern technology combined with insights from positive psychology interventions such as life crafting, shows great potential in optimizing student well-being and (academic) achievement.

10 Discussion

As many students struggle with academic underperformance and mental health problems during their transition to tertiary education, we sought to outline possible solutions that involve both the use of contemporary AI solutions and combine this with the latest insights from effective positive psychology interventions, specifically a promising life-crafting intervention. The advantages of such a solution are that it is scalable, has a low threshold, would contribute to early detection of academic or mental health problems, and would be interactive and personalised. We proposed an inclusive approach: all students could potentially benefit from the resulting intervention. Combining insights from two lines of research, namely the life-crafting (goal-setting) literature, and the literature on online mental health care, we proposed integrating a life-crafting intervention with a mental health chatbot could offer a solution for all students.

Our focus on scalability as an important criterion has to do with the fact universities are currently not able to cater to be 24/7 responsive to all the personal needs and mental health problems of their students. A chatbot is a scalable solution that is constantly available, because all students can individually take part in this intervention online. Only students with serious academic or mental health problems would be referred to the student advisor for further coaching or to, for example, psychologists. Our focus on a low threshold had to do with the fact that mental health problems bear a stigma that prevents many students from seeking help for these problems. Using a chatbot is anonymous, which is related to more self-disclosure and rapport (Lucas et al., 2014; Lucas et al., 2017).

We proposed an inclusive approach, in which all students within a certain study program receive access to the intervention at the beginning of their first year of tertiary education. The main focus of the intervention is not mental health problems, but life crafting and setting personal goals, which can be beneficial to all students. Having this positive focus will probably also decrease the association with stigma on mental health problems. Only students who need it will also be able to receive mental health coaching via the chatbot. This touches another important criterion that we set for the intervention: differentiation. With a chatbot, it is possible to offer interactive and personalised coaching, based on the students' individual needs. Moreover, the chatbot can also follow-up and interact with the students in later stages on what they have written in their intervention.

Finally, the chatbot can assist in early recognition of academic and mental health problems in two ways. First off, we expect that the life-crafting intervention integrated into the chatbot will make students more aware of their goals and potential obstacles. This will help them to set priorities for themselves, and may also encourage them to seek help for their problems in an early stage. Secondly, the chatbot itself can also recognize signals of academic or mental health problems, and offer in-app coaching (for mild problems) or refer to external help (for more severe problems) in early stages, if necessary. An additional advantage is that mental health chatbots often have higher adherence rates than other internet-based mental health care.

Future research should experimentally test the effects of interventions that combine insights from positive psychology which lend itself for curriculum wide implementation with the interactive potential of a chatbot. In line with Lattie et al., (2019) we propose that it would be of great value if these experiments were conducted in professional colleges or community colleges as well, besides research universities. It would also be highly recommended, to monitor technology acceptance, usability and implementation feasibility with validated scales. As Harrer et al. (2019) concluded, research on the effects of chatbots has so far not yet defined student subsets for which the intervention is most effective. A large scale experiment in which different student subsets are followed could, therefore, open up valuable new vistas which can further explore the added value of differentiation that a chatbot can offer.

In short, we expect that the proposed AI-enhanced life-crafting intervention will help students to overcome the difficulties they face when transitioning into tertiary education. We anticipate that it will increase students' academic performance and decrease the development of mental health problems. Future studies will need to uncover the specific effects of this intervention. Ideally, this intervention will be able to optimize both student well-being and academic achievement.

Chapter 4

The Right Job Pays: Effects of Work on the Study Progress of Pre-service Teachers

Spending time on work during a full-time study might compete with class attendance or self-study and slow study progress. At the same time, a domain-relevant job may grant beneficial effects that enhance academic outcomes. Prior research showed contradictory findings, possibly because of a lack of distinction between types of work and the different years of college. The current study analyzed the effect of different types of work on the study progress of 132 Dutch pre-service teachers with repeated measures at 25 points in time over a 4-year timespan using growth models. Students who spent more time on a paid job as a teacher obtained significantly more study credits. The optimal number of hours spent on paid work outside of education changes during college. These findings support the importance of study-job-congruence and add the roles of timing (year of college) and remuneration (getting paid) as relevant variables to role-based resource theory.

1 Introduction

Several countries cope with a shortage of qualified teachers (Donitsa-Schmidt & Zuzovsky, 2016; European Commission, 2014; Garcia & Weiss, 2019; Sutcher et al., 2016), particularly in disadvantaged areas (OECD, 2005). Subsequently, schools in need of teaching staff may opt to offer pre-service teachers a contract before they finish college. Hiring pre-service teachers could alleviate the shortage and provide pre-service teachers with valuable experience. However, it could also strain their study progress, compete with study hours or demotivate students to obtain a degree they no longer seem to require. Therefore, it might even backfire and increase the shortage of qualified teachers in the long run.

The two most recent systematic literature reviews about the effects of student employment on educational outcomes reported contradictory, but mainly non-positive effects (Riggert et al., 2006; Neyt et al., 2019). Both reviews expect that some of the differences in the findings could be due to the focus and quality of the included studies. They address several shortcomings of the current literature.

First off, the lack of experimental data leads to a risk of endogeneity bias. The statistical methods in the literature could use more rigor and need to take this risk into account (Riggert et al., 2006; Neyt et al., 2019). Students who work (more) beside their study might have certain traits in common that distinguish them from others and that explain differences in study outcomes. Neyt et al. (2019) propose using instrumental variables (IV) techniques or longitudinal data to lower this risk.

The second shortcoming is that studies should distinguish how different types of work or choices of track relate to outcomes (Neyt et al., 2019). In the case of teacher education: how does the additional time that students spent on paid and unpaid jobs in- and outside of education influence their study progress?

Third, previous studies did not distinguish between the different college years, making it unclear whether correlations are similar for different college years (Riggert et al., 2006). Given that internships are often integrated in specific years of the teacher education curriculum and can lead to a paid job as a teacher, this should be relevant. Studying this interaction requires repeated measures throughout the four-year college duration, instead of the single outcome measures (e.g., first year GPA) that are mostly used.

Finally, most studies thus far have lacked a theoretical foundation, which leads to a myriad of piecemeal exploratory findings that lack integration (Riggert et al., 2006). These concerns should be addressed in order to truly answer the practical concerns of pre-service teachers, teacher educators and policymakers, and in order to deepen our scientific understanding of the interaction between different types of work on study outcomes throughout college.

2 Theoretical Framework

In their systematic literature review, Neyt et al. (2019) conclude that student employment, on average, seems to have a negative effect on dropping out, but a non-negative effect on test and exam scores. They report three heterogeneous effects. Studies in a European context found relatively more and larger negative effects than studies in the North-American context. Working more than 15 hours predominantly relates to lower academic outcomes, while working a little can even relate to positive outcomes. Work-oriented students work more and have worse educational outcomes compared to study-oriented students.

All studies face the risk of endogeneity bias, meaning that students who work (more) might be different from unemployed students in observable and unobservable characteristics. This makes it particularly hard to discern whether student employment 'causes' any effects, or whether these are effects that should be attributed to confounding variables such as motivation or financial need. Neyt et al. (2019) use three different theories to interpret the findings from studies in the past two decades. According to the human capital theory (Becker, 1964), student employment can be a beneficial complement to education. Employment could enable the acquisition of skills (e.g., Buscha et al., 2012) and allow students to practice what they learn in theory (Geel & Backes-Gellner, 2012). Using the allocation of time or 'zero-sum' theory (Becker, 1965), on the other hand, would mean that employment can substitute time spend on classes and self-study, thereby negatively affecting educational outcomes (e.g., Stinebrickner & Stinebrickner, 2004). Human capital theory and zero-sum theory may interact: it could be that the first hours of employment generate the largest marginal benefits, while only crowding out the least important hours of studying.

Apart from these two theories that originate from the field of economics, Neyt et al. (2019) also applied one sociological theory to interpret their findings. Primary orientation theory (Warren, 2002) suggests that students whose primary orientation is towards work instead of college will perform worse than their counterparts because it reflects a form of disengagement.

Neyt et al.'s review omits theories about student employment used in the educational and psychological field. Within educational research, the theories of Tinto (1993), Bean and Metzner (1985) and their adaptation by Riggert et al. (2006), all predict a negative indirect effect of student employment on academic outcomes. In line with Zero-Sum Theory, they argue that students who spent time on work have less time to spend on 'on-campus activities', thereby experience lower social integration, which in turn leads to lower psychological and academic outcomes. In Tinto's model this can lead to the eventual 'departure decision' of dropping out. Their models can be seen as more elaborate applications of Zero-Sum Theory to the educational context.

In the field of psychology, Butler (2007) applied the theory of role-based resources to explain effects of student employment. This theory proposes that performance in multiple domains is beneficial for individuals when certain conditions are met (Marks, 1977; Greenhaus & Powell, 2006). Butler extended this theory to student employment, stating that 'job-school congruence' enriches resources, leading to work-school facilitation, study effort, and better study performance. Job demands and number of working hours, on the other hand, lead to work-school conflict and subsequently lower study effort and study performance. Evidence from Butler's study with 253 full-time American college students, and results from a few additional cross-sectional studies (Creed et al., 2015; Meeuwisse et al., 2017) support this model. In these samples of heterogeneous (different types of courses) university students, they found that job-congruence relates to work-school facilitation, which subsequently relates to study effort and study performance.

Butlers theory combines elements of human capital theory and zero-sum theory with a psychological mechanism that can explain which conditions predict the result of the trade-off. This theory is supported by empirical evidence from recent cross-sectional studies (Butler, 2007; Creed et al., 2015; Meeuwisse et al., 2017) and aligns with findings from several studies that underscore that job-characteristics matter (McNeal, 1997; Tuononen et al., 2016; Wang et al., 2010).

Butler's model could clarify some of the ambiguous results thus far. Job congruence and working hours may define whether student employment is negative or positive. Potentially this could also explain why so many studies thus far found a 'curvilinear relationship'. A curvilinear relationship entails that working a limited number of hours is better than both not working or working more hours. For instance, Wikan and Bugge (2014) reported that working 1-15 hours related to better academic outcomes for Norwegian pre-service teachers. This suggests that there might be an ideal balance between time spent on work and study. A bonus granted by control and or job-school congruence could initially lead to a positive effect, which can become harmful when too many hours lead to work-school conflict.

To substantiate these findings, in particular for teacher education, more specific research is needed. The study from Wikan and Bugge (2014) did not distinguish between types of work. Additionally, even the few studies that did distinguish between types of jobs did not specify which jobs were job-congruent (e.g., in education) and did not take the role of unpaid internships into account (Tuononen et al., 2016; Wang et al., 2010). Especially in higher vocational education, internships are both a part of the curriculum and offer a work-like experience. The responsibilities and demands of internships can lead to requests for unpaid overtime. In studies about student employment, this type of unpaid work should be taken into account. Finally, none of the models or studies thus far takes the specific character of the different years of college into account. If internships function as a stepping stone for a paid position, then it should be expected that the importance of the types of work that students engage in changes during the study program. Using a longitudinal approach could chart the development of the relationship and lower the risk of endogeneity bias.

2.1 Research Questions

This study analysed the effects of unpaid internship overtime hours, as well as hours spent on paid work in- and outside the educational field, on study progress. Effects were studied with a longitudinal approach that enabled us to assess the effects of different types of work on study progress with more precision. Specifically, it allowed us to test whether, when, and how much hours spent on student employment affected study progress. In line with Butler (2007) and Wang et al. (2010), we predicted that domain-relevance (i.e., 'job congruence') and the number of working hours matter. Additionally, we expect that the types of jobs that students have, change during college time, and we explore the effect of these different types of jobs for each separate semester. Therefore, we formulated the following research questions:

RQ 1: How does the allocation of time spent by pre-service teachers on unpaid internship overtime, paid jobs outside of education, and paid jobs as a teacher develop during the span of their study?

RQ 2: How does time spent on unpaid overtime during internships, paid jobs outside of education, and paid jobs as a teacher, relates with study progress during the 4-year span of college?

RQ 3: How much time spent on either unpaid internships, paid jobs outside of education, or paid jobs as a teacher, relates to optimal study progress during each specific semester of 4-year college?

3 Methods and Materials

3.1 Design

To measure the effect of different types of paid work and unpaid internship overtime on study progress, we used a dataset that contained the accumulated study credits of a full cohort of 132 pre-service teachers in the Netherlands at 25 time points (repeated measures) over a 4-year timespan. We combined this dataset with a survey about the average number of hours that students spent per week on different types of (un)paid work for every semester over the same 4-year period.

3.2 Inclusion and Exclusion

The studied cohort consisted of 330 pre-service teachers from 13 Bachelor study programmes within a Faculty of Education at a Dutch university of applied sciences. All pre-service teachers who started a full-time teacher education study in 2016 and still attended university in 2020 received an email with a link to an online survey at the end of their fourth year. The email stated the purpose of the study and the key elements of the data management plan. Students who accepted the online informed-consent statement were directed to the survey. All students who finished the survey received €10 for their effort. 189 students started the survey, and 142 students completed the survey. After data cleaning, 10 students who interrupted their study and therefore had incomplete data were removed, and 132 students were used in the actual dataset.

3.3 Participants Characteristics

Within the chosen cohort, 36% of the pre-service teachers were male. In the Netherlands, students from different types of previous education are admissible to teacher education at a university of applied science. Most of the respondents followed ‘higher general secondary education’ (HAVO), followed by students from a vocational track (MBO), and students who followed an academic track (VWO) prior to becoming a pre-service teacher (Table 1). These percentages correspond nearly precisely with the dispersion among previous education in the sample and are similar to national averages. The students in the sample followed 13 different teacher education courses (Elementary school, Dutch, English, French, German, Mathematics, Physics, Biology, Economics, Business Administration, Sociology, Geology, and History).

Table 1. Sample and Response Characteristics

Characteristic	Sample		Response	
	N	%	*N*	%
Gender				
Female	212	64	103	78
Male	118	36	29	22
Previous education				
HGSE	177	54	71	54
Vocational track	89	27	35	27
Academic track	63	19	26	20

Note. This table shows the characteristics of the sample compared to the realised response. HGSE stands for Higher General Secondary Education.

3.4 Instruments

The university at which the study took place records the study progress of students in a 'data warehouse'. Students received ECTS 'study credits' for the courses that they finish. Each year's programme contains 60 study credits, and 240 credits are needed to obtain a teaching degree. Each time a student receives new study credits, the new total amount of credits is recorded together with the associated date. The university information department provided us a dataset that included data about enrolment, gender, previous education, and the records of cumulatively obtained study credit at 25 repeatedly measured moments. Additionally, a survey asked students whether they have received paid employment outside or within the field of education and whether they have completed an internship during the past eight semesters. Subsequently, they filled in how much time they spent on average per week on different types of jobs and unpaid internship overtime during the past eight semesters, resulting in eight repeatedly measured indicators for hours spent on jobs and overtime. Because internships are integrated into the curriculum of the study programmes, we specifically asked how much extra time – i.e., more time than required by the study programme - they spent on unpaid internships within education. Paid work in education was also clearly separated from paid internships that were part of the curriculum. This allowed us to study the effects of both unpaid work in education and paid work in- and outside of education as separated from the study programme on the accumulation of study credits.

3.5 Analytic Strategy

Growth models were fitted by means of multilevel modelling using the programme MLwiN (Rashbash et al., 2020). In the random parts of these models, the 25 repeated measures represent the lowest variance level, which is nested within students, implying that the random parts of the growth models contain at least two levels. Because the sample consisted of students from different courses, for every analysis, we tested whether the intra–course correlation was significantly larger than zero. If so, analyses were conducted with three variance levels in the random parts of the models, a repeated measures level, a student level, and a course level. Only random intercept models were used, random slopes were not added since random slopes do not change the fixed part of the model. The significance of the fit improvement after adding the course level to the random part of the model is evaluated with the chi-square distributed difference in the deviances (-2*loglikelihood) of both nested models. As indicated in Hox et al. (2018), the probability of this chi-square must be divided by two, since variances cannot be negative.

In the growth models, time is included as an independent variable. Since occasions at which the 25 repeated measures of study credits are collected vary, the actual dates are used to construct the time variable. The first measurement for each student's time is set to zero. For each subsequent measurement of each student, time represents the days that have passed since their first measurement. To allow using the same time variable for hours weekly spent on jobs or unpaid internship overtime, the number of hours spent in a semester is repeated for all measurement moments within that semester.

The effects of the overtime during internships and work in- and outside of education are estimated as the interaction between the time factor in the growth model and the number of hours spent on respective overtime for the internship or work inside or outside education. Variances in the random parts of two nested models are compared to estimate the effect sizes; one model with the main effects of time and the number of hours worked or spent on overtime and the same model but with an added interaction between time and overtime or work hours. This interaction represents the effects of weekly hours worked on growth in study credits over time. After fitting these growth models with time, hours worked, and the interaction between both as independent variables, we fitted new growth models in which we controlled for gender and previous education. We analysed these models separately to ascertain if and how adding control variables changes the effects, given that adding covariates can spuriously diminish estimated effects when covariates are correlated with the number of hours worked.

To measure whether the relationship between different types of paid work, unpaid internship overtime hours, and study credits is curvilinear, a different set of analyses was conducted. We performed a separate regression analysis in MLwiN for each semester and for each type of work on a cross-sectional dataset. We obtained study credits per semester that functioned as dependent variables. In the first fitted models, the time spent on work outside of education, paid work as a teacher, or unpaid internship overtime hours during the first semester were included separately as independent variables. Subsequently, time squared is added to the models. By means of Wald tests (ratios of regression coefficients and their standard error) and through testing model fit improvement with the difference in deviances, we tested whether adding the squared time variable to the model significantly improved model fit. If so, the relation between time spent on the chosen type of work and study credits is curvilinear. This allowed us to infer not only if there is a curvilinear relationship but also at how many hours the exact break-even point is located for each separate semester and type of work.

4 Results

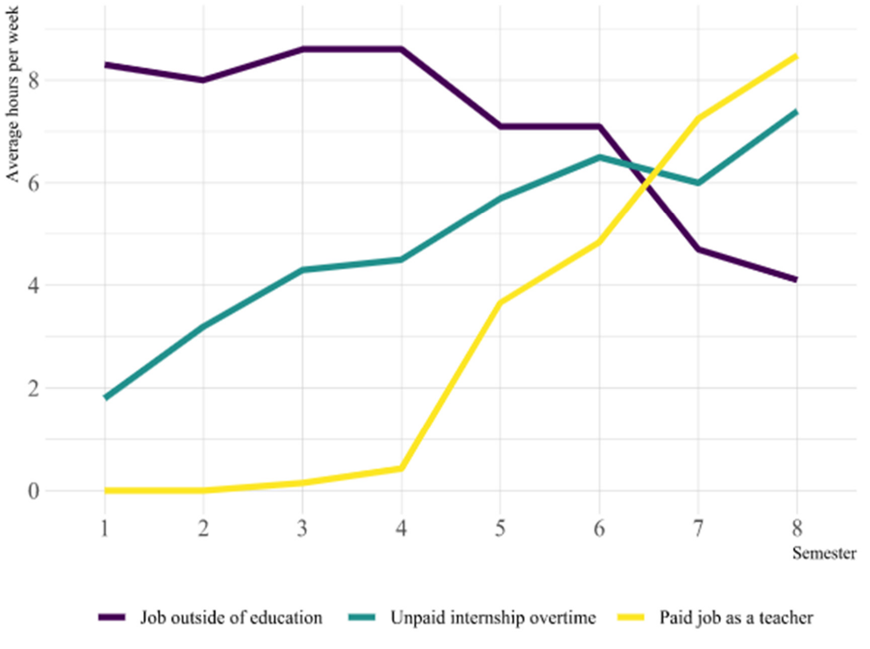
4.1 Trends in Types of Work During College

The results in Table 2 and Figure 1 show the descriptive analysis of the amount of time that students in our sample spent on overtime during internships, paid work outside of education, and paid work as a teacher over the course of eight semesters. At the start of their study, none of the students had a paid teaching job, one in four students reported overtime hours during internships, and a majority of 70.5 percent had a paid job outside of education. Throughout the four years of college, the balance gradually shifted; in the final year, 54.5 percent had a paid teaching job and 37.9 percent has a paid job outside of education. The average number of hours of unpaid overtime that students do during their internship slowly increased from *M* 1.8 hours (*SD* 5.2) in the first semester to *M* 7.4 hours (*SD* 9.0) per week in the eighth. The combined number of hours of paid work and unpaid overtime reported by students in addition to their study during an average week, gradually rose from 10.1 hours per week during the first semester up to 20 hours in the eighth semester. This suggests that on average, students partly replaced work outside the educational domain for (un)paid work within the educational domain. It also shows that the total number of hours spent on work increased during the study and that the percentage of students who work and study increased (during the last semester only 8.3% of the students did not work besides their full-time study).

Table 2. Descriptive Statistics

Semester	Paid job as a teacher %			Unpaid internship overtime %			Job outside education %		
	0 hours	1-15 hours	>15 hours	0 hours	1-15 hours	>15 hours	0 hours	1-15 hours	>15 hours
1	100	0	0	75	20.5	4.5	29.5	50	20.5
2	100	0	0	60.6	31.8	7.6	31.8	48.5	19.7
3	97	3	0	52.3	39.4	8.3	27.3	50.8	22
4	94.7	4.5	0.8	46.2	44.7	9.1	28	52.3	19.7
5	69.7	19.7	10.6	34.8	53.8	11.4	40.9	41.7	17.4
6	62.9	22	15.2	31.8	53.8	14.4	40.9	40.9	18.2
7	51.5	22	26.5	41.7	44.7	13.6	59.8	28.8	11.4
8	45.5	23.5	31.1	31.8	51.5	16.7	62.1	30.3	7.6

Figure 1. Time Spent on Different Types of Student Employment During 4 Years of College



4.2 Findings from Growth Models

Internship overtime hours show a significant but small positive effect on growth in study credits (Table 3, model 3; interaction overtime*time). Paid work outside of education has a small but nonsignificant negative effect on growth in study credits over time (Table 3, model 4 and 5). Paid work as a teacher has a significant positive effect on growth in study credits during the last two years of college (Table 3, model 6 and 7). This is in accordance with the fact that almost no student reported having a paid job as a teacher during the first two years of college (Table 2). When both the interaction effects of growth in time with respectively internship overtime, paid work outside of education and paid work as a teacher are added to the model, only paid work as a teacher shows a significant positive effect on growth in study credits (model 9). This means that paid work in education did not hinder study progress of these preservice teachers. The positive effect is small though, given that the aggregate proportion of explained variance of model 8 and 9 amounts to only 1.76%.

Table 3. Effects of Different Types of Work on Growth in Study Credits Over Time

Effect	Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
		Fixed effects								
Intercept	γ_{00}	-1.23 (1.02)	-1.34 (1.02)	-0.89 (1.04)	0.03 (1.10)	-0.69 (1.20)	-0.72 (1.02)	-0.49 (1.02)	-0.33 (1.42)	-0.04 (1.22)
Time	γ_{01}	54.11*** (0.19)	54.03*** (0.20)	53.81*** (0.22)	53.96*** (0.19)	54.26*** (0.27)	53.52*** (0.22)	53.43*** (0.22)	52.98*** (0.28)	53.36*** (0.33)
Internship overtime	γ_{10}		0.06 (0.05)	-0.08 (0.08)					0.03 (0.05)	-0.08 (0.08)
Internship overtime*Time	γ_{11}			0.06* (0.03)						0.04 (0.03)
Job outside of education	γ_{20}				-0.13* (0.05)	-0.04 (0.08)			-0.06 (0.05)	-0.02 (0.08)
Job o.o.e. *Time	γ_{21}					-0.04 (0.03)				-0.02 (0.03)
Paid job as teacher	γ_{30}						0.24*** (0.05)	-0.39 (0.20)	0.21*** (0.05)	-0.36 (0.21)
Paid job as teacher.*Time	γ_{31}							0.19*** (0.06)		0.17*** (0.06)
		Random effects								
Student variance	μ_{0j}	112.15 (14.71)	111.95 (14.69)	111.87 (14.67)	110.64 (14.40)	110.33 (14.38)	110.20 (14.40)	110.16 (14.38)	109.37 (14.19)	109.32 (14.23)
Repeated measures variance	e_{0ij}	156.92 (3.96)	156.85 (3.96)	156.64 (3.96)	156.58 (3.97)	156.48 (3.96)	155.71 (3.94)	155.21 (3.93)	155.65 (3.95)	155.02 (3.93)
Total variance	$\mu_{0j}+ e_{0ij}$	269.07	268.80	268.52	267.22	266.81	265.91	265.37	265.01	264.34
% expl. var. student level				0	1.35		1.74	0.04	2.49	0.04
% expl. var. rep. meas. level				0.13	0.22		0.77	0.32	0.81	0.40
% expl. var. total				0.10	0.68		1.18	0.14	1.51	0.25
		Goodness of fit								
Deviance		26000.31	25998.77	25994.56	25991.87	25989.57	25973.97	25964.06	25971.82	25959.14
Sig. difference of fit compared to model			Model 1 $\chi^2_{(1)} =$ 1.54	Model 2 $\chi^2_{(1)} =$ 4.22*	Model 1 $\chi^2_{(1)} =$ 8.45**	Model 4 $\chi^2_{(1)} =$ 2.30	Model 1 $\chi^2_{(1)} =$ 26.34***	Model 6 $\chi^2_{(1)} =$ 9.91**	Model 1 $\chi^2_{(3)} =$ 28.49***	Model 8 $\chi^2_{(3)} =$ 12.69**

Note. Dependent variable is study credits, measured 25 times (repeated measures N = 3,245; student N = 132; Course N = 13) (*SE* between brackets). Independent variables are number of hours spent on unpaid internship overtime, a job outside of education and paid job as teacher, respectively, each measured 8 times and added to the time factor of the model by repeating the reported hours of a semester on each repeated measure within that semester. The time variable represents the dates of each of the 25 repeated measures of study credits (dependent variable).*$p < .05$ **$p < .01$ ***$p < .001$

Table 4. Covariates Effects on Growth in Study Credits

Effect	Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
		Fixed effects				
Intercept	γ_{00}	-1.23 (1.02)	-0.61 (1.13)	-1.85 (1.15)	-3.19 (1.40)	-1.11 (1.44)
Time	γ_{01}	54.12*** (0.19)	54.11*** (0.19)	54.73*** (0.21)	54.73*** (0.21)	53.70*** (0.27)
Male	γ_{10}		-2.83 (2.28)	2.91 (2.45)	3.44 (2.40)	3.14 (2.40)
Male*Time	γ_{11}			-2.83*** (0.45)	-2.83*** (0.45)	-2.69*** (0.44)
Vocational track	γ_{20}				-0.29 (2.17)	-2.45 (2.35)
Academic track	γ_{30}				6.59*** (2.41)	-0.81 (2.60)
Vocational*Time	γ_{21}					1.07*** (0.43)
Academic*Time	γ_{31}					3.67*** (0.48)
		Random effects				
Student variance	μ_{0j}	112.15 (14.71)	110.78 (14.55)	110.88 (14.55)	103.86 (13.56)	104.22 (13.59)
Repeated measures variance	e_{0ij}	156.92 (3.96)	156.92 (3.96)	154.90 (3.91)	154.90 (3.93)	152.08 (3.86)
Total variance	$\mu_{0j}+ e_{0ij}$	269.07	267.70	265.79	258.76	256.30
% expl. var. student level				-	6.33	0.35
% expl. var. rep. meas. level				1.28	0	1.82
% expl. var. total				0.71	2.64	0.95
		Goodness of fit				
Deviance		26000.31	25998.77	25958.62	25950.45	25893.56
Sig. difference of fit compared to model			Model 1 $\chi^2_{(1)} = 1.54$	Model 2 $\chi^2_{(1)} = 41.69$***	Model 3 $\chi^2_{(2)} = 8.17$*	Model 4 $\chi^2_{(2)} = 56.90$***

Note. Dependent variable is study credits, measured 25 times (repeated measures $N = 3{,}245$; student $N = 132$; Course $N = 13$) (*SE* between brackets). Independent variables are gender and previous education (general higher secondary education is the comparison). *$p < .05$ **$p < .01$ ***$p < .001$

Table 5. Effects of Different Types of Work on Growth in Study Credits Over Time with Covariates

Effect	Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed part							
Intercept	γ_{00}	-1.11 (1.44)	-1.19 (1.44)	0.34 (1.41)	-0.58 (1.42)	0.03 (1.48)	0.94 (1.56)
Time	γ_{01}	53.70*** (0.27)	53.57*** (0.28)	53.53*** (0.27)	53.79*** (0.28)	53.01*** (0.29)	52.57*** (0.37)
Male	γ_{10}	3.14 (2.40)	3.21 (2.40)	3.78 (2.40)	3.38 (2.37)	3.71 (2.37)	3.96 (2.38)
Male*Time	γ_{11}	-2.69*** (0.44)	-2.70*** (0.44)	-2.85*** (0.45)	-2.90*** (0.44)	-2.96*** (0.44)	-3.08*** (0.45)
Vocational track	γ_{20}	-2.45 (2.35)	-2.60 (2.35)	-2.55 (2.34)	-2.35 (2.32)	-2.51 (2.32)	-2.51 (2.32)
Academic track	γ_{30}	-0.81 (2.60)	-1.10 (2.61)	-1.19 (2.60)	-0.86 (2.57)	-1.22 (2.57)	-1.39 (2.58)
Vocational*Time	γ_{21}	1.07*** (0.43)	1.09** (0.43)	1.00** (0.43)	0.91* (0.43)	0.91* (0.43)	0.94* (0.43)
Academic*Time	γ_{31}	3.67*** (0.49)	3.73*** (0.49)	3.66*** (0.48)	3.76*** (0.48)	3.78*** (0.48)	3.86*** (0.49)
Overtime internship	γ_{40}		0.09 (0.05)			0.05 (0.05)	0.03 (0.08)
Overtime Internship*Time	γ_{41}						0.01 (0.03)
Student employment o.e.	γ_{50}			-0.16*** (0.05)		-0.08 (0.05)	-0.16* (0.08)
Student employment*Time	γ_{51}						0.04 (0.03)
Work in education	γ_{60}				0.28*** (0.05)	0.26*** (0.05)	-0.46** (0.21)
Work in education*Time	γ_{61}						0.22*** (0.06)
Random part							
Student variance	μ_{0j}	104.22 (13.59)	104.27 (13.72)	103.46 (13.62)	101.34 (13.26)	101.03 (13.20)	101.25 (13.15)
Repeated measures variance	e_{0ij}	152.08 (3.86)	151.91 (3.84)	151.54 (3.83)	150.44 (3.81)	150.27 (3.81)	149.57 (3.80)
Total variance	$\mu_{0j}+ e_{0ij}$	256.30	256.18	255.00	251.78	251.30	250.83
Deviance		25893.56	25890.07	25881.48	25856.23	25852.19	25838.00
% expl. var. student level			-	0.70	2.76	3.06	-
% expl. var. rep. meas. level			0.11	0.36	1.08	1.19	0.47
% expl. var. total			0.004	0.50	1.76	1.95	0.19
Sig. difference of fit compared to model			Model 1 $\chi^2_{(1)} = 3.49$*	Model 1 $\chi^2_{(1)} = 12.08$**	Model 1 $\chi^2_{(1)} = 37.33$***	Model 1 $\chi^2_{(3)} = 41.37$***	Model 5 $\chi^2_{(3)} = 14.19$**

Note. Dependent variable is study credits, measured 25 times (repeated measures $N = 3{,}245$; student $N = 132$; Course $N = 13$) (*SE* between brackets). Independent variables gender, previous education, work outside of education, overtime internships, and work in education based on input per semester (8 times). *$p < .05$ **$p < .01$ ***$p < .001$

Because we wanted to control for gender and previous education, we tested the effect of gender and previous education on growth in study credits over the four-year time span (Table 4). Especially previous education showed to be a predictor of growth in study credits, it explained 6.7 percent of the variance.

While controlling for gender and previous education, we tested the same models that included work outside of education, unpaid internship overtime hours, and paid work as a teacher (Table 5). This confirmed our earlier findings. Only paid work as a teacher proved to show a significant but limited positive effect on growth in study credits (Table 5, model 6). Again, this positive effect applied only to the last two years of college.

Finally, to answer RQ 3, we wanted to test if there exists a curvilinear relationship between either of the types of work and study progress and accordingly define the break-even point. We analysed this for each individual semester, which allowed us to ascertain whether effects differ between semesters. The results in Table 6 (Model 3) show that working outside of education during the first semester does not significantly relate to more study progress. But when student employment hours squared is added to the model, both the first and second-order of employment hours significantly relate to study progress, which confirms a curvilinear relationship. The largest increase in study credits is found for students working 7.75 hours per week outside of education. There is no difference in terms of study credits between students working 15.5 hours per week outside of education and students that do not have a side-job. Students working more than 15.5 hours per week outside of education receive fewer credits than students who do not work outside of education and the more hours these students work per week, the more negative the relation between work and the number of study credits becomes (Figure 2). Adding paid work outside of education squared in model 4 (Table 6) explains 6.4% of all variance at the student level, but leads to more course and total variance.

Figure 2. Effect of Hours per Week Spent on Paid Work Outside of Education on Obtained Study Credits in the First Semester

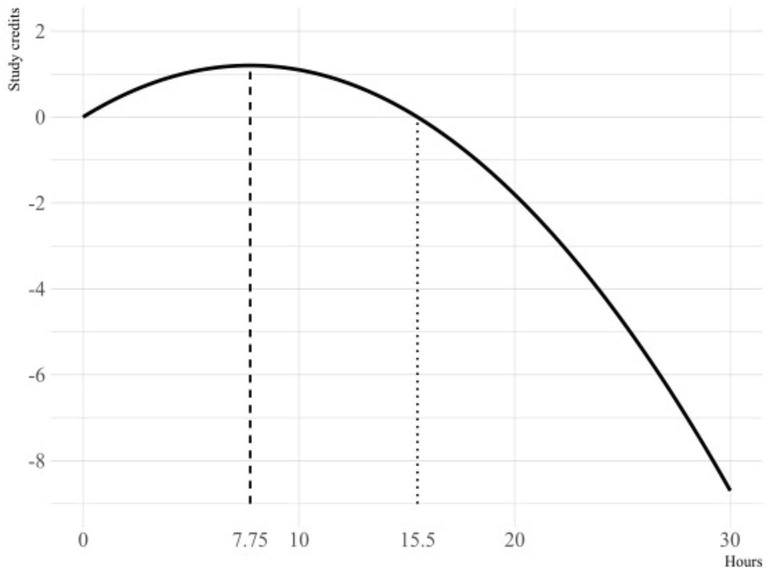


Table 6. Effect of a Job Outside of Education on Study Progress in the First Semester

Effect	Parameter	Model 1	Model 2	Model 3	Model 4
		Fixed effects			
Intercept	γ_{00}	26.55 (0.56)	26.35 (1.26)	27.02 (1.40)	26.01 (1.51)
Job o.o.e.	γ_{10}			-0.08 (0.07)	0.31* (0.17)
Job o.o.e.^2	γ_{20}				-0.02** (0.01)
		Random effects			
Course variance	μ_{0j}		15.81 (7.92)	15.91 (7.95)	18.58 (8.94)
Student variance	e_{0ij}	42.61 (5.12)	27.20 (3.52)	26.90 (3.45)	25.28 (3.27)
Total variance	$\mu_{0j+}\ e_{0ij}$	42.61	43.01	42.81	43.86
% expl. var. course level					n.a.
% expl. var. student level					6.40
% expl. var. total					n.a.
		Goodness of fit			
Deviance		866.75	832.61	831.34	825.35
Sig. difference of fit compared to model			Model 1 $\chi^2_{(1)}$ = 34.14***	Model 2 $\chi^2_{(1)}$ = 1.27	Model 3 $\chi^2_{(1)}$ = 5.99*

Note. Dependent variable is study credits at Semester 1 (student N = 132; Course N = 13) (*SE* between brackets). Independent variable is overtime internships measured 25 times based on input per semester (8 times). $^*p < .05$ $^{**}p < .01$ $^{***}p < .001$

We conducted the same analysis for other semesters and also found a significant curvilinear correlation between time spent on a paid job outside of education and obtained course credits during the third semester (Table 7, Model 4). In the third semester, paid work outside of education and its squared version together predict 7.93% of all variance in study credits at the student level and 5.68% of all total variance in study credits (Table 7, Model 5). In this case, 8.25 hours of paid work outside of education correlated with the largest net gain in study credits, and the break-even point is 16.5 hours. Internship overtime did not correlate significantly with obtained credits during any of the semesters.

Having a paid (congruent) job as a teacher shows a positive significant effect on the gain in study credits especially during the fifth semester (Table 8, Model 3). Interestingly, this relation is not curvilinear within semesters (Table 8, Model 4). As far as the range of our dataset permits (with 30 hours as the highest reported amount), more hours spent on paid work as a teacher during the fifth semester simply relates to more obtained study credits.

Table 7. Effect of a Job Outside of Education on Study Progress in the Third Semester

Effect	Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
		Fixed effects				
Intercept	γ_{00}	24.16 (0.72)	24.39 (1.36)	25.86 (1.60)	24.36 (1.68)	24.36 (1.68)
Job o.o.e.	γ_{10}			-0.17* (0.09)	0.33 (0.22)	0.33 (0.22)
Job o.o.e.^2	γ_{20}				-0.02* (0.01)	-0.02* (0.01)
		Random effects				
Course variance	μ_{0j}		15.07 (8.94)	16.00 (9.27)	15.37 (8.89)	15.37 (8.89)
Student variance	e_{0ij}	67.82 (8.35)	56.51 (7.30)	54.82 (7.08)	52.36 (6.76)	52.36 (6.76)
Total variance	$\mu_{0j}+ e_{0ij}$	67.82	71.58	70.82	67.73	67.73
% expl. var. course level					4.10	n.a.
% expl. var. student level				3.08	4.70	7.93
% expl. var. total				1.07	4.56	5.68
		Goodness of fit				
Deviance		931.22	922.05	918.78	912.78	912.78
Sig. difference of fit compared to model			Model 1 $\chi^2_{(1)} =$ 9.17**	Model 2 $\chi^2_{(1)} =$ 3.27#	Model 3 $\chi^2_{(1)} = 6.00$*	Model 2 $\chi^2_{(2)} = 9.27$**

Note. Dependent variable is study credits at Semester 3 (student *N* = 132; Course *N* = 13) (*SE* between brackets). Independent variable is overtime internships measured 25 times based on input per semester (8 times). #$p < .1$ *$p < .05$ **$p < .01$ ***$p < .001$

Table 8. Effect of a Paid Job as a Teacher on Study Progress in the Fifth Semester

Effect	Parameter	Model 1	Model 2	Model 3	Model 4
		Fixed effects			
Intercept	γ_{00}	21.69	22.47	20.30	29.87
		(0.80)	(1.33)	(0.90)	(1.27)
Paid job as teacher	γ_{10}			0.38***	0.32
				(0.12)	(0.40)
Paid job as teacher^2	γ_{20}				0.00
					(0.02)
		Random effects			
Course variance	μ_{0j}		11.93		
			(8.38)		
Student variance	e_{0ij}	84.85	76.34	79.17	79.15
		(10.44)	(9.83)	(9.75)	(9.74)
Total variance	$\mu_{0j+}\ e_{0ij}$	84.85	88.27	79.17	79.15
% expl. var. course level					
% expl. var. student level				7.17	
% expl. var. total				7.17	
		Goodness of fit			
Deviance		960.80	957.74	951.64	951.62
Sig. difference of fit compared to model			Model 1 $\chi^2_{(1)}$ = 3.06	Model 1 $\chi^2_{(1)}$ = 9.16**	Model 5 $\chi^2_{(1)}$ = 0.02

Note. Dependent variable is study credits at Semester 5 (student N = 132; Course N = 13) (*SE* between brackets). Independent variable is overtime internships measured 25 times based on input per semester (8 times). $*p < .05$ $**p < .01$ $***p < .001$

5 Discussion

In this article, we described how much time students in teacher education spent on different types of work during four years of college. We also analysed how time spent on either type of work relates to study progress and specified how many hours related to study progress during every specific semester of college. Our results show that many pre-service teachers take on a paid job as a teacher by the third year and that this relates to significantly more study progress over time. Time spent on unpaid internship overtime or paid work outside of education does not significantly relate to study progress over the span of four years of college. However, we did find that working respectively 7.75 and 8.25 hours per week in a (non-congruent) job outside of education is connected to obtaining the optimal amount of study credits during the first and third semester of college.

In accordance with Wang et al. (2010) and Tuononen et al. (2016), we found that it does matter what type of work students engage in during their study. Wang et al. surveyed third-year students and found that those who chose a job that they considered relevant to their study averaged a higher GPA. Their results align with our finding that a paid job as a teacher significantly improves study progress in the last two years of college. Wikan and Bugge (2014) reported a curvilinear relationship between the average number of hours spent on paid work and the most recently received grade (self-report) during the first year of teacher education. Their results coincide with the curvilinear relationship we found between paid work outside of education and earned study credits in the first semester. This study confirms both previous findings and shows how these correlations apply to specific semesters during college. Thereby this study offers a more suitable statistical methodology for future research and a more comprehensive explanation of how different types of work relate differently to study progress through time. Findings from both this study and the studies by Tuononen et al. (2016), Wang et al. (2010), and Wikan and Bugge (2014) contradict the main assumptions about the influence of student employment on academic outcomes in the theoretical models of Tinto (1993), Bean and Metzner (1985), and their adaptation by Riggert et al. (2006). These three models all predicted a negative indirect effect of student employment on academic outcomes, because students who spent time on work have less time to spend on 'on-campus activities', thereby experience lower social integration, which in turn leads to lower psychological and academic outcomes. The models from Tinto (1993), Bean and Metzner (1985), and Riggert et al. (2006) are based mostly on research about North-American research universities with on-campus residents. The interaction between different types of student-employment and academic outcomes at vocational or applied universities might need a different model that takes internships and the different types of employment into account. We, therefore, propose adding remuneration, intensity and timing (semester) as factors to the conceptual model of Butler (2007). During their study, students make choices on how to allocate their time. Given that most students face financial needs, they will seek available employment (Humphrey 2006). Students will seek to optimise the balance between employment and their study. Once the opportunity arises, we found that many pre-service teachers opt to trade their job outside of education for a domain-relevant job. For those who need the income from work, paid work in education can replace paid work outside of education, but unpaid work cannot. This might explain the positive effect of paid work in education compared to unpaid work in education. It also offers an addition to Butler's model, which did not distinguish between voluntary and paid work. The positive relation between a paid job as a teacher and study progress might be explained by several potential explanations. It could be explained by a positive spill-over effect in accordance with the role-based resources theory (Butler 2007), or it could be that this type of job is offered to more successful students. Future studies should seek to clarify which direction the correlation is headed (e.g., by studying which students are offered a domain-relevant job) and could integrate these findings in a broader theory that is suited for vocational education and universities of applied science. Before generalising to a broader theory that includes fields outside of the educational domain, studies should first explore what specific type of work students perform in a specific domain during the different years of their study.

Several aspects of this study influence what conclusions can be drawn. The sample in our study only contained students who did not drop out of the study programme. Consequently, potential effects of student employment on dropping-out have not been included. The demographic and study progress variables in this study are directly generated from the university administration, which makes their reliability optimal. But the number of hours that students spend on work is based on self-report through a survey at one point in time, which is generally less reliable. Students might overreport or underreport how much time they spent on work because of desirability or because they may have trouble remembering exactly how many hours they worked in a given time span. During the time span of the study, there was a teacher shortage in The Netherlands. With fewer vacant positions, the percentage of students who had a paid job as a teacher during their study may be significantly lower.

5.1 Practical implications

This study has two main implications. First, we falsified the assumption that accepting a paid job as a teacher hinders study progress. On the contrary, a paid job as a teacher during the study seems to slightly enhance study progress. Possibly because of job congruence and because it allows students to quit their paid job outside of education. This is relevant information for policymakers who deal with a teacher shortage, teacher educators who worry about study progress, school boards that consider hiring pre-service teachers, and pre-service teachers who might wonder whether they should accept the job offer. Secondly, we found a curvilinear relationship for paid work outside of education during specific semesters of college. With our method of analysis, we could deduce that respectively 7.75 and 8.25 hours per week is optimal and that more than 15.5 and 16.5 hours relates to less study progress during the first and third semester. This is useful information for study advisors and aspiring pre-service teachers who wonder how much time they should preferably spend on paid work outside of education at the start of their study.

6 Conclusion

This study found that many pre-service teachers swap a job outside of education for a paid job as a teacher during the course of their study. Hardly any of the students have a paid job as a teacher during the first two years, but during the third and fourth year, most do. Additionally, the majority of pre-service teachers log extra hours of unpaid overtime during their internship. During the first semester and the third semester, when paid jobs as a teacher do not yet occur, having a paid job outside of education roughly one day a week relates to optimal study progress. During the fifth semester, time spent on a paid job as a teacher relates to more study progress. Although further examinations of replicability in other types of education are needed, our findings suggest two important additions to role-based resource theory. Not only does it matter whether or not student employment is congruent with the study, but it also matters whether students get paid and in which of the four years of college these effects are studied. Students and policymakers alike should take note of these findings in order to optimise both study progress and student employment.

General Discussion

The chapters that comprise this dissertation share a commonality in their potential contribution to 'academic thriving' through evidence-based higher education. Academic thriving stands for a combination of academic outcomes, such as course credits or retention, with positive developments in other important life domains during college, such as well-being, or finding 'the right' job. In this general discussion, I aim to slightly deviate from a 'traditional' discussion, by contextualising three of the conducted studies. In addition to restating the most important scientific contributions, I will also describe how three studies were conducted and how they impacted educational practice, i.e., the process of 'valorisation'. In the educational domain, there is much ado about the 'research-practice gap' (Akkerman et al., 2021). In order to bridge this gap, researchers are expected to improve the impact or valorisation of their research and educational professionals are expected to apply scientific findings in their practice. I was an educational professional before and during my part-time PhD project, and tried to be mindful of how my research should connect to educational practice. Yet, bridging the alleged gap indeed sometimes proved challenging. By explicitly describing and reflecting on the process and impact of three studies as 'cases' (I'm excluding chapter 1 because it did not have the chance to have any impact yet), I am trying to learn from the pitfalls and problems my colleagues and I encountered. In doing so, I also hope to contribute to constructive future research-practice collaborations.

The first chapter of this dissertation described how the scientific (epistemic, economic, normative) challenges that evidence-based education (EBE) recently faced can be overcome. It also described what types of research could take the criticism on EBE into account and still contribute to EBE. Thereby, this chapter prescribed what the studies in this thesis should aspire to do. It did not, however, discuss the organizational challenges that implementing scientific outcomes entail. As mentioned before, this discussion will therefore focus on chapters 2, 3, and 4.

Chapter 2: Will Goal-Setting Work Here?

Scientific Contribution

Chapter two described the results of a large-scale field experiment into the effects of a reflective type of goal-setting. We tested the effects of the intervention on the academic performance, well-being, self-regulated learning (SRL), grit, resilience, and engagement of first-year students in teacher education and business studies of a Dutch university of applied science. Students in the treatment group earned significantly more course credits and had a markedly reduced risk of dropping out of the programme compared to the control group. These effects were independent of domain (teacher vs business), gender, ethnicity, and previous education. We found no treatment effects on SRL, grit, resilience, engagement, or well-being. Although well-being was not improved by the intervention, apparently the gains in academic performance did also not appear to have adverse side-effects on well-being.

This study contributed to the literature in several ways. Contrary to the Dobronyi et al. (2019) study, this study successfully replicated the effect found by Morisano et al. (2010) and Schippers et al. (2015). The effect sizes of the treatment effects were lower, but this should be expected with a large-scale rigorous field experiment (Greenberg & Abenavola, 2017). In line with Kraft (2020), an effect of .11 standard deviation can even be considered large when: 1) it's derived at with a large-scale experiment that used broad performance measures months after the intervention took place 2) the intervention is scalable: it can be send to ten students as easily as to one million students 3) the intervention is low cost: the intervention costs students less than 2 hours. Additionally, our results indicate that the intervention can also work in the domain of teacher education. Previously the intervention was studied mainly with business and economics students. Finally, we found that the intervention had a durable effect that improved over time. The students in the treatment group averaged more course credits both after the first semester and second semester. The effect on dropout became significant after a year and not after one semester. This result most likely means that the treatment improved course credits, which then allowed the students to continue their enrolment.

The differences in the findings of Dobronyi et al. (2019) and this study could be due to several causes. RCTs ideally create a *ceteris paribus* situation in which the only difference between the treatment and control group is the intervention. However, as Morrison (2021), Joyce and Cartwright (2020), and others pointed out, there can be many contextual derailers or support factors in place that influence the hypothesized causal link. This is sometimes referred to as the 'causal cake' or 'web' within which a causal mechanism functions (Cartwright & Hardie, 2012). Causal cakes or webs might play a role when results from an intervention cannot consistently be replicated in different contexts. Context-centred research and implementation science can be used as methods to account for local factors and bridge the gap between general efficacy in research and local effectiveness. This approach entails that local educators and school leaders influence the research agenda from the start, resulting in both scientific outcomes as well as practical answers to local questions. Implementation science stresses the need for monitoring if the intervention is executed and experienced as intended.

Context-Centred Research Approach. In line with Joyce and Cartwright (2020) we organised a context-centred research approach and monitored implementation fidelity in addition to our RCT design. The execution of the field experiment in chapter 2 was done in cooperation with school leaders and teachers from the different participating courses of study. This started well before the PhD project began. Before choosing my PhD topic, I talked with school leaders and teachers, they mentioned being interested in the intervention in the Schippers et al. (2015) study because of its large impact on academic performance. I organised meetings at every organizational level, and included only the courses in which both teachers, coordinators, managers and directors agreed with cooperating in the experiment. Cooperating meant using a rigorous experimental design which would make the study relevant to the scientific field, but also allowing and facilitating teachers who were interested to join the research team. Six teachers joined the research team as a form of professional development or used this study or some parts of additional data for their own masters courses. For example, two teachers wanted to find out what strategies coaches used to increase academic performance. We 'teamed up' and interviewed 19 coaches about their strategies while also asking what types of elements from the goal setting intervention they might already applied before the experiment. This allowed us to monitor 'program differentiation' (an aspect of implementation fidelity) while also achieving the goals of the teachers. Together we presented the findings from this sub-study on a scientific conference (Dekker et al., 2021).

Implementation Science. As mentioned before, many contextual factors can interact with the hypothesized causal mechanism. Carefully monitoring implementation fidelity can provide a method to render these contextual variables less opaque and help explain effects (Durlak, 2015; Durlak & DuPre, 2008). For the study in chapter 2 we used the model that Horowitz et al. (2018) adapted from Carroll et al. (2007), and Dane and Schneider (1998). This entails monitoring: 1) program differentiation, 2) dosage, 3) adherence, 4) quality of delivery, 5) student responsiveness, and 6) fidelity of receipt. Based on our quantitative and qualitative measures of these factors, we assessed the fidelity as moderate. Differences in the findings from our study, the Morisano et al. (2010), the Schippers et al. (2015; 2020), and the Dobronyi et al. (2019) study could be due to differences in implementation fidelity. Our application of implementation science can be applied by future replications to increase the understanding of the degree to which different types of implementation matter when introducing this goal-setting intervention.

Advice for Practitioners. The picture that emerges from the findings in this study, is that the reflective type of goal-setting that we tested is a low-cost and scalable intervention that can offer significant improvements to academic performance and retention if implemented with moderate or higher fidelity. Although future studies should test whether the findings can be replicated across different contexts it should be warranted to implement the intervention (whilst monitoring fidelity) in the curriculum in similar contexts given the beneficial cost-benefits ratio. Reasoned from the evidence-based education framework from Davies (1999), this contributed to the establishment of sound evidence for the application of this intervention in similar (and, at the very least, the same) contexts. The following section will discuss whether this also led to a durable application of these findings thus far.

Practical Impact

The experiment took place at the start of the 2018/2019 academic year. Based on preliminary results, it seemed to have a positive effect, and the participating courses wanted all first-year students in the 2019/2020 year to receive the intervention as well. Practical success, which later tuned out to be significant improvements according to the analyses, led to a demand for further implementation.

The chosen context-centred research approach eased the process of implementation. The involvement of teachers and coordinators from each course, allowed them to organise aspects of the implementation themselves. Working closely together with the original investigators, as well having a team of student assistants played a crucial role in the execution of the experiment and further implementation. There were several responsibilities, however, that still required my involvement. Someone needed to coordinate further implementation across the different courses (in order for the intervention to be send out to the right students at the right time). There was turnover in key positions among the teaching staff of participating courses of study. Teachers that were involved in the implementation of the intervention in 2018 were often assigned to different tasks in 2019. Moreover, someone needed to tend to additional courses who showed interest in implementing the interventions.

Because of my background in educational policy, I had combined teaching positions with a job as senior policy advisor for a university, I looked for organizational solutions. The university where the research took place had gone through several policy developments that were relevant to the matter at hand. In the 80's, the Dutch universities of applied sciences merged from 348 smaller 'schools' into 85, and later 36, large universities (Goedegebuure & Lynn Meek, 1991; Vereniging Hogescholen, 2021). During and after this period, the university of applied sciences where this research took place, had prioritised centralising processes in order to increase efficiency. Services were centralised, as well as the educational design: every course of study was organised according to one centrally developed blue-print

(Bruggeman, 2012). If we wanted to perform an RCT or scale-up findings from an experiment, this would have fitted neatly within that strategy.

However, from 2012 on, a new chairman of the board had altered the strategy. In line with aspects of New Public Management (NPM), teams were made accountable for outcome targets (student evaluations, academic performance, external quality assessments, and alumni evaluations), but gained more freedom to choose their own methods and design (Ferlie et al., 2008). From 2016 on, the emphasis shifted to the autonomy and responsibility of the professionals, in line with post-NPM (Reiter & Klenk, 2019), this meant a more horizontal and less hierarchical culture and more room for decentral initiatives (Bormans & Dekker, 2016). Within this context, professional teams can experience more control and responsibility for the quality of the services they render (Laloux, 2015), but mounting large-scale experiments and scaling-up evidence-based interventions becomes more complicated. The context-centred research approach had made mounting the experiment possible, what could be an organizational solution for scaling up? According to Deiglmeier and Greco (2018), many proven social innovations struggle to scale up because of 1) inadequate funding 2) fragmented ecosystems and 3) talent gaps. The matter of funding was relevant, in order to delegate my temporary task of scaling-up, I needed funding to assign a coordinator or project manager. Although the board, different directors, managers and teachers agreed we should organise and fund this, it took more than a year before the directors and middle managers agreed which budget should be used for this purpose. This relates to the second point of 'fragmented ecosystems'. Several courses wanted to use the intervention but did not feel for appointing a university-wide project-manager. The university had a central department that was responsible for advising and coordinating educational change and quality assurance, but because of a more decentralised strategy this department no longer designed or coordinated any central educational designs. Its budget was partly relocated to courses and faculties who could buy their services through service-contracts. One central program from this university supported 'experiments' (innovations which did not involve experimental research designs) that were smaller in size and could therefore not be used. Another central program had initially supported this project but ceased to exist. In its stead, the university organised a new central program that intended to increase academic performance by offering courses of study that were willing to implement an integrated approach (both course-design, didactics, assessment-policies, etc.) advice and support for innovation. The program manager wanted to include this intervention in her program, but was not allowed to offer the intervention separately, which was the desire of the courses who wanted to use it. The third problem -talent gap- also applied to this case: the teachers who were responsible for implementing the intervention changed position in several of the courses, the advisor from the central department who was suited to take on this project retired, and I was trying to delegate my involvement as soon and as well as I could.

Fortunately this was not the end of the story. The manager of the central department, the program manager of a central program, and I, created a special temporary 'unit' that organised the further scaling-up of this intervention and future interventions. This unit eventually appointed a project-manager and organised the necessary financial and legal requisites for further implementation. At the start of the academic year of 2021-2022 several courses from various faculties plan to use the intervention.

From the aftermath of this study I've learned that local research-practice collaborations can prove fruitful grounds for evidence-based innovation, but that additional central organisational infrastructures are required for the durable implementation and scaling-up of evidence-based interventions. Modern horizontal organisational structures can be favourable to small-scale experimentation and flexibility, but can be adverse to the scaling-up of evidence-based innovations. Context-centred research approaches can increase the likeliness of the eventual implementation of studied interventions, but even before knowing whether an intervention will work, it is advisable to device a contingency plan that will allow for central coordination of potential follow-up and scale-up projects. Especially in research-practice collaborations, central policies are needed to stimulate and coordinate scaling-up. These lessons are particularly relevant in contexts such as The Netherlands, with highly decentralised policies and distributed autonomy (Frankowski et al., 2018).

Chapter 3: AI-Enhanced Goal-Setting

Scientific Contribution

During the qualitative focus-groups that evaluated the RCT in chapter 2, students remarked that they would have liked better and more personal follow-up on the initial goal-setting assignments. In the narrative review in chapter 3, we explored the potential of combining the latest findings from reflective goal-setting literature with findings from other fields that could offer a suitable form of follow-up.

Recently, online mental health interventions for university students showed rapid and promising developments (Abd-Alrazaq et al., 2019). AI-enhanced chatbots, in particular, can offer personalised experiences and support (Fulmer, 2019). Most interventions aimed at improving mental health of students, however, bear a stigma and do not reach the right students in time (Clement et al., 2015). Integrating a goal-setting intervention, with the technology and personalisation and follow-up potential of mental health chatbots, could enhance the impact of goal-setting and the reach of mental-health chatbots (Schippers & Ziegler, 2019; Dekker et al., 2021). Research on this new combination of interventions should use design principles that increase user-friendliness and monitor the technology acceptance of its participants. This chapter synthesized developments from different fields and offered a launchpad for new research with the introduction of a conceptual model and guidelines for research into the effects of interventions that integrates aspects from these fields.

Practical Impact

Chapter 3 offered a narrative review on the potential of delivering educational and psychological interventions through an AI-enhanced chatbot. With the use of a research grant, the IT-team from the Rotterdam School of Management (RSM) developed the chatbot together with students from both involved universities. A team of researchers from the Erasmus Centre for Study and Career success' conducted the first trials with small samples of students. After two test-trials, the chatbot was functional enough to be experimented with at the start of the 2020-2021 academic year. Our team organised a large-scale field experiment with RSM first-year students (n = 1,402) and a parallel large-scale field experiment with first-year students from the Rotterdam University of Applied Sciences (RUAS) (n = 1,281). The students received the same treatment and surveys simultaneously, but at RSM it would be part of the curriculum, and at RUAS it would be a voluntary follow-up of the goal-setting intervention. In a similar fashion to chapter two, we used a context-centered research collaboration with teachers and school leaders from the participating courses of studies and faculties. Initially, the COVID-19 pandemic increased the interest of courses and faculties to offer online interventions that could help students improve their self-regulated learning and mental health. However, many teams became overburdened by the demands of having to instantly develop online education. This included the developing team responsible for programming the different modules of the chatbot. Three courses (out of the original 20 courses) that could no longer commit to any extraneous engagements had to pull out of the trial last-minute. Other courses did participate, but had trouble executing the designed plans. Frankly, so did I. During the experiment in chapter 2, I would visit the different courses regularly, to stay in touch and see if everything went all right. I would work on the different locations of the courses to prevent me from being an anonymous researcher. This time though, I was left with sending emails and videocalls. Combined with the time that was required for me to 'flip' the courses that I taught and two young kids at home, I fell short in managing what was needed for optimal fidelity. From the developers side, any scheduled modules and components of the chatbot were not ready at the right time, resulting in delays. Somehow, we managed to pull through. Currently we are finalizing the data analysis. Although we think the quality of the data in the end did not suffer, they came in just too late to be part of this thesis.

Based on what I had learned during the field experiment in chapter 2, I tried to organise structures that could organise further implementation if the intervention proved useful. Early on I made sure that the team of university wide student councillors and the team responsible for an innovative university wide 'student-app' were in on the project and willing to invest in the next phase after the research project. Both parties are currently discussing how the developed modules can be used in the RUAS 'student-app'. In conclusion, it seems that early engagement improved the potential for scaling up. The availability of a central program that was already working on a student app, can also be considered key.

Chapter 4: The Right Job Pays

Scientific Contribution

In the longitudinal study in chapter 4 we found a significant positive relation between paid congruent student-employment on study progress in the context of teacher education and non-negative effects of paid non-congruent work and unpaid congruent work. We also found evidence for a curvilinear relationship of the effects of paid non-congruent work during the first and third semester of college. Working one day per week related with the most study progress, while more than two days related to less study progress. The findings from our study provide support for the Role-Based Resource theory (Butler, 2007), but also suggest two new dimensions that should be taken into account. Differentiating between paid and unpaid congruent student employment allowed us to infer if remuneration matters. Paid congruent job hours significantly related with more study progress, while unpaid congruent job hours did not. Furthermore, we found that effects of different types of student employment can be time-specific. Some effects and interactions occur only during specific semesters of college. In our sample, the majority of students (70.5%) had a paid non-congruent job during the first semester of college, while none reported having a paid congruent job. We observed that a third of the students 'traded' their non-congruent job for a congruent job during their third year. This stresses the importance for future studies and reviews to distinguish between types of jobs, remuneration, intensity, and timing (semester), when studying the effects of student employment on educational outcomes. Our findings also stipulate the need for distinguishing between universities of applied sciences and research universities, given that the former often include multiple internships in their curricula. Finally, study designs should account for the differences between domains and faculties, because the availability of congruent paid work is likely to be dependent on the domain and the business cycle.

From a methodological perspective, chapter 4 offered a statistical approach that can readily be applied by other researchers who want to account for the differences in semesters and types of work. Our approach tried to limit endogeneity bias by controlling for gender and previous education, and by using a longitudinal design with 25 repeated measures which increased the within-student explanatory power. However, this does not principally exclude endogeneity bias, leaving the door open for a reverse causal relationship. Other studies used forms of matching, Cox proportional hazard models, quasi-experimental designs, or instrumental variables (e.g., local labour market conditions) to account for this problem (Neyt et al., 2019). Future studies could combine our statistical approach with a combination of matching and a suitable instrumental variable. Studying which students are offered a job and why would also be of great added value and could shed more light on whether congruent employment causes study progress or whether this should be attributed to a confounding variable. In the following section I will elaborate which practical advice we deduced from our findings and how this was received.

Practical Impact

Chapter 4 had two main practical implications. First, we falsified the assumption that accepting a paid job as a teacher hinders study progress. On the contrary, a paid job as a teacher during the study relates to faster study progress. These findings can be interpreted in different ways. It could be that job congruence offers benefits and allows students to quit their paid job outside of education, or it could be that students who perform better in terms of study progress are more often offered paid congruent jobs.

Second, we found a curvilinear relationship for paid work outside of education during specific semesters of college. With our method of analysis, we could deduce that respectively 7.75 and 8.25 hours per week is optimal and that more than 15.5 and 16.5 hours relates to less study progress during the first and third semesters.

The results from this study were presented to the management team of the teacher education faculty of the university where the research took place, and they were published in a popular magazine for teachers and school leaders. The managers and director were enthusiastic about the insights, but admitted having trouble interpreting them. Because of the risk of endogeneity bias, and the non-experimental design, it could not be concluded that getting a paid job led to more study progress, let alone that it could predict future behavioural effects. Yet, the maxim that it was bad *perse* had to be abandoned. We advised the managers that if they wanted students to work less, it could be a good idea to increase the remuneration for internships and see if they could ask schools to pay for overtime during an internship. Paid work in education seems to compete with paid work outside of education, while it does not seem to compete with study progress.

One month later, two investigative journalists published a compelling story that appeared in several large media outlets. The headlines were respectively: “Thrown before the lions unprepared”, “Schools rob interns from teacher education, which increases the teacher shortage”, “Schools lure young pre-service teacher away from teacher education too soon”, and “How starting teachers are burned up”. Based on interviews and a survey among the division for young members of the labour union for teachers, the articles claimed that pre-service teachers experienced longer study durations because of student employment in education. The articles mentioned that more than 600 teachers responded to the survey, but failed to mention the response rate of only 19.5%. To measure the effects on study progress, they asked teachers who reported a longer study duration, whether they thought that this was partly due to student employment, 53% of these respondents agreed. The investigative journalists study used a biased sample, had lower response rates, less reliable outcome variables, and did not align with the findings of our study, with recent systematic literature reviews and a meta-analysis, but probably did have a larger audience in The Netherlands than the scientific articles on this topic. This does not further the ambition of evidence-based education to establish sound evidence where existing evidence is lacking or uncertain to say the least.[12]

[12] The investigative journalists declined an invitation to talk about the differences between our findings because they preferred to answer questions by email. Unfortunately they did not respond to the emailed questions about their sample.

Evidence-based higher education

In a recent synthesis of higher education research, Tight (2021) describes the relationship between higher education and work as "a topic area that needs more research" (p. 235). Much indeed needs to be further investigated before it will constitute a solid evidence-base that can be used for clear policy advise. However, this experience also reminded me of the importance of the first principle of evidence-based education according to Davies (1999, p. 109):

"Educators at all levels need to be able to

- pose an answerable question about education;
- know where and how to find evidence systematically and comprehensively using the electronic (computer-based) and non-electronic (print) media;
- retrieve and read such evidence competently and undertake critical appraisal and analysis of that evidence according to agreed professional and scientific standards;
- organise and grade the power of this evidence;
- and determine its relevance to their educational needs and environments."

This aspiration, sometimes, seems at least as distant as the degree to which we have established sound evidence regarding this topic. Most colleague teacher educators I've spoken with know more about the articles from the investigative journalists than about any of the scientific research done in this field. The case might not be different for school leaders: Neeleman (2019) studied the decision making of Dutch secondary school leaders and found that they attached greater value to tacit knowledge and their intuition than to evidence when it came to choosing interventions. Whenever evidence was used, it often came through personal networks, knowledge brokers such as teachers who studied for a master's degree (Neeleman, 2019). I am grateful that being able to do a part-time PhD as a teacher allowed me to offer a small contribution to both aims of EBE.

Conclusion

The research in this dissertation suggests that higher education scholars should strive to study outcome measures such as study progress or retention in combination with other outcome measures that capture the relevant educational goods and values at stake. Transparently formulating outcome measures and a research agenda in cooperation with educational practitioners, and mounting experimental studies that investigate which interventions further these educational goods will provide sound evidence that is useful for practitioners and students.

In line with this endeavour, the studies in this dissertation made several contributions. Letting students in business and teacher education reflect on their future and goals in life significantly improved academic outcomes without negative effects on well-being. Offering them personalised follow-up coaching through a chatbot with accessible therapy when needed, could further prevent academic underperformance, anxiety, and depression. At the start of college, when most teacher education students have a side-job that is not related to their study domain, combining one day of work per week with studying relates to more study progress. During the third and fourth year, those who combine teacher education with a paid job as a teacher earn more course credits. Future research might further unravel the ways in which goal-setting, choosing the right (side)job and AI can contribute to academic thriving.

References

Abd-alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, *132*, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Abdul-Kader, S. A., & Woods, J. (2015). Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications, 6*(7), 72-80. https://doi.org/10.14569/IJACSA.2015.060712

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl, & J. Beckmann (Eds.), *Action control: From cognition to behavior* (pp. 11-39). Springer Verlag.

Akkerman, S. F., Bakker, A., & Penuel, W. R. (2021). Relevance of educational research: An ontological conceptualization. *Educational Researcher*. https://doi.org/10.3102/0013189X211028239

Andersson, G., Cuijpers, P., Carlbring, P., Riper, H., & Hedman, E. (2014). Guided internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: A systematic review and meta□analysis. *World Psychiatry, 13*(3), 288-295. https://doi.org/10.1002/wps.20151

Arnett, J. J. (2006). Emerging adulthood: Understanding the new way of coming of age. In J. J. Arnett, & J. L. Tanner (Eds.), *Emerging adults in America: Coming of age in the 21st century* (pp. 3-19). American Psychological Association. https://doi.org/10.1037/11381-001

Auerbach, R. P., Alonso, J., Axinn, W. G., Cuijpers, P., Ebert, D. D., Green, J. G., … Bruffaerts, R. (2016). Mental disorders among college students in the world health organization world mental health surveys. *Psychological Medicine, 46*(14), 2955-2970. https://doi.org/10.1017/S0033291716001665

Auerbach, R. P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., Demyttenaere, K., Ebert, D. D., Greif Green, J., Hasking, P., Murray, E., Nock, M. K., Pinder-Amaker, S., Sampson. N. A., Stein, D. J., Vilagut, G., Zaslavsky, A. M., & Kessler, R. C. (2018). WHO world mental health surveys international college student project: Prevalence and distribution of mental disorders. *Journal of Abnormal Psychology, 127*(7), 623-638. https://doi.org/10.1037/abn0000362

Ayyagari, R., Grover, V., & Purvis, R. (2011). Technostress: Technological antecedents and implications. *MIS Quarterly, 35*(4), 831-858. https://doi.org/10.2307/41409963

Bakker, A. B., Sanz Vergel, A. I., & Kuntze, J. (2014). Student engagement and performance: A weekly diary study on the role of openness. *Motivation and Emotion, 39*(1), 49-62. https://doi.org/10.1007/s11031-014-9422-5

Baldwin, C., Bensimon, E. M., Dowd, A. C., & Kleiman, L. (2011). Measuring Student Success. *New Directions for Community Colleges*, *153*, 75-88. https://doi.org/10.1002/cc.438

Bassett, C. (2019). The computational therapeutic: Exploring Weizenbaum's ELIZA as a history of the present. *AI & Society, 34*, 803-812. https://doi.org/10.1007/s00146-018-0825-9

Baumeister, R. F., Vohs, K. D., Aaker, J. L., & Garbinsky, E. N. (2013). Some key differences between a happy life and a meaningful life. *The Journal of Positive Psychology*, *8*(6), 505-516. https://doi.org/10.1080/17439760.2013.830764

Baumel, A., Muench, F., Edan, S., & Kane, J. M. (2019). Objective user engagement with mental health apps: Systematic search and panel-based usage analysis. *Journal of Medical Internet Research, 21*(9), e14567. https://doi.org/10.2196/14567

Bayram, N., & Bilgel, N. (2008). The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. *Social Psychiatry and Psychiatric Epidemiology*, *43*(8), 667-672. https://doi.org/10.1007/s00127-008-0345-x

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485–540. https://doi.org/10.3102/00346543055004485

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. National Bureau of Economic Research.

Becker, G. S. (1965). A theory of the allocation of time. *The Economic Journal*, *75*(299), 493-517. https://doi.org/10.2307/2228949

Bendig, E., Erb, B., Schulze-Thuesing, L., & Baumeister, H. (2019). The next generation: chatbots in clinical psychology and psychotherapy to foster mental health–a scoping review. *Verhaltenstherapie*, 1-13. http://doi.org/10.1159/000501812

Bettinger, E. P., & Baker, R. B. (2014). The effects of student coaching: An evaluation of a randomized experiment in student advising. *Educational Evaluation and Policy Analysis*, *36*(1), 3-19. https://doi.org/10.3102/0162373713500523

Bickmore, T., Gruber, A., & Picard, R. (2005). Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient Education and Counseling, 59*(1), 21-30. https://doi.org/10.1016/j.pec.2004.09.008

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347-364. https://doi.org/10.1007/BF00138871

Biesta, G. J. (2007). Why "what works" won't work: Evidence□based practice and the democratic deficit in educational research. *Educational Theory*, *57*(1), 1-22. https://doi.org/10.1111/j.1741-5446.2006.00241.x

Biesta, G. J. (2010). Why 'what works' still won't work: From evidence-based education to value-based education. *Studies in Philosophy and Education*, *29*(5), 491-503. https://doi.org/10.1007/s11217-010-9191-x

Bipp, T., Kleingeld, A., Van Den Tooren, M., & Schinkel, S. (2015). The effect of self-set grade goals and core self-evaluations on academic performance: A diary study. *Psychological Reports*, *117*(3), 917-930. https://doi.org/10.2466/11.07.PR0.117c26z0

Bormans, M J. G., & Dekker, I. (2016). *Samen leven in de moderne samenleving.* Rotterdam University Press.

Bowman, N. A. H., Patrick, L., Denson, Nida, & Bronkema, R. (2015). Keep on truckin' or stay the course? Exploring grit dimensions as differential predictors of educational achievement, satisfaction, and intentions. *Social Psychological and Personality Science*, *6*(6), 639–645. https://doi.org/10.1177/1948550615574300

Braun, D., & Matthes, F. (2019). Towards a framework for classifying chatbots. In J. Filipe, M. Smialek, A. Brodsky & S. Hammoudi (Eds.), *Proceedings of the 21th international conference on enterprise information systems (ICEIS 2019) - volume 2* (pp. 496-501). Heraklion, Greece. https://doi.org/10.5220/0007772704960501

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2016). Educational goods and values: A framework for decision makers. *Theory and Research in Education*, *14*(1), 3-25. https://doi.org/10.1177/1477878515620887

Brighouse, H. & McPherson, M. (2015). Introduction: Problems of morality and justice in higher education. In H. Brighouse & M. McPherson (Eds.), *The aims of higher education: Problems of morality and justice* (pp. 1-6). The University of Chicago Press. https://doi.org/107208/chicago/9780226259512.001.0001

Bruffaerts, R., Mortier, P., Kiekens, G., Auerbach, R. P., Cuijpers, P., Demyttenaere, K., ... Kessler, R. C. (2018). Mental health problems in college freshmen: Prevalence and academic functioning. *Journal of Affective Disorders, 225*, 97-103. https://doi.org/10.1016/j.jad.2017.07.044

Bruggeman, J. (2012). *(Hoge)school maken in Rotterdam: 25 jaar Hogeschool Rotterdam*. Stad en Bedrijf.

Buscha, F., Maurel, A., Page, L., & Speckesser, S. (2012). The effect of employment while in high school on educational attainment: A conditional difference‐in‐differences approach. *Oxford Bulletin of Economics and Statistics*, *74*(3), 380–396. https://doi.org/10.1111/j.1468-0084.2011.00650.x

Bulle, N. (2018). What is wrong with Dewey's theory of knowing. *Ergo, an Open Access Journal of Philosophy*, *5*(21). https://doi.org/10.3998/ergo.12405314.0005.021

Burris, J. L., Brechting, E. H., Salsman, J., & Carlson, C. R. (2009). Factors associated with the psychological well-being and distress of university students. *Journal of American College Health, 57*(5), 536-544. https://doi.org/10.3200/ JACH.57.5.536-544

Butler, A. B. (2007). Job characteristics and college performance and attitudes: A model of work-school conflict and facilitation. *Journal of Applied Psychology, 92*(2), 500. https://doi.org/10.1037/0021-9010.92.2.500

Campbell-Sills, L., & Stein, M. B. (2007). Psychometric analysis and refinement of the Connor–Davidson resilience scale (CD-RISC): Validation of a 10-item measure of resilience. *Journal of Traumatic Stress*, *20*, 1019–1028. https://doi.org/10.1002/jts.20271

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: An updated systematic review and meta-analysis. *Cognitive Behaviour Therapy, 47*(1), 1-18. https://doi.org/10.1080/16506073.2017.1401115

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*(1), 40. https://doi.org/10.1186/1748-5908-2-40

Carroll, J. M. (1997). Human-computer interaction: Psychology as a science of design. *Annual Review of Psychology*, *48*, 61-83. https://doi.org/10.1146/annurev.psych.48.1.61

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Centre for Education Statistics and Evaluation. (2015). *Student wellbeing.* Sydney, Australia: NSW Department of Education and Communities. Retrieved at April 2, 2020 from https://www.cese.nsw.gov.au//images/stories/PDF/student_wellbeing_LR_AA.pdf

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293-332. https://doi.org/10.1207/s1532690xci0804_2

Chang, E., London, R. A., & Foster, S. S. (2019). Reimagining student success: Equity-oriented responses to traditional notions of success. *Innovative Higher Education*, *44*(6), 481-496. https://doi.org/10.1007/s10755-019-09473-x

Chau, P. Y. K., & Hu, P. J. (2002). Examining a model of information technology acceptance by individual professionals: An exploratory study. *Journal of Management Information Systems, 18*(4), 191-229. https://doi.org/10.1080/07421222.2002.11045699

Cheung, R., O'Donnell, S., Madi, N., & Goldner, E. M. (2017). Factors associated with delayed diagnosis of mood and/or anxiety disorders. *Health Promotion and Chronic Disease Prevention in Canada, 37*(5), 137-148. https://doi.org/10.24095/hpcdp.37.5.02

Choi, A. (2018). *Emotional well-being of children and adolescents: Recent trends and relevant factors.* (No. 169). OECD Publishing. https://doi.org/10.1787/41576fb2-en

Clark, D., Gill, D., Prowse, V., & Rush, M. (2020). Using goals to motivate college students: Theory and evidence from field experiments, *Review of Economics and Statistics*, *102*(4), 648-663. https://doi.org/10.1162/rest_a_00864

Clement, S., Schauman, O., Graham, T., Maggioni, F., Evans-Lacko, S., Bezborodovs, N., Morgan, C., Rüsch, N., Brown, J. S. L., & Thornicroft, G. (2015). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine*, *45*(1), 11–27. https://doi.org/10.1017/S0033291714000129

Compton, W. C., Smith, M. L., Cornish, K. A., & Qualls, D. L. (1996). Factor structure of mental health measures. *Journal of Personality and Social Psychology*, *71*(2), 406. https://doi.org/10.1037/0022-3514.71.2.406

Connor, K. M., & Davidson, J. R. (2003). Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). *Depression and Anxiety, 18*(2), 76-82. https://doi.org/10.1002/da.10113

Cook, T. D. (2007). Randomized experiments in education: Assessing the objections to doing them. *Economics of Innovation and New Technology*, *16*(5), 331-355. https://doi.org/10.1080/10438590600982335

Corpus, J. H., Robinson, K. A., & Wormington, S. V. (2020). Trajectories of motivation and their academic correlates over the first year of college. *Contemporary Educational Psychology*, *63*, Article 101907. https://doi.org/10.1016/j.cedpsych.2020.101907

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*(3), 175-199. https://doi.org/10.3102/01623737024003175

Cowen, N. (2019). For whom does "what works" work? The political economy of evidence-based education. *Educational Research and Evaluation*, *25*(1-2), 81-98. https://doi.org/10.1080/13803611.2019.1617991

Curran, T., Hill, A. P., Appleton, P. R., Vallerand, R. J., and Standage, M. (2015). The psychology of passion: a meta-analytical review of a decade of research on intrapersonal outcomes. *Motivation and Emotion, 39*, 631–655. https://doi.org/10.1007/s11031-015-9503-0

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23-45. https://doi.org/10.1016/S0272-7358(97)00043-3

Dancy, M., Henderson, C., & Turpen, C. (2016). How faculty learn about and implement research-based instructional strategies: The case of peer instruction. *Physical Review Physics Education Research*, *12*(1), 010110. https://doi.org/10.1103/PhysRevPhysEducRes.12.010110

Davies, E. B., Morriss, R., & Glazebrook, C. (2014). Computer-delivered and web-based interventions to improve depression, anxiety, and psychological well-being of university students: A systematic review and meta-analysis. *Journal of Medical Internet Research, 16*(5):e130. https://doi.org/10.2196/jmir.3142

Davies, P. (1999). What is evidence□based education?. *British Journal of Educational Studies*, *47*(2), 108-121. https://doi.org/10.1111/1467-8527.00106

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340. https://doi.org/10.2307/249008

De Bruijn-Smolders, M., (2017). *Self-regulated learning and academic performance; a study among freshmen*. [Doctoral dissertation, Erasmus University Rotterdam]. hdl.handle.net/1765/

De Bruijn-Smolders, M., Timmers, C. F., Gawke, J. C. L., Schoonman, W., & Born, M. P. (2016). Effective self-regulatory processes in higher education: Research findings and future directions. A systematic review. *Studies in Higher Education, 41*(1), 139-158. https://doi.org/10.1080/03075079.2014.915302

De Girolamo, G., Dagani, J., Purcell, R., Cocchi, A., & McGorry, P. D. (2012). Age of onset of mental disorders and use of mental health services: Needs, opportunities and obstacles. *Epidemiology and Psychiatric Sciences*, *21*, 47-57. https://doi.org/10.1017/S2045796011000746

De Luca, S. M., Franklin, C., Yueqi, Y., Johnson, S., & Brownson, C. (2016). The relationship between suicide ideation, behavioral health, and college academic performance. *Community Mental Health Journal*, *52*(5), 534-540. https://doi.org/10.1007/s10597-016-9987-4

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21. https://doi.org/10.1016/j.socscimed.2017.12.005

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*(6), 627-668. https://doi.org/10.1037/0033-2909.125.6.627

Deiglmeier, K., & Greco, A. (2018). Why proven solutions struggle to scale up. *Stanford Social Innovation Review.* Available at. https://ssir. org/articles/entry/why_proven_solutions_struggle_to_scale_up.

Dekker I., De Jong E. M., Schippers M. C., De Bruijn-Smolders, M., Alexiou A. & Giesbers, B. (2020). Optimizing students' mental health and academic performance: AI-enhanced life crafting. *Frontiers in Psychology*, *11*, Article 1063. https://doi.org/10.3389/fpsyg.2020.01063

Derksen, F., Bensing, J., & Lagro-Janssen, A. (2013). Effectiveness of empathy in general practice: A systematic review. *British Journal of General Practice*, *63*(606):e76-e84. https://doi.org/10.3399/bjgp13X660814

Diefenbach, S., & Niess, J. (2015). Vom Wunsch zum Ziel?! Potential von Technologien zur Selbstverbesserung. *Mensch Und Computer 2015–Proceedings*, 391-394. http://doi.org/10.1515/9783110443929-060

Dobronyi, C. R., Oreopoulos, P., & Petronijevic, U. (2019). Goal setting, academic reminders, and college success: A large-scale field experiment. *Journal of Research on Educational Effectiveness, 12*(1), 38-66. https://doi.org/10.1080/19345747.2018.1517849

Donitsa-Schmidt, S., & Zuzovsky, R. (2016). Quantitative and qualitative teacher shortage and the turnover phenomenon. *International Journal of Educational Research*, 77, 83-91. https://doi.org/10.1016/j.ijer.2016.03.005

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087-1101. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*(2), 166-174. https://doi.org/10.1080/00223890802634290

Duflo, E., & Banerjee, A. (Eds.). (2017). *Handbook of field experiments*. Elsevier.

Duncan, T., & McKeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educational Psychologist*, *40*(2), 117–128. https://doi.org/10.1207/s15326985ep4002_6

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58. https://doi.org/10.1177/1529100612453266

Durlak, J. A. (2015). Studying program implementation is not easy but it is essential. *Prevention Science*, *16*(8), 1123-1127. https://doi.org/10.1007/s11121-015-0606-3

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3-4), 327-350. https://doi.org/10.1007/s10464-008-9165-0

Ebert, D. D., Van Daele, T., Nordgreen, T., Karekla, M., Compare, A., Zarbo, C., ... Baumeister, H. (2018). Internet-and mobile-based psychological interventions: Applications, efficacy, and potential for improving mental health: A report of the EFPA E-health taskforce. *European Psychologist, 23*(2), 167-187. https://doi.org/10.1027/1016-9040/a000318

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., and Jardeleza, S. E. (2011). What we say is not what we do: effective evaluation of faculty professional development programs. *BioScience*, *61*(7), 550–558. https://doi.org/10.1525/bio.2011.61.7.9

Epton, T., Currie, S., & Armitage, C. J. (2017). Unique effects of setting goals on behavior change: Systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *85*(12), 1182-1198. https://doi.org/10.1037/ccp0000260

European Commission (2014). Study on policy measures to improve the attractiveness of the teaching profession in Europe. Volume 2. https://doi.org/10.2766/41166

Evans, N. J., Forney, D. S., Guido, F. M., Patton, L. D., & Renn, K. A. (2009). *Student development in college: Theory, research, and practice*. John Wiley & Sons.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. https://doi.org/10.3758/BF03193146

Fingerson, L., & Troutman, D. R. (2019). Measuring and Enhancing Student Success. *New Directions for Institutional Research*, *2019*(184), 33-46. https://doi.org/10.1002/ir.20320

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2):e19. https://doi.org/10.2196/mental.7785

Frankowski, A., Van der Steen, M., Bressers, D., Schulz, M., Shewbridge, C., Fuster, M., & Rouw, R. (2018). "Dilemmas of central governance and distributed autonomy in education", *OECD Education Working Papers*, No. 189. https://doi.org/10.1787/060260bf-en

Freund, A. M., & Baltes, P. B. (2002). Life-management strategies of selection, optimization and compensation: Measurement by self-report and construct validity. *Journal of Personality and Social Psychology*, *82*(4), 642-662. https://doi.org/10.1037/0022-3514.82.4.642

Froyd, J. E., Borrego, M., Cutler, S., Henderson, C., & Prince, M. J. (2013). Estimates of use of research-based instructional strategies in core electrical or computer engineering courses. *IEEE Transactions on Education*, *56*(4), 393-399. https://doi.org/10.1109/TE.2013.2244602.

Fulmer, R. (2019). Artificial intelligence and counseling: Four levels of implementation. *Theory & Psychology, 29*(6), 807-819. https://doi.org/10.1177/0959354319853045

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health, 5*(4):e64. https://doi.org/10.2196/mental.9782

Garcia, E., & Weiss, E. (2019). The teacher shortage is real, large and growing, and worse than we thought. The first report in "The perfect storm in the teacher labor market" Series. *Economic Policy Institute*. https://epi.org/163651

Gateshill, G., Kucharska-Pietura, K., & Wattis, J. (2011). Attitudes towards mental disorders and emotional empathy in mental health and other healthcare professionals. *The Psychiatrist*, *35*, 101-105. https://doi.org/10.1192/pb.bp.110.029900

Geel, R., & Backes-Gellner, U. (2012). Earning while learning: When and how student employment is beneficial. *Labour, 26*(3), 313–340. https://doi.org/10.1111/j.1467-9914.2012.00548.x

Gettinger, M., & Seibert, J. K. (2002). Contributions of study skills to academic competence. *School Psychology Review*, *31*(3), 350-365. https://doi.org/10.1080/02796015.2002.12086160

Goedegebuure, L. C. J., & Meek, V. (1991). Restructuring higher education. A comparative analysis between Australia and The Netherlands. *Comparative Education, 27*(1), 7-22. https://doi.org/10.1080/0305006910270103

Gollwitzer, P. M. (1993). Goal achievement: The role of intentions. *European Review of Social Psychology*, *4*(1), 141-185. https://doi.org/10.1080/14792779343000059

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *American Psychologist, 54*(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493

Gollwitzer, P. M., & Brandstätter, V. (1997). Implementation intentions and effective goal pursuit. *Journal of Personality and Social Psychology, 73*(1), 186-199. https://doi.org/10.1037/0022-3514.73.1.186

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology*, *38*, 69-119. https://doi.org/10.1016/S0065-2601(06)38002-1

Graybiel, A. M., and Smith, K. S. (2014). Good habits, bad habits. *Scientific American*, *310*, 38–43. https://doi.org/10.1038/scientificamerican0614-38

Greenberg M. T. & R. Abenavoli (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts, *Journal of Research on Educational Effectiveness*, *10*(1), 40-67. https://doi.org/10.1080/19345747.2016.1246632

Greenhaus, J. H., & Powell, G. N. (2006). When work and family are allies: A theory of work-family enrichment. *Academy of Management Review, 31*(1), 72-92. https://doi.org/10.5465/amr.2006.19379625

Grossi, E., Groth, N., Mosconi, P., Cerutti, R., Pace, F., Compare, A., & Apolone, G. (2006). Development and validation of the short version of the Psychological General Well-Being Index (PGWB-S). *Health and Quality of Life Outcomes*, *4*, Article 88. https://doi:10.1186/1477-7525-4-88

Hargreaves, D. H. (2000). Teaching as a research-based profession: possibilities and prospects. In B. Moon, J. Butcher, & E. Bird (Eds.), *Leading professional development in education*. (pp. 200-210). Psychology Press.

Harrer, M., Adam, S. H., Baumeister, H., Cuijpers, P., Karyotaki, E., Auerbach, R. P., ... Ebert, D. D. (2019). Internet interventions for mental health in university students: A systematic review and meta□analysis. *International Journal of Methods in Psychiatric Research, 28*(2):e1759. https://doi.org/10.1002/mpr.1759

Hartley, M. T. (2010). Increasing resilience: Strategies for reducing dropout rates for college students with psychiatric disabilities. *American Journal of Psychiatric Rehabilitation, 13*(4), 295-315. https://doi.org/10.1080/15487768.2010.523372

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching*, *48*(8), 952-984. https://doi.org/10.1002/tea.20439

Henderson, C., and Dancy, M. H. (2009). Impact of physics education research on the teaching of introductory quantitative physics in the United States. *Physical Review Physics Education Research*. *5*, Article 020107. https://doi.org/10.1103/PhysRevSTPER.5.020107

Henderson, H. (2007). *Artificial intelligence: Mirrors for the mind*. Infobase Publishing.

Hodge, B., Wright, B., & Bennett, P. (2018). The role of grit in determining engagement and academic outcomes for university students. *Research in Higher Education*, *59*(4), 448-460. https://doi.org/10.1007/s11162-017-9474-y

Holland, R. W., Aarts, H., & Langendam, D. (2006). Breaking and creating habits on the working floor: a field-experiment on the power of implementation intentions. *Journal of Experimental Social Psychology*, *42*, 776–783. https://doi.org//j.jesp.2005.11.006

Honicke, T., & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review*, *17*, 63-84. https://doi.org/10.1016/j.edurev.2015.11.002

Horowitz, E., Sorensen, N., Yoder, N., & Oyserman, D. (2018). Teachers can do it: Scalable identity-based motivation intervention in the classroom. *Contemporary Educational Psychology*, *54*, 12-28. https://doi.org/10.1016/j.cedpsych.2018.04.004

Hoyle, R. H., & Sherrill, M. R. (2006). Future orientation in the self□system: Possible selves, self□regulation, and behavior. *Journal of Personality*, *74*(6), 1673-1696. https://doi.org/10.1111/j.1467-6494.2006.00424.x

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Humphrey, R. (2006). Pulling structured inequality into higher education: the impact of part-time working on English university students. *Higher Education Quarterly, 60*(3), 270–286. https://doi.org/10.1111/j.1468-2273.2006.00317.x

Hunt, J., & Eisenberg, D. (2010). Mental health problems and help-seeking behavior among college students. *Journal of Adolescent Health, 46*(1), 3-10. https://doi.org/10.1016/j.jadohealth.2009.08.008

Huppert, F. A., Baylis, N., Keverne, B., Ryff, C. D., Singer, B. H., & Love, G. D. (2004). Positive health: connecting well–being with biology. *Philosophical Transactions of the Royal Society B: Biological Science*, *359*, 1383–1394. https://doi.org/10.1098/rstb.2004.1521

Ibrahim, A. K., Kelly, S. J., Adams, C. E., & Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. *Journal of Psychiatric Research, 47*(3), 391-400. https://doi.org/10.1016/j.jpsychires.2012.11.015

Johnson, M. L., Taasoobshirazi, G., Kestler, J. L., & Cordova, J. R. (2015). Models and messengers of resilience: a theoretical model of college students' resilience, regulatory strategy use, and academic achievement. *Educational Psychology*, *35*(7), 869-885. https://doi.org/10.1080/01443410.2014.893560

Jacques, R., Følstad, A., Gerber, E., Grudin, J., Luger, E., Monroy-Hernández, A., & Wang, D. (2019). Conversational agents: Acting on the wave of research and development. Paper presented at the *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-8. https://doi.org/10.1145/3290607.3299034

Jongbloed, B., Kaiser, F. & Westerheijden, D. F. (2020). Improving study success and diversity in Dutch higher education using performance agreements. *Tertiary Education Management*, *26*, 329–343. https://doi.org/10.1007/s11233-019-09055-8

Joyce, K. E., & Cartwright, N. (2020). Bridging the gap between research and practice: Predicting what will work locally. *American Educational Research Journal*, *57*(3). 1045-1082. https://doi.org/10.3102/0002831219866687

Kahneman, D. (1999). Objective Happiness. In D. Kahneman, E. Diener, & N. Schwarz (Eds.). *Well-being: Foundations of hedonic psychology*. Russell Sage Foundation.

Kamita, T., Ito, T., Matsumoto, A., Munakata, T., & Inoue, T. (2019). A chatbot system for mental healthcare based on SAT counseling method. *Mobile Information Systems*, Article 9517321. https://doi.org/10.1155/2019/9517321

Kessler, R. C., Foster, C. L., Saunders, W. B., & Stang, P. E. (1995). Social consequences of psychiatric disorders, I: Educational attainment. *American Journal of Psychiatry*, *152*(7), 1026-1032. https://doi.org/10.1176/ajp.152.7.1026

Kim, K. R., & Seo, E. H. (2015). The relationship between procrastination and academic performance: A meta-analysis. *Personality and Individual Differences*, *82*, 26-33. https://doi.org/10.1016/j.paid.2015.02.038

Kleingeld, A., Van Mierlo, H., & Arends, L. (2011). The effect of goal setting on group performance: A meta-analysis. *Journal of Applied Psychology, 96*(6), 1289–1304. https://doi.org/10.1037/a0024315

Klinger, E. (1977). *Meaning and void: Inner experience and the incentives in peoples lives*. University of Minnesota Press.

Klug, H. J., & Maier, G. W. (2015). Linking goal progress and subjective well-being: A meta-analysis. *Journal of Happiness Studies*, *16*(1), 37-65. https://doi.org/10.1007/s10902-013-9493-0

Koestner, R., Lekes, N., Powers, T. A., & Chicoine, E. (2002). Attaining personal goals: Self-concordance plus implementation intentions equals success. *Journal of Personality and Social Psychology, 83,* 231–244.

Koutsouris, G., & Norwich, B. (2018). What exactly do RCT findings tell us in education research?. *British Educational Research Journal*, *44*(6), 939-959. https://doi.org/10.1002/berj.3464

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253. https://doi.org/10.3102/0013189X20912798

Kristjánsson, K. (2017). Recent work on flourishing as the aim of education: A critical review. *British Journal of Educational Studies*, *65*(1), 87-107. https://doi.org/10.1080/00071005.2016.1182115

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2007). *Piecing together the student success puzzle*: *Research, propositions, and recommendations* (ASHE Higher Education Report, Vol. 32). Jossey-Bass.

Kuh, G. D., Kinzie, J., Schuh, J. H., & Whitt, E. J. (2005). *Student success in college: Creating conditions that matter*. Jossey-Bass.

Kvillemo, P., Brandberg, Y., & Bränström, R. (2016). Feasibility and outcomes of an internet-based mindfulness training program: A pilot randomized controlled trial. *JMIR mental health*, *3*(3):e33. https://doi.org/10.2196/mental.5457

Laloux, F. (2015). *Reinventing organizations.* Lannoo Meulenhoff.

Lambert, M. (2018). *Chatbot decision trees*. Retrieved at April 2, 2020 from https://chatbotslife.com/chatbot-decision-trees-a42ed8b8cf32

Lane, J., Lane, A. M., & Kyprianou, A. (2004). Self-efficacy, self-esteem and their impact on academic performance. *Social Behavior and Personality*, *32*(3), 247-256. https://doi.org/10.2224/sbp.2004.32.3.247

Latham, G. P., & Brown, T. C. (2006). The effect of learning vs. outcome goals on self□efficacy, satisfaction and performance in an MBA program. *Applied Psychology: An International Review*, *55*(4), 606-623. https://doi.org/10.1111/j.1464-0597.2006.00246.x

Latham, G. P., & Seijts, G. H. (1999). The effects of proximal and distal goals on performance on a moderately complex task. *Journal of Organizational Behavior*, *20*, 421–429. https://www.jstor.org/stable/3100381

Lattie, E. G., Adkins, E. C., Winquist, N., Stiles-Shields, C., Wafford, Q. E., & Graham, A. K. (2019). Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: Systematic review. *Journal of Medical Internet Research*, *21*(7):e12869. https://doi.org/10.2196/12869

Locke, E. A. (2015). Theory building, replication, and behavioral priming: Where do we need to go from here?. *Perspectives on Psychological Science, 10*(3), 408-414. https://doi.org/10.1177/1745691614567231

Locke, E. A., & Latham, G. (2002). Building a practically useful theory of goal-setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705-717. https://doi.org/10.1037/0003-066X.57.9.705

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned?. *Educational Researcher*, *48*(3), 158-166. https://doi.org/10.3102/0013189X19832850

Lucas, G. M., Gratch, J., King, A., & Morency, L. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94-100. https://doi.org/10.1016/j.chb.2014.04.043

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304-316. https://doi.org/10.3102/0013189X14545513

Marks, S. R. (1977). Multiple roles and role strain: Some notes on human energy, time and commitment. *American Sociological Review*, *42*(6) 921-936. https://doi.org/10.2307/2094577

Martin, A. J., Bottrell, D., Armstrong, D., Mansour, M., Ungar, M., Liebenberg, L., & Collie, R.J. (2015). The role of resilience in assisting the educational connectedness of at-risk youth: A study of service users and non-users. *International Journal of Educational Research*, *74*, 1–12. https://doi.org/10.1016/j.ijer.2015.09.004

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?. *American Psychologist*, *70*(6), 487-498. https://doi.org/10.1037/a0039400

McNeal, R. B. (1997). Are students being pulled out of high school? The effect of adolescent employment on dropping out. *Sociology of Education*, 206-220. https://doi.org/10.2307/2673209

Mento, A. J., Steel, R. P., & Karren, R. J. (1987). A meta-analytic study of the effects of goal setting on task performance: 1966–1984. *Organizational Behavior and Human Decision Processes*, *39*(1), 52-83. https://doi.org/10.1016/0749-5978(87)90045-8

Meurk, C., Leung, J., Hall, W., Head, B. W., & Whiteford, H. (2016). Establishing and governing e-mental health care in Australia: A systematic review of challenges and a call for policy-focussed research. *Journal of Medical Internet Research*, *18*(1):e10. https://doi.org/10.2196/jmir.4827

Moir, T. (2018). Why is implementation science important for intervention design and evaluation within educational settings?. *Frontiers in Education*, *3*(61). https://doi.org/10.3389/feduc.2018.00061

Moon, J., & Kim, Y. (2001). Extending the TAM for a world-wide-web context. *Information & Management*, *38*(4), 217-230. https://doi.org/10.1016/S0378-7206(00)00061-6

Morisano, D., Hirsh, J. B., Peterson, J. B., Pihl, R. O., Shore, B. M. (2010). Setting, elaborating, and reflecting on personal goals improves academic performance. *Journal of Applied Psychology*, *95*(2), 255-264. https://doi.org/10.1037/a0018478

Morosanu, L., Handley, K., & O’Donovan, B. (2010). Seeking support: researching first□year students’ experiences of coping with academic life. *Higher Education Research & Development*, *29*(6), 665-678. https://doi.org/10.1080/07294360.2010.487200

Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research*, *20*(6):e10148. https://doi.org/10.2196/10148

Morrison, K. (2021). *Taming randomized controlled trials in education: Exploring key claims, issues and debates*. Routledge.

Mortier, P., Demyttenaere, K., Auerbach, R. P., Green, J. G., Kessler, R. C., Kiekens, G., … Bruffaerts, R. (2015). The impact of lifetime suicidality on academic performance in college freshmen. *Journal of Affective Disorders*, *186*, 254-260. https://doi.org/10.1016/j.jad.2015.07.030

Muthén, L. K., & Muthén, B. O. (1998-2006). *Mplus user's guide* (4th ed.). Muthén & Muthén.

Neeleman, M. B. M. (2019). School autonomy in practice: School intervention decision-making by Dutch secondary school leaders [Doctoral dissertation, Maastricht University]. https://doi.org/10.26481/dis.20190628mn

Newman, J. (2017). Deconstructing the debate over evidence-based policy. *Critical Policy Studies*, *11*(2), 211-226. https://doi.org/10.1080/19460171.2016.1224724

Newton, P. M., Da Silva, A., & Berry, S. (2020). The case for pragmatic evidence-based higher education: A useful way forward?. *Frontiers in Education*, *5*:583157. https://doi.org/10.3389/feduc.2020.583157

Neyt, B., Omey, E., Verhaest, D., & Baert, S. (2019). Does student work really affect educational outcomes? A review of the literature. *Journal of Economic Surveys*, *33*(3), 896-921. https://doi.org/10.1111/joes.12301

Oettingen, G. (2000). Expectancy effects on behavior depend on self-regulatory thought. *Social Cognition*, *18*(2), 101-129. https://doi.org/10.1521/soco.2000.18.2.101

Oettingen, G. (2012). Future thought and behaviour change. *European Review of Social Psychology*, *23*(1), 1-63. https://doi.org/10.1080/10463283.2011.643698

Oettingen, G., Mayer, D., and Brinkmann, B. (2010). Mental contrasting of future and reality. *Journal of Personnel Psychology*, *9*(3), 138–144. https://doi.org/10.1027/1866-5888/a000018

Oettingen, G., & Sevincer, A. T. (2018). Fantasy about the future as friend and foe. In G. Oettingen, A. T. Sevincer & P. M. Gollwitzer (Eds.), *The psychology of thinking about the future* (pp. 127-149). The Guilford Press.

Organization for Economic Co-operation and Development (OECD). (2005). *Attracting, developing and retaining effective teachers - final report: Teachers matter.* http://www.oecd.org/education/school/attractingdevelopingandretainingeffectiveteachers-finalreportteachersmatter.htm

Organisation for Economic Co-operation and Development (OECD). (2010). *Education at a glance 2010: OECD indicators*. OECD.

Organisation for Economic Co-operation and Development (OECD). (2013). *Education at a glance 2013: OECD indicators.* OECD.

Organization for Economic Co-operation and Development (OECD). (2019). *Education at a glance 2019: OECD indicators*. OECD.

Oyserman, D., Bybee, D., & Terry, K. (2006). Possible selves and academic outcomes: How and when possible selves impel action. *Journal of Personality and Social Psychology*, *91*(1), 188-204. https://doi.org/10.1037/0022-3514.91.1.188

Oyserman, D., & Destin, M. (2010). Identity-based motivation: Implications for intervention. *The Counseling Psychologist*, *38*(7), 1001-1043. https://doi.org/10.1177/0011000010374775

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, Article 422. https://doi.org/10.3389/fpsyg.2017.00422

Parasuraman, A. (2000). Technology readiness index (TRI) a multiple-item scale to measure readiness to embrace new technologies. *Journal of Service Research*, *2*(4), 307-320. https://doi.org/10.1177/109467050024001

Park, M., Aiken, M., & Salvador, L. (2018). How do humans interact with chatbots?: An analysis of transcripts. *International Journal of Management and Information Technology*, *14*, 3338-3350. https://doi.org/10.24297/ijmit.v14i0.7921

Pennebaker, J. W. (2004). Theories, therapies, and taxpayers: On the complexities of the expressive writing paradigm. *Clinical Psychology: Science and Practice*, *11*(2), 138-142. https://doi.org/10.1093/clipsy.bph063

Pennebaker, J. W., & Chung, C. K. (2011). *Expressive writing: Connections to physical and mental health.* In H. S. Friedman (Ed.), *Oxford library of psychology. The Oxford handbook of health psychology* (p. 417–437). Oxford University Press.

Pennebaker, J. W., Colder, M., & Sharp, L. K. (1990). Accelerating the coping process. *Journal of Personality and Social Psychology*, *58*(3), 528-537. https://doi.org/10.1037//0022-3514.58.3.528

Perry, R. P. (1991). Perceived control in college students: Implications for instruction in higher education. In J. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 7, pp. 1–56). Agathon Press.

Peterson, J. B., & Mar, R. A. (2004). *Self-authoring program: The ideal future.* Unpublished manuscript, University of Toronto, Ontario, Canada. Retrieved from www.selfauthoring.com

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33-40. https://doi.org/10.1037/0022-0663.82.1.33

Plant, E. A., Ericsson, K. A., Hill, L., & Asberg, K. (2005). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology, 30*(1), 96-116. https://doi.org/10.1016/j.cedpsych.2004.06.001

Powers, T. A., Koestner, R., & Topciu, R. A. (2005). Implementation intentions, perfectionism, and goal progress: Perhaps the road to hell is paved with good intentions. *Personality and Social Psychology Bulletin*, *31*(7), 902-912. https://doi.org/10.1177/0146167204272311

Provoost, S., Lau, H. M., Ruwaard, J., & Riper, H. (2017). Embodied conversational agents in clinical psychology: A scoping review. *Journal of Medical Internet Research*, *19*(5) https://doi.org/10.2196/jmir.6553

Radziwill, N. M., & Benton, M. C. (2017). *Evaluating Quality of Chatbots and Intelligent Conversational Agents.* ArXiv. https://arxiv.org/abs/1704.04579

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2020). *A User's Guide to MLwiN Version 3.05*. University of Bristol.

Reis, S. M., & McCoach, D. B. (2000). The underachievement of gifted students: What do we know and where do we go? *Gifted Child Quarterly*, *44*(3), 152–170. https://doi.org/10.1177/001698620004400302

Reiter, R., & Klenk, T. (2019). The manifold meanings of 'post-New Public Management' - a systematic literature review. *International Review of Administrative Sciences*, *85*(1), 11-27. https://doi.org/10.1177/0020852318759736

Respondek, L., Seufert, T., Hamm, J. M., & Nett, U. E. (2020). Linking changes in perceived academic control to university dropout and university grades: A longitudinal approach. *Journal of Educational Psychology*, *112*(5), 987-1002. http://dx.doi.org/10.1037/edu0000388

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353-387. https://doi.org/10.1037/a0026838

Riggert, S. C., Boyle, M., Petrosko, J. M., Ash, D., & Rude-Parkins, C. (2006). Student employment and higher education: Empiricism and contradiction. *Review of Educational Research, 76*(1), 63-92. https://doi.org/10.3102/00346543076001063

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*(2), 261-288. https://doi.org/10.1037/0033-2909.130.2.261

Royal College of Psychiatrists. (2011). *Mental health of students in higher education college*. (College report CR166). Royal College of Psychiatrists. https://www.rcpsych.ac.uk/docs/default-source/improving-care/better-mh-policy/college-reports/college-report-cr166.pdf?sfvrsn=d5fa2c24_2

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small-and medium-scale evaluations. *American Journal of Evaluation*, *33*(3), 414-430. https://doi.org/10.1177/1098214012441499

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology*, *52*(1), 141-166. https://doi.org/10.1146/annurev.psych.52.1.141

Ryff, C. D., & Singer, B. H. (2008). Know thyself and become what you are: A eudaimonic approach to psychological well-being. *Journal of Happiness Studies*, *9*, 13-39. https://doi.org/10.1007/s10902-006-9019-0

Saddler, C. D., & Sacks, L. A. (1993). Multidimensional perfectionism and academic procrastination: Relationships with depression in university students. *Psychological Reports*, *73*(3), 863-871. https://doi.org/10.2466/pr0.1993.73.3.863

Schaufeli, W. B., & Bakker, A. B. (2004). Job demands, job resources, and their relationship with burnout and engagement: a multi-sample study. *Journal of Organizational Behavior*, *25*(3), 293-315. https://doi:10.1002/job.248

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The measurement of work engagement. *Educational and Psychological Measurement*, *66*(4), 33-40. https://doi.org/10.1177/0013164405282471

Schechtman, E. (2002). Odds ratio, relative risk, absolute risk reduction, and the number needed to treat-which of these should we use?. *Value in Health*, *5*(5), 431-436. https://doi.org/10.1046/J.1524-4733.2002.55150.x

Schippers, M. C. (2017). *Ikigai: Reflection on life goals optimizes performance and happiness.* (Inaugural address). https://repub.eur.nl/pub/100484/27710_Oratie_Boekje_Micheala_Schippers_ONLINE.PDF

Schippers, M. C., Morisano, D., Locke, E. A., Scheepers, A. W. A., Latham, G. P., de Jong, E. M., (2020). Writing about personal goals and plans regardless of goal type boosts academic performance, *Contemporary Educational Psychology*, *60*, Article 101823. https://doi.org/10.1016/j.cedpsych.2019.101823

Schippers, M. C., Scheepers, W. A. & Peterson, J. B. (2015). A scalable goal-setting intervention closes both the gender and ethnic minority achievement gap. *Palgrave Communications*, *1*(1), 1-12. https://doi:10.1057/palcomms.2015.14

Schippers, M. C., & Ziegler, N. (2019). Life crafting as a way to find purpose and meaning in life. *Frontiers in Psychology*, *10*, Article 2778. https://doi.org/10.3389/fpsyg.2019.02778

Schmidt, H. G., Cohen-Schotanus, J., Van Der Molen, H. T., Splinter, T. A. W., Bulte, J., Holdrinet, R., & Van Rossum, H. J. M. (2010). Learning more by being taught less: A "time-for-self-study" theory explaining curricular effects on graduation rate and study duration. *Higher Education*, *60*, 287-300. https://doi.org/10.1007/s10734-009-9300-3

Searle, J. R. (1999). *Mind, language and society: Philosophy in the real world.* Basic Books.

Sheldon, K. M., & Houser-Marko, L. (2001). Self-concordance, goal attainment, and the pursuit of happiness: can there be an upward spiral? *Journal of Personality and Social Psychology*, *80*, 152–165. https://doi.org/10.1037/0022-3514.80.1.152

Sheldon, K. M. (2002). The self-concordance model of healthy goal striving: when personal goals correctly represent the person. In E. L. Deci & R. M. Ryan (Eds.). *Handbook of self-determination research*. pp. 65-86. The University of Rochester Press.

Sheldon, K. M., & Kasser, T. (1998). Pursuing personal goals: skills enable progress, but not all progress is beneficial. *Personality and Social Psychology Bulletin*, *24*, 1319–1331. https://doi.org/10.1177/01461672982412006

Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., Tomassetti, S., Nussbaum, A. D., & Cohen, G. L. (2013). Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology*, *104*(4), 591–618. https://doi.org/10.1037/a0031495

Shum, H., He, X., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*, 10-26. https://doi.org/10.1631/FITEE.1700826

Sitzmann, T. & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin*, *137*(3), 421–442. https://doi.org/10.1037/a0022777

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15-21. https://doi.org/10.3102/0013189X031007015

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5-14. https://doi.org/10.3102/0013189X08314117

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education, *Educational Psychologist, 55*(1), 21-31. https://doi.org/10.1080/00461520.2019.1611432

Sneyers, E., & De Witte, K. (2017). The effect of an academic dismissal policy on dropout, graduation rates and student satisfaction. Evidence from the Netherlands. *Studies in Higher Education*, *42*(2), 354-389. https://doi.org/10.1080/03075079.2015.1049143

Sone, T., Nakaya, N., Ohmori, K., Shimazu, T., Higashiguchi, M., Kakizaki, M., … Tsuji, I. (2008). Sense of life worth living (ikigai) and mortality in japan: Ohsaki study. *Psychosomatic Medicine*, *70*(6), 709-715. https://doi.org/10.1097/PSY.0b013e31817e7e64

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2016). A coming crisis in teaching?: Teacher supply, demand, and shortages in the U.S. Palo Alto, CA: *Learning Policy Institute.* https://lemanncenter.stanford.edu/sites/default/files/A_Coming_Crisis_in_Teaching_BRIEF.pdf

Steel, P., Brothen, T., & Wambach, C. (2001). Procrastination and personality, performance, and mood. *Personality and Individual Differences*, *30*(1), 95-106. https://doi.org/10.1016/S0191-8869(00)00013-1

Stes, A., & Van Petegem, P. (2011). Instructional development for early career academics: An overview of impact. *Educational Research*, *53*(4), 459-474. https://doi.org/10.1080/00131881.2011.625156

Stewart, G., Kamata, A., Miles, R., Grandoit, E., Mandelbaum, F., Quinn, C., & Rabin, L. (2019). Predicting mental health help seeking orientations among diverse undergraduates: An ordinal logistic regression analysis. *Journal of Affective Disorders*, *257*, 271-280. https://doi.org/10.1016/j.jad.2019.07.058

Stinebrickner, R., & Stinebrickner, T. R. (2003). Working during school and academic performance. *Journal of labor Economics, 21*(2), 473-491. https://doi.org/10.1086/345565

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*, 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Taherdoost, H. (2018). A review of technology acceptance and adoption models and theories. *Procedia Manufacturing*, *22*, 960-967. https://doi.org/10.1016/j.promfg.2018.03.137

The Netherlands Association of Universities of Applied Sciences 'Vereniging Hogescholen' (2020). *Cijfers en feiten databank*. https://www.vereniginghogescholen.nl/kennisbank/feiten-en-cijfers

Theune, K. (2015). The working status of students and time to degree at German universities. *Higher Education, 70*(4), 725-752. https://doi.org/10.1007/s10734-015-9864-z

Tight, M. (2021). *Syntheses of higher education research: What we know*. Bloomsbury.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89-125. https://doi.org/10.3102/00346543045001089

Tinto, V. (1993). Building community. *Liberal Education, 79*(4), 16-21.

Tinto, V. (1998). Colleges as communities: Taking research on student persistence seriously. *The Review of Higher Education*, *21*(2), 167-177. muse.jhu.edu/article/30046

Tinto, V. (1999). Taking retention seriously: Rethinking the first year of college. *NACADA Journal*, *19*(2), 5-9. https://doi.org/10.12930/0271-9517-19.2.5

Travers, C. J., Morisano, D., & Locke, E. A. (2015). Self-reflection, growth goals, and academic outcomes: A qualitative study. *British Journal of Educational Psychology*, *85*(2), 224-241. https://doi.org/10.1111/bjep.12059

Tuononen, T., Parpala, A., Mattsson, M., & Lindblom-Ylänne, S. (2016). Work experience in relation to study pace and thesis grade: Investigating the mediating role of student learning. *Higher Education, 72*(1), 41-58. https://doi.org/10.1007/s10734-015-9937-z

UNESCO (2017). *Education for Sustainable Development Goals: Learning objectives.* Paris: UNESCO. Retrieved February 7, 2019, from https://www.unesco.de/sites/default/files/2018-08/unesco_education_for_sustainable_development_goals.pdf.

Van der Heijden, H. (2004). User acceptance of hedonic information systems. *MIS Quarterly*, *28*(4), 695-704. https://doi.org/10.2307/25148660

Van Eerde, W., & Klingsieck, K. B. (2018). Overcoming procrastination? A meta-analysis of intervention studies. *Educational Research Review, 25*, 73-85. https://doi.org/10.1016/j.edurev.2018.09.002

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186-204. https://doi.org/10.1287/mnsc.46.2.186.11926

Van der Zanden, P. J., Denessen, E., Cillessen, A. H., & Meijer, P. C. (2018). Domains and predictors of first-year student success: A systematic review. *Educational Research Review*, *23*, 57-77. https://doi.org/10.1016/j.edurev.2018.01.001

Van Lent, M. (2019). Goal Setting, information, and goal revision: A field experiment. *German Economic Review*, *20*(4), 949-972. https://doi.org/10.1111/geer.12199

Van Lent, M., & Souverijn, M. (2020). Goal setting and raising the bar: A field experiment. *Journal of Behavioral and Experimental Economics*, *87*, Article 101570. https://doi.org/10.1016/j.socec.2020.101570

Van Merriënboer, J. J. G., Kirschner, P. A. & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist*, *38*(1), 5-13, https://doi.org/10.1207/S15326985EP3801_2

Vosniadou, S. (2020). Bridging secondary and higher education. The importance of self-regulated learning. *European Review*, *28*(S1), 103. https://doi.org/10.1017/S1062798720000939

Vossensteyn, H., Kottmann, A., Jongbloed, B., Kaiser, F., Cremonini, Stensaker, B., et al. (2015). *Dropout and completion in higher education in Europe main report*. Retrieved from http://ec.europa.eu/dgs/education_culture/repository/education/library/study/2015/dropout-completion-he-summary_en.pdf.

Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science*, *23*(1), 73-82. https://doi.org/10.1177/0963721413512856

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, *18;331*(6023):1447-1451. https://doi:10.1126/science.1198364

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *Journal of Educational Psychology*, *107*(2), 468-485. https://doi.org/10.1037/a0037461

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 3-14). https://doi.org/10.1145/2632048.2632054

Wang, H., Kong, M., Shan, W., & Vong, S. K. (2010). The effects of doing part□time jobs on college student academic performance and social life in a Chinese society. *Journal of Education and Work*, *23*(1), 79–94. https://doi.org/10.1080/13639080903418402

Warren, J. R. (2002). Reconsidering the relationship between student employment and academic outcomes: A new theory and better data. *Youth and Society, 33*(3), 366–393. https://doi.org/10.1177/0044118X02033003002

Warwick, K., & Shah, H. (2014). Good machine performance in Turing's imitation game. *IEEE Transactions on Computational Intelligence and AI in Games*, *6*(3), 289-299. https://doi.org/10.1109/TCIAIG.2013.2283538

Waterman, A. S. (1993). Two conceptions of happiness: Contrasts of personal expressiveness (eudaimonia) and hedonic enjoyment. *Journal of Personality and Social Psychology*, *64*(4), 678–691. https://doi.org/10.1037/0022-3514.64.4.678

Weidauer, A. (2018, September 27). Conversational AI: Your guide to five levels of AI assistants in enterprise. *Rasa*. https://blog.rasa.com/conversational-ai-your-guide-to-five-levels-of-ai-assistants-in-enterprise/

Weizenbaum, J. (1966). A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45. https://doi.org/10.1145/365153.365168

Willcoxson, L. (2010). Factors affecting intention to leave in the first, second and third year of university studies: a semester-by-semester investigation. *Higher Education Research and Development*, *29*(6), 623–639. https://doi.org/10.1080/07294360.2010.501071

Wikan, G., & Bugge, L. (2014). Student performance in teacher education in Norway: The impact of student, institutional and structural factors. *European Journal of Teacher Education, 37*(4), 442-452. https://doi.org/10.1080/02619768.2014.912626

Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In: *Academy of Management Annual Meeting*. Chicago, IL, United States. https://doi.org/10.5465/AMBPP.2018.15903abstract

Wolbert, L. S., De Ruyter, D. J., & Schinkel, A. (2018). What kind of theory should theory on education for human flourishing be? *British Journal of Educational Studies, 67*(1), 25-39. https://doi.org/10.1080/00071005.2017.1390061

Wrigley, T. (2018). The power of 'evidence': Reliable science or a set of blunt tools?. *British Educational Research Journal*, *44*(3), 359-376. https://doi.org/10.1002/berj.3338

Zarouali, B., Van den Broeck, E., Walrave, M., & Poels, K. (2018). Predicting consumer responses to a chatbot on Facebook. *Cyberpsychology, Behavior, and Social Networking*, *21*(8), 491-497. https://doi.org/10.1089/cyber.2017.0518

Zhao, Y. (2017). What works may hurt: Side effects in education. *Journal of Educational Change*, *18*(1). 1-19. https://doi.org/10.1007/s10833-016-9294-4

Zivin, K., Eisenberg, D., Gollust, S. E., & Golberstein, E. (2009). Persistence of mental health problems and needs in a college student population. *Journal of Affective Disorders*, *117*(3), 180-185. https://doi.org/10.1016/j.jad.2009.01.001

Appendix A
Literature Overview

Table A.1

Summary of Empirical Findings about the Effects of Goal-setting Interventions in Higher Education

Authors (year)	Type of goal-setting	Study design	Sample	Outcome	Limitations
Bettinger & Baker (2014)	Coaching and Goal setting	Field experiment	13,555 American from private, public and proprietary Ba and Ad undergrad. students	Significant ($p <$ .01) positive effects on retention: 5 percentage point less drop-out	Block randomization; No multilevel analysis or subgroup analysis; Untransparent fidelity
Clark et al. (2019)	Grade goals and Task goals	Field experiment	3,971 first-year Microeconomics course students from an American public University	Robust significant ($p <$ .001) positive effect of task-goals on performance for male students. No effect of grade goals	Untransparent fidelity measures; Effect size not reported
Dobronyi et al. (2019)	Reflective goal setting	Field experiment	1,356 First-year Economics students from a Canadian university	No effect on GPA, course credits or persistence. Enough power for a 7% standardized performance effect at $p <$ .05	Untransparent fidelity measures. Data analysis of retention (binary) done with OLS instead of logistic regression
Latham & Brown (2006)	Learning goals and Outcome goals	Field experiment	125 international Mba students at a Canadian university	Setting a distal outcome goal that includes proximal goals, or setting learning goals resulted in higher GPA than urging people to 'do your best'	Small sample. Relatively high risk of contamination because the control group participants were aware of the goal-setting treatments
Morisano et al. (2010)	Reflective goal-setting	Field experiment	85 Struggling undergraduate students from a Canadian University	Positive significant ($p <$ 0.01) effect (d = .65) on GPA.	Small sample of only struggling students who voluntarily joined the experiment
Schippers et al. (2020)	Reflective goal-setting.	Time-lagged quasi-experiment with	2.928 first-year students of a Dutch Business School. Treatment cohorts: 1,409.	21.6% (d = .34, p < .001) more course credits obtained in treatment cohorts.	No randomization. Partly transparent fidelity measures.

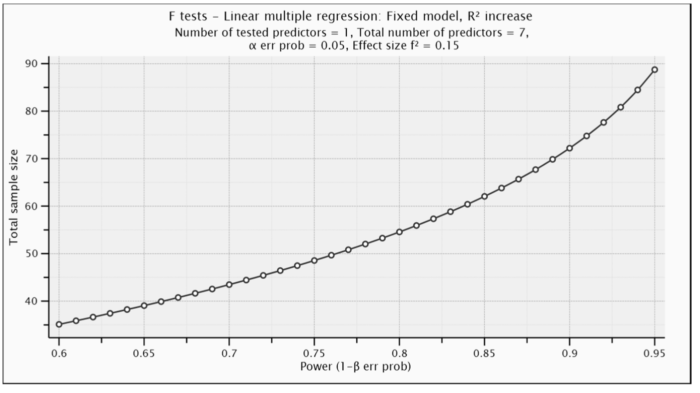
		two treatment and two control cohorts.	Control cohorts: 1,519		
Schippers et al. (2015)	Reflective goal-setting	Time-lagged quasi-experiment with three control cohorts.	703 (treatment cohort) compared to 2,441 first-year students (control cohorts) from a Dutch business school	On average 22% more obtained course credits. Large effect (d = .56, p < .001) for minority males (44% more course credits)	No randomization. Partly transparent fidelity measures
Van Lent (2019)	Grade goals (+optional different goals)	Field experiment	2,100 undergraduate Economics students from a Dutch university	No main treatment effect. Does work for students with low GPA. No effect for students who procrastinate or are present biased	The grade for one course is used as dependent variable, not succeeding in multiple courses. Randomization at group level
Van Lent & Souverijn (2020)	Grade goals and mentor encouragement	Field experiment	1,092 first-year Economics students from a Dutch university	Small (9 % of a standard deviation) positive effects on grades through drop-out. Significant for females (p < .01)	Randomization at mentor level; Suboptimal delivery fidelity

Note. The search terms "Goal setting" AND "Academic Performance" OR "Academic Success" yielded 144 peer-reviewed records in the Worldcat search engine. All titles and abstracts were screened. Reference lists from relevant articles were screened for additional hits. All 9 records that used a (quasi-)experimental design to measure the effects of a goal-setting intervention on academic performance (course credits, GPA, or retention) were included in this table.

Appendix B
Extended Data Analysis

Figure B.1

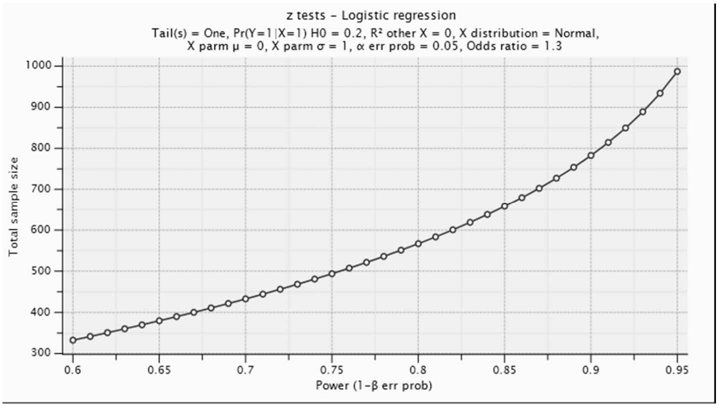
Power Analysis for Measuring Intervention Effects on Course Credits



Note. This power analysis was used for our tests of the direct effects of the intervention on course credits and potential moderator effects. The sample was 1,134. Whenever we corrected for multilevel structure the effective *n* was: 1,134 / (1 + 86 * 0.1) = 118. This is still leads to a power level of .95 for a f = .15 effect size or a .90 power level for finding a f = .10 effect size.

Figure B.2

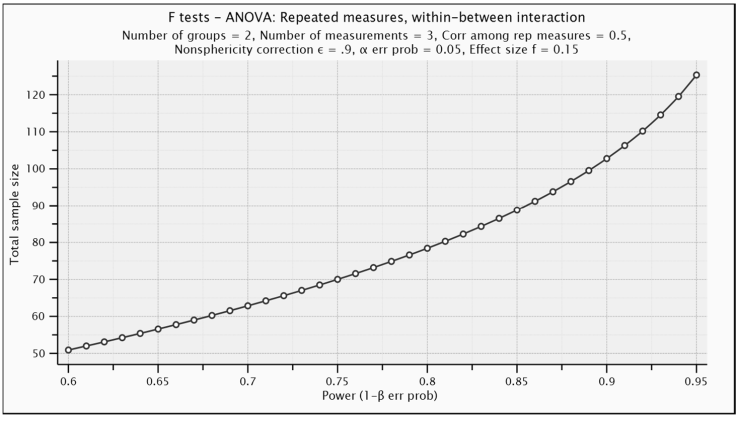
Power Analysis for Measuring Intervention Effect on Drop-out with Logistic Regression



Note. This power analysis was used for our tests of the direct effects of the intervention on dropout with logistic regression. The sample was 1,134. No correction for multilevel structure was required.

Figure B.3

Power Analysis for Repeated Measures Intervention Effect on Psychological Constructs (f = 0.15)



Note. This power analysis was used for our tests of the effects of the intervention on the psychological constructs. The smallest sample was n = 1,046. After correcting for multilevel structure this leads to an effective n of 108.

Figure B.4

Population Pyramid Frequency Course Credits T1

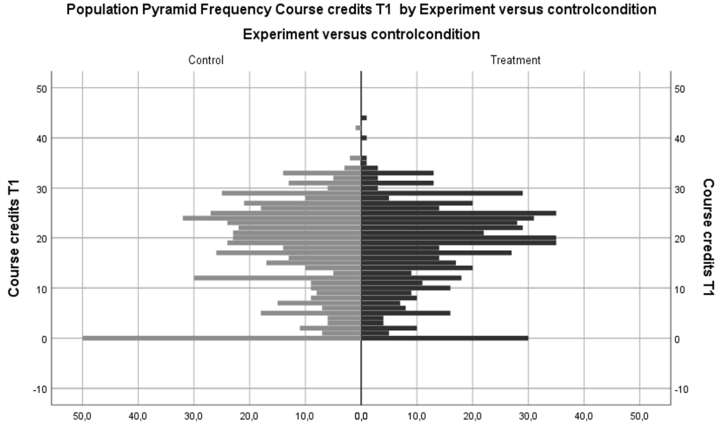


Figure B.5

Population Pyramid Frequency Course Credits T2

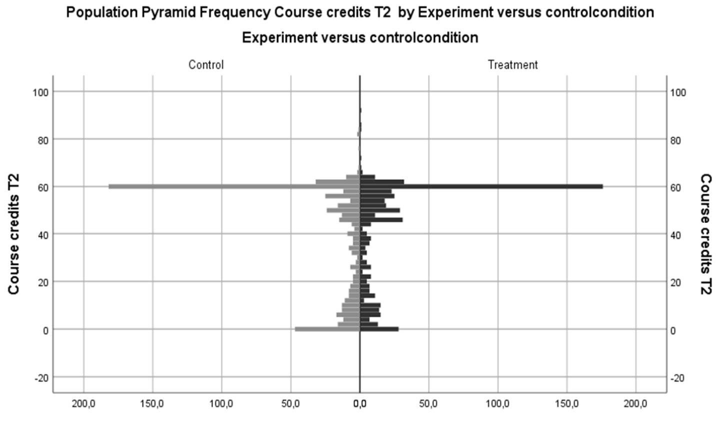


Table B.1

Results CFA of T0-T2

Measures	1st model T0	1st model T1	1st model T2	2nd model T0	2nd model T1	2nd model T2
χ^2	5810.25	4991.94	6484.11	5388.90	4359.32	5496.47
df	1,916	1,916	1,916	1,793	1,793	1,793
p	.000	.000	.000	.000	.000	.000
RMSEA	.046	.056	.060	.046	.053	.056
(90% CI)	(.045-.047)	(.054-.058)	(.058-.061)	(.044-.047)	(.051-.055)	(.054-.057)
CFI	.89	.84	.77	.90	.86	.81
TLI	.88	.83	.76	.89	.85	.80

RMSEA = root mean squared error of approximation; CI = confidence interval; CFI = comparative fit index; TLI = Tucker Lewis index). Sample sizes: T0 = 960; T1 = 505; T2 = 666.

Note. Original models with 64 items and final models with 62 items. Models fitted with option 'Categorical', estimated with Weighted Least Squares with Means and Variances (Muthén & Muthén, 1998-2006).

Table B.2.0

Establishing Random Part with Multilevel Analyses for Course Credits at T1

Effect	Course credits		
	Model 1	Model 2	Model 3
	Fixed effects		
Intercept	17.75 (0.28)	17.24 (0.94)	18.71 (2.02)
	Random effects		
Student variance	85.72 (3.62)	77.04 (3.27)	77.06 (3.27)
Course variance		10.13 (4.46)	6.36 (3.27)
Faculty variance			6.26 (8.06)
Total variance		87.17	89.69
	Goodness of fit		
Deviance	8,192.84	8,102.86	8,101.35
Model of reference		Model 1	Model 2
Chi-square fit improvement		$\chi^2 = 89.97$	$\chi^2 = 1.51$
P-value		$df = 1$	$df = 1$
		$p < .001$	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is course credits at T1. Student n = 1,134; course of study n = 13; faculty n = 2. n.s. = non-significant; df = degrees of freedom.

Table B.2.1

Results Multilevel Analyses of Moderator Effects with Course Credits at T1

Effect	Parameter	Course credits			
		Model 1	Model 2	Model 3	Model 4
		Fixed effects			
Intercept	γ_{00}	20.33 (0.99)	20.45 (1.00)	20.50 (1.00)	20.59 (1.01)
Intervention (= 1)	γ_{10}	1.09* (0.50)	0.855 (0.59)	0.76 (0.60)	0.57 (0.69)
Vocational background (= 1)	γ_{20}	-3.60*** (0.59)	-4.01*** (0.82)	-3.57*** (0.60)	-3.58*** (0.59)
Ethnic minority b. (= 1)	γ_{30}	-3.54*** (0.59)	-3.53*** (0.59)	-4.09*** (0.81)	-3.56*** (0.59)
Male (= 1)	γ_{40}	-3.20*** (0.55)	-3.20*** (0.25)	-3.21*** (0.55)	-3.74*** (0.74)
Intervention*Vocational backgr.	γ_{50}		0.84 (1.12)		
Intervention*Ethnic min.	γ_{60}			1.08 (1.09)	
Intervention*Male	γ_{70}				1.07 (1.00)
		Random effects			
Student variance	e_{0ij}	69.73 (2.96)	69.70 (2.96)	69.68 (2.96)	69.67 (2.96)
Course variance	μ_{0j}	8.99 (3.99)	8.94 (3.97)	8.89 (3.95)	8.85 (3.94)
Total variance	$\mu_{0j}+ e_{0ij}$	78.72	78.64	78.567	78.52
% expl. var. course level			0.04	0.07	0.09
% expl. var. student level			0.56	1.12	1.58
% expl. var. total			0.10	0.20	0.26
		Goodness of fit			
Deviance		7,990.59	7,990.04	7,989.62	7,989.46
Model of reference			Model 1	Model 1	Model 1
Chi-square fit improvement			$\chi^2_{(1)} = 0.55$	$\chi^2_{(1)} = 0.97$	$\chi^2_{(1)} = 1.13$
P-value			p = n.s.	p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is course credits at T1. Student n = 1,134; course of study n = 13; faculty n = 2. *$p < .05$ **$p < .01$ ***$p < .001$

Table B.2.2

Results Multilevel Analyses of Moderator Effect of Domain with Course Credits at T1

Effect	Parameter	Course credits	
		Model 1	Model 2
		Fixed effects	
Intercept	γ_{00}	21.39 (1.63)	21.37 (1.68)
Intervention (= 1)	γ_{10}	1.07* (0.53)	1.33 (1.02)
Teacher Education (= 1)	γ_{01}	-5.93** (1.63)	-5.77** (1.85)
Intervention*Teacher Education	γ_{11}		-0.36 (1.12)
		Random effects	
Student variance	e_{0ij}	78.00 (3.30)	77.99 (3.30)
Course variance	μ_{0j}	4.31 (2.16)	4.32 (2.164)
Total variance	$\mu_{0j+}\ e_{0ij}$	82.31	81.31
% expl. var. course level			0
% expl. var. student level			0
% expl. var. total			0
		Goodness of fit	
Deviance		8,179.76	8,179.67
Model of reference			Model 1
Chi-square fit improvement			$\chi^2_{(1)} = 0.09$
P-value			p = n.s.

Note. Standard errors are in parentheses. Dependent variable is course credits at T1. Student n = 1,134; course of study n = 13; faculty n = 2. *p < .05 **p < .01 ***p < .001

Table B.3.0

Establishing Random Part with Multilevel Analyses for Course Credits at T2

Effect	Parameter	Course credits		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	42.01 (0.67)	41.82 (0.90)	41.82 (0.67)
		Random effects		
Student variance	e_{0ijk}	508.26 (21.35)	505.05 (21.32)	505.05 (21.32)
Course variance	μ_{0jk}		3.74 (3.86)	3.74 (3.86)
Faculty variance	ν_{0k}			0.00 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	508.26	508.79	508.79
		Goodness of fit		
Deviance		10,284.11	10,282.96	10,282.96
Model of reference			Model 1	Model 1
Chi-square fit improvement			$\chi^2_{(1)} = 1.15$	$\chi^2_{(1)} = 1.15$
P-value			p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is course credits at T2. Student n = 1,134; course of study n = 13; faculty n = 2. *$p < .05$ **$p < .01$ ***$p < .001$

Table B.3.1

Results Multilevel Analyses of Moderator Effects on Credits at T2

Effect	Parameter	Course credits			
		Model 1	Model 2	Model 3	Model 4
		Fixed effects			
Intercept	γ_0	49.21 (1.27)	49.72 (1.33)	49.57 (1.34)	50.31 (1.42)
Intervention (=1)	γ_1	2.53* (1.28)	1.46 (1.50)	1.83 (1.53)	0.41 (1.77)
Vocational background (=1)	γ_2	-9.95*** (1.49)	-4.01*** (.82)	-9.90*** (1.49)	-9.90*** (1.49)
Ethnic minority backg. (=1)	γ_3	-7.01*** (1.46)	-3.53*** (.59)	-8.18*** (2.04)	-7.07*** (1.459)
Male (=1)	γ_4	-7.50*** (1.30)	-7.47*** (1.30)	-7.52*** (1.30)	-9.73*** (1.83)
Intervention*Vocational backg.	γ_5		3.88 (2.86)		
Intervention*Ethnic min. backg.	γ_6			2.29 (2.78)	
Intervention*Male	γ_7				4.43 (2.55)
		Random effects			
Student variance	e_{0i}	462.26 (19.41)	461.51 (19.38)	461.98 (19.40)	461.04 (19.36)
% expl. var. student level (= total variance)			0.16	0.06	0.27
		Goodness of fit			
Deviance		10,176.53	10,174.68	10,175.84	10,173.53
Model of reference			Model 1	Model 1	Model 1
Chi-square fit improvement			$\chi^2_{(1)} = 1.85$	$\chi^2_{(1)} = 0.68$	$\chi^2_{(1)} = 3.00$
P-value			p = n.s.	p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is course credits at T2. Student n = 1,134; course of study n = 13; faculty n = 2. *$p < .05$ **$p < .01$ ***$p < .001$

Table B.3.2

Results Multilevel Analyses of Moderator Effects of Domain on Credits at T2

Effect	Parameter	Course credits	
		Model 1	Model 2
		Fixed effects	
Intercept	γ_0	41.62 (1.63)	41.81 (1.77)
Intervention (= 1)	γ_1	2.76* (1.34)	2.36 (2.60)
Teacher Education (= 1)	γ_2	-1.36 (1.51)	-1.62 (2.09)
Intervention*Teacher Ed.	γ_3		0.55 (3.03)
		Random effects	
Student variance	e_{0i}	506.08 (21.25)	506.06 (21.25)
% expl. var. student level (= total variance)			0.00
		Goodness of fit	
Deviance		10,279.22	10,279.18
Model of reference			Model 1
Chi-square fit improvement			$\chi^2_{(1)} = 0.04$
P-value			p = n.s.

Note. Standard errors are in parentheses. Dependent variable is course credits at T1. Student n = 1,134; course of study n = 13; faculty n = 2. n.s. = non-significant $^*p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table B.4

Logistic Regression Results

Effect	Model 1	Model 2	Model 3	Model 4
Constant	-0.56 (0.06)	-0.43 (0.09)	-1.38 (0.12)	-1.25 (0.13)
Intervention		-0.26* (0.12)		-0.26* (0.13)
Vocational background			0.83***	0.83***
			(0.15)	(0.15)
Ethnic minority backg.			0.64***	0.64***
			(0.14)	(0.14)
Male			0.76***	0.76***
			(0.13)	(0.13)
Nagelkerke r^2		0.01	0.10	0.11
Δ Nagelkerke r^2				0.01
Model of reference		Model 1	Model 1	Model 3
Chi-square fit improvement		$\chi^2_{(1)} = 4.38$	$\chi^2_{(3)} = 87.28$	$\chi^2_{(1)} = 4.14$
P-value		$p < .05$	$p < .001$	$p < .05$

Note. Dependent variable is 'Drop-out': higher regression coefficients indicate higher chances of dropping out. Student n = 1,134. *$p < .05$ **$p < .01$ ***$p < .001$

Table B.5.1

Results Multilevel Repeated Measures 'Effort regulation'

Effect	Parameters of Model 4 and 5	Effort regulation				
		Model 1	Model 2	Model 3	Model 4	Model 5
		Fixed effects				
Intercept	γ_{000}	4.02 (0.03)	4.08 (0.05)	4.06 (0.06)	4.05 (0.05)	4.04 (0.06)
Trend	γ_{100}	-0.29*** (0.01)	-0.29*** (0.01)	-0.29*** (0.01)	-0.29*** (0.01)	-0.28*** (0.02)
Intervention (= 1)	γ_{010}				0.07 (0.04)	0.08 (0.06)
Intervention*Trend	γ_{110}					-0.01 (0.02)
		Random effects				
Rep. meas. variance	e_{0ijk}	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)
Student variance	μ_{0jk}	0.22 (0.02)	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)
Course variance	ν_{0k}		0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Faculty variance				0.00 (0.00)		
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.42	0.42	0.42	0.42	0.42
		Goodness of fit				
Deviance		3,740.37	3,711.37	3,711.39	3,707.78	3,707.68
Model of reference			Model 1	Model 2	Model 2	Model 4
Chi-square fit improvement			$\chi^2_{(1)} =$ 29	$\chi^2_{(1)} = 0$ $p =$ n.s.	$\chi^2_{(1)} =$ 3.59	$\chi^2_{(1)} =$ 0.07
P-value			$p < .001$		$p < .05$	$p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is effort regulation, measured 3 times (repeated measures n = 2,102; student n = 1,050; course of study n = 13; faculty n = 2). n.s. = non-significant *$p < .05$ **$p < .01$ ***$p < .001$

Table B.5.2

Results Multilevel Repeated Measures ‘Effort regulation’ with Gender as Moderator

Effect	Parameter	Effort regulation		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	4.15 (0.05)	4.13 (0.06)	4.13 (0.06)
Trend	γ_{100}	-0.29*** (0.01)	-0.27*** (0.02)	-0.26*** (0.02)
Intervention (= 1)	γ_{010}	0.06 (0.03)	0.03 (0.06)	0.05 (0.08)
Male (= 1)	γ_{020}	-0.24*** (0.04)	0.21** (0.07)	-0.20* (0.08)
Intervention*Trend	γ_{110}		-0.01 (0.02)	-0.01 (0.03)
Intervention*Male	γ_{030}		0.08 (0.07)	0.05 (0.11)
Trend*Male	γ_{120}		-0.04 (0.02)	-0.05 (0.03)
Intervention*Trend*Male	γ_{140}			0.02 (0.05)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)
Student variance	μ_{0jk}	0.19 (0.01)	0.19 (0.01)	0.19 (0.01)
Course variance	ν_{0k}	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.40	0.40	0.40
		Goodness of fit		
Deviance		3,666.20	3,662.27	3,662.16
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 3.93$	$\chi^2_{(1)} = 0.11$
P-value			p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is effort regulation, measured 3 times (repeated measures n = 2,102; student n = 1,050; course of study n = 13; faculty n = 2). n.s. = non-significant *p < .05 **p < .01 ***p < .001

Table B.5.3

Results Multilevel Repeated Measures 'Effort regulation' with Domain as Moderator

Effect	Parameter	Effort regulation		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.87 (0.10)	4.02 (0.11)	4.01 (0.11)
Trend	γ_{100}	-0.29*** (0.01)	-0.37*** (0.03)	-0.37*** (0.03)
Intervention (= 1)	γ_{010}	0.06 (0.04)	0.12 (0.08)	0.13 (0.11)
Teacher ed. (= 1)	γ_{020}	0.21* (0.10)	-0.01 (0.12)	-0.01 (0.13)
Intervention*Trend	γ_{110}		-0.02 (0.02)	-0.02 (0.05)
Intervention*Teacher ed.	γ_{030}		-0.04 (0.08)	-0.05 (0.13)
Trend*Teacher ed.	γ_{120}		0.13*** (0.03)	0.13** (0.04)
Intervention*Trend*Teacher ed.	γ_{140}			0.01 (0.05)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.20 (0.01)	0.19 (0.01)	0.19 (0.01)
Student variance	μ_{0jk}	0.20 (0.01)	0.21 (0.01)	0.21 (0.01)
Course variance	ν_{0k}	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.41	0.40	0.40
		Goodness of fit		
Deviance		3,704.17	3,679.69	3,679.68
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 24.48$	$\chi^2_{(1)} = 0.01$
P-value			p = <.001	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is effort regulation, measured 3 times (repeated measures n = 2,102; student n = 1,050; course of study n = 13; faculty n = 2). n.s. = non-significant *$p < .05$ **$p < .01$ ***$p < .001$

Table B.5.4

Results Multilevel Repeated Measures 'Effort regulation' with Ethnicity as Moderator

Effect	Parameter	Effort regulation		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	4.05 (0.05)	4.06 (0.06)	4.05 (0.06)
Trend	γ_{100}	-0.29*** (0.01)	-0.29*** (0.02)	-0.28*** (0.02)
Intervention (= 1)	γ_{010}	0.07 (0.04)	0.07 (0.06)	0.08 (0.07)
Ethnic minority (= 1)	γ_{020}	-0.01 (0.04)	-0.07 (0.07)	-0.04 (0.09)
Intervention*Trend	γ_{110}		-0.01 (0.02)	-0.02 (0.03)
Intervention*Ethnic minority	γ_{030}		0.04 (0.08)	-0.02 (0.12)
Trend*Ethnic minority	γ_{120}		0.02 (0.03)	-0.01 (0.04)
Intervention*Trend*Ethnic minority	γ_{140}			0.03 (0.05)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)
Student variance	μ_{0jk}	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)
Course variance	ν_{0k}	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.42	0.42	0.42
		Goodness of fit		
Deviance		3,707.70	3,706.78	3,706.48
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 0.92$	$\chi^2_{(1)} = 0.30$
P-value			p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is effort regulation, measured 3 times (repeated measures n = 2,102; student n = 1,050; course of study n = 13; faculty n = 2). n.s. = non-significant *p < .05 **p < .01 ***p < .001

Table B.6.1

Results Multilevel Repeated Measures 'Self-efficacy'

Effect	Parameters of Model 4 and 5	Self-efficacy				
		Model 1	Model 2	Model 3	Model 4	Model 5
		Fixed effects				
Intercept	γ_{000}	3.99 (.03)	3.99 (0.03)	3.99 (0.04)	3.99 (0.05)	3.97 (0.05)
Trend	γ_{100}	-0.08***	-0.08***	-0.08***	-0.08***	-0.07**
		(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Intervention (= 1)	γ_{010}				-0.01 (0.03)	0.04 (0.05)
Intervention *Trend	γ_{110}					-0.02 (0.02)
		Random effects				
Repeated measures variance	e_{0ijk}	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)
Student variance	μ_{0jk}	0.14 (0.01)	0.13 (0.01)	0.13 (0.01)	0.13 (0.01)	0.13 (0.01)
Course variance	ν_{0k}		0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Faculty variance				0.00 (0.00)		
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.34	0.34	0.34	0.34	0.34
		Goodness of fit				
Deviance		3,494.94	3,468.31	3,468.31	3,468.28	3,467.26
Model of reference			Model 1	Model 2	Model 2	Model 4
Chi-square fit improvement			$\chi^2_{(1)} = 26.63$	$\chi^2_{(1)} = 0$	$\chi^2_{(1)} = 0.03$	$\chi^2_{(1)} = 1.02$
P-value						
			$p < .001$	p = n.s.	p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is self-efficacy, measured 3 times (repeated measures n = 2,098; student n = 1,045; course of study n = 13; faculty n = 2). n.s. = non-significant *$p < .05$ **$p < .01$ ***$p < .001$

Table B.6.2

Results Multilevel Repeated Measures 'Self-efficacy' with Gender as Moderator

Effect	Parameter	Self-efficacy		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.93 (0.05)	3.93 (0.06)	3.91 (0.06)
Trend	γ_{100}	-0.08*** (0.01)	-0.07*** (0.02)	-0.06** (0.02)
Intervention (= 1)	γ_{010}	-0.00 (0.03)	0.02 (0.06)	0.05 (0.07)
Male (= 1)	γ_{020}	-0.14*** (0.03)	0.09 (0.06)	0.13 (0.08)
Intervention*Trend	γ_{110}		-0.03 (0.02)	-0.04 (0.03)
Intervention*Male	γ_{030}		0.05 (0.06)	-0.02 (0.10)
Trend*Male	γ_{120}		0.01 (0.02)	-0.01 (0.03)
Intervention*Trend*Male	γ_{140}			0.04 (0.05)
		Random effects		
Repeated measures variance	e_{0ijk}	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)
Student variance	μ_{0jk}	0.13 (0.01)	0.13 (0.01)	0.13 (0.01)
Course variance	ν_{0k}	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.34	0.34	0.34
		Goodness of fit		
Deviance		3,450.55	3,448.49	3,447.82
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 2.06$	$\chi^2_{(1)} = 0.67$
P-value			p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is self-efficacy, measured 3 times (repeated measures n = 2,098; student n = 1,045; course of study n = 13; faculty n = 2). n.s. = non-significant $*p < .05$ $**p < .01$ $***p < .001$

Table B.6.3

Results Multilevel Repeated Measures 'Self-efficacy' with Domain as Moderator

Effect	Parameter	Self-efficacy		
		Model 1	Model 2	Model 3
Fixed effects				
Intercept	γ_{000}	3.90	3.88	3.83
		(0.09)	(0.10)	(0.11)
Trend	γ_{100}	-	-0.06*	-0.03
		0.08***	(0.03)	(0.03)
		(0.01)		
Intervention (= 1)	γ_{010}	-0.01	-0.02	0.09
		(0.03)	(0.08)	(0.10)
Teacher ed. (= 1)	γ_{020}	0.11***	0.11	0.18
		(0.01)	(0.11)	(0.12)
Intervention*Trend	γ_{110}		-0.02	-0.08
			(0.02)	(0.05)
Intervention*Teacher ed.	γ_{030}		0.07	-0.08
			(0.07)	(0.12)
Trend*Teacher ed.	γ_{120}		-0.02	-0.06
			(0.03)	(0.04)
Intervention*Trend*Teacher ed.	γ_{140}			0.08
				(0.05)
Random effects				
Repeated measures variance	e_{0ijk}	0.20	0.20	0.20
		(0.01)	(0.01)	(0.01)
Student variance	μ_{0jk}	0.13	0.13	0.13
		(0.01)	(0.01)	(0.01)
Course variance	ν_{0k}	0.01	0.01	0.01
		(0.01)	(0.01)	(0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.34	0.34	0.34
Goodness of fit				
Deviance		3,467.08	3,464.43	3,462.20
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)}$ = 2.65	$\chi^2_{(1)}$ = 2.23
P-value			*p* = n.s.	*p* = n.s.

Note. Standard errors are in parentheses. Dependent variable is self-efficacy, measured 3 times (repeated measures *n* = 2,098; student *n* = 1,045; course of study *n* = 13; faculty *n* = 2). n.s. = non-significant **p* < .05 ***p* < .01 ****p* < .001

Table B.6.4

Results Multilevel Repeated Measures 'Self-efficacy' with Ethnicity as Moderator

Effect	Parameter	Self-efficacy		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.98 (0.05)	3.89 (0.05)	3.90
Trend	γ_{100}	-0.08*** (0.01)	-0.03 (0.02)	-0.03 (0.02)
Intervention (= 1)	γ_{010}	-0.01 (0.03)	0.01 (0.06)	0.00 (0.06)
Ethnic minority (= 1)	γ_{020}	0.02 (0.04)	0.28 (0.07)	0.26** (0.08)
Intervention*Trend	γ_{110}		-0.02 (0.02)	-0.02 (0.03)
Intervention*Ethnic minority	γ_{030}		0.08 (0.07)	0.11 (0.12)
Trend*Ethnic minority	γ_{120}		0.16*** (0.03)	-0.15*** (0.04)
Intervention*Trend*Ethnic minority	γ_{140}			-0.02 (0.05)
		Random effects		
Repeated measures variance	e_{0ijk}	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)
Student variance	μ_{0jk}	0.13 (0.01)	0.13 (0.01)	0.13 (0.01)
Course variance	ν_{0k}	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.34	0.34	0.34
		Goodness of fit		
Deviance		3,467.83	3,428.21	3,428.09
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 39.62$	$\chi^2_{(1)} = 0.12$
P-value			$p = < .001$	$p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is self-efficacy, measured 3 times (repeated measures $n = 2{,}098$; student $n = 1{,}045$; course of study $n = 13$; faculty $n = 2$). n.s. = non-significant $^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table B.7.1

Results Multilevel Repeated Measures 'Intrinsic Goal Orientation'

Effect	Parameter	Intrinsic goal orientation				
		Model 1	Model 2	Model 3	Model 4	Model 5
		Fixed effects				
Intercept	γ_{0000}	4.74 (0.03)	4.51 (0.04)	4.42 (0.10)	4.41 (0.10)	4.39 (0.10)
Trend	γ_{1000}	-0.26*** (0.01)	-0.26*** (0.01)	-0.26*** (0.01)	-0.26*** (0.01)	-0.25*** (0.02)
Intervention (= 1)	γ_{0100}				0.03 (0.03)	0.06 (0.05)
Intervention *Trend	γ_{1100}					-0.02 (0.02)
		Random effects				
Repeated measures variance	e_{0ijkl}	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)
Student variance	μ_{0jkl}	0.15 (0.01)	0.14 (0.01)	0.14 (0.01)	0.14 (0.01)	0.14 (0.01)
Course variance	ν_{0kl}		0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
Faculty variance	f_{0l}			0.02 (0.02)	0.02 (0.02)	0.02 (0.02)
Total variance	$e_{0ijkl} + \mu_{0jkl} + \nu_{0kl} + f_{0l}$	0.31	0.32	0.33	0.33	0.33
		Goodness of fit				
Deviance		3,208.93	3,161.69	3,156.55	3,155.82	3,155.17
Model of reference			Model 1	Model 2	Model 3	Model 4
Chi-square fit improvement *P*-value			$\chi^2_{(1)} =$ 47.24 $p <$ .001	$\chi^2_{(1)} =$ 5.14 $p <$ .05	$\chi^2_{(1)} =$ 0.73 $p =$ n.s.	$\chi^2_{(1)} =$ 0.65 $p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is intrinsic goal orientation, measured 3 times (repeated measures $n = 2{,}098$; student $n = 1{,}049$; course of study $n = 13$; faculty $n = 2$). n.s. = non-significant

*$p < .05$ **$p < .01$ ***$p < .001$

Table B.7.2

Results Multilevel Repeated Measures Analyses 'Intrinsic Goal Orientation' with Gender as Moderator

Effect	Parameter	Intrinsic goal orientation		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{0000}	4.55	4.53	4.53
		(0.08)	(0.09)	(0.09)
Trend	γ_{1000}	-	-	-
		0.26***	0.25***	0.25***
		(0.01)	(0.02)	(0.02)
Intervention (= 1)	γ_{0100}	0.02	0.05	0.04
		(0.03)	(0.06)	(0.07)
Male (= 1)	γ_{0200}	-	-	-0.24
		0.25***	0.23***	(0.07)
		(0.03)	(0.06)	
Intervention*Trend	γ_{1100}		-0.02	-0.01
			(0.02)	(0.03)
Intervention*Male	γ_{0300}		0.01	0.02
			(0.06)	(0.10)
Trend*Male	γ_{1200}		-0.01	-0.01
			(0.02)	(0.03)
Intervention*Trend*Male	γ_{1400}			-0.01
				(0.04)
		Random effects		
Repeated measures variance	e_{0ijkl}	0.16	0.16	0.16
		(0.01)	(0.01)	(0.01)
Student variance	μ_{0jkl}	0.12	0.12	0.12
		(0.01)	(0.01)	(0.01)
Course variance	ν_{0kl}	0.01	0.01	0.01
		(0.00)	(0.00)	(0.00)
Faculty variance	f_{0l}	0.01	0.01	0.01
		(0.01)	(0.01)	(0.01)
Total variance	$e_{0ijkl} + \mu_{0jkl} + \nu_{0kl} + f_{0l}$	0.30	0.30	0.30
		Goodness of fit		
Deviance		3,092.09	3,091.41	3,091.38
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} =$ 0.68	$\chi^2_{(1)} =$ 0.03
P-value			p = n.s.	p = n.s.

Note. Idem to Table B.7.1

Table B.7.3

Results Multilevel Repeated Measures Analyses 'Intrinsic Goal Orientation' with Domain as Moderator

Effect	Parameter	Intrinsic goal orientation		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	4.26 (0.06)	4.30 (0.07)	4.24 (0.08)
Trend	γ_{100}	-0.26*** (0.01)	-0.29*** (0.02)	-0.26*** (0.03)
Intervention (= 1)	γ_{010}	0.02 (0.03)	0.11 (0.07)	0.24** (0.10)
Teacher education (= 1)	γ_{020}	0.28*** (0.06)	0.20** (0.08)	0.28** (0.09)
Intervention*Trend	γ_{110}		-0.02 (0.02)	-0.09* (0.04)
Intervention*Teacher ed.	γ_{030}		-0.06 (0.07)	-0.24* (0.11)
Trend*Teacher ed.	γ_{120}		-0.06 (0.02)	0.01 (0.03)
Intervention*Trend*Teacher ed.	γ_{140}			0.10* (0.05)
		Random effects		
Repeated measures variance	e_{0ijk}	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)
Student variance	μ_{0jk}	0.14 (0.01)	0.14 (0.01)	0.14 (0.01)
Course variance	ν_{0k}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.30	0.30	0.30
		Goodness of fit		
Deviance		3,148.88	3,141.66	3,137.71
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 7.22$	$\chi^2_{(1)} = 3.95$
P-value			p = < n.s.	p = < .05

Note. Idem to Table B.7.1

Table B.7.4

Results Multilevel Repeated Measures Analyses 'Intrinsic Goal Orientation' with Ethnicity as Moderator

Effect	Parameter	Intrinsic goal orientation		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{0000}	4.37 (0.09)	4.36 (0.10)	4.35 (0.10)
Trend	γ_{1000}	-0.26*** (0.01)	-0.25*** (0.02)	-0.25 (0.02)
Intervention (= 1)	γ_{0100}	0.02 (0.03)	0.04 (0.05)	0.06 (0.06)
Ethnic minority (= 1)	γ_{0200}	0.16*** (0.03)	0.14** (0.06)	0.17* (0.08)
Intervention*Trend	γ_{1100}		-0.02 (0.02)	-0.03 (0.03)
Intervention*Ethnic min.	γ_{0300}		0.05 (0.07)	-0.01 (0.11)
Trend*Ethnic min.	γ_{1200}		-0.00 (0.04)	-0.02 (0.04)
Intervention*Trend*Ethnic min.	γ_{1400}			0.03 (0.05)
		Random effects		
Repeated measures variance	e_{0ijkl}	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)
Student variance	μ_{0jkl}	0.13 (0.01)	0.13 (0.01)	0.13 (0.01)
Course variance	ν_{0kl}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Faculty variance	f_{0l}	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)
Total variance	$e_{0ijkl} + \mu_{0jkl} + \nu_{0kl} + f_{0l}$	0.30	0.30	0.30
		Goodness of fit		
Deviance		3,132.88	3,131.66	3,131.23
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 1.22$	$\chi^2_{(1)} = 0.43$
P-value			p = n.s.	p = n.s.

Note. Idem to Table B.7.1

Table B.8.1

Results Multilevel Repeated Measures Analyses 'Metacognition'

Effect	Parameters of Models 4 and 5	Metacognition				
		Model 1	Model 2	Model 3	Model 4	Model 5
Fixed effects						
Intercept	γ_{000}	3.53 (0.03)	3.59 (0.06)	3.59 (0.06)	3.57 (0.06)	3.54 (0.06)
Trend	γ_{100}	-0.19*** (0.01)	-0.19*** (0.01)	-0.19*** (0.01)	-0.19*** (0.01)	-0.17*** (0.02)
Intervention (= 1)	γ_{010}				0.03 (0.04)	0.09 (0.06)
Intervention* Trend	γ_{110}					-0.03 (0.03)
Random effects						
Repeated measures variance	e_{0ijk}	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)
Student variance	μ_{0jk}	0.21 (0.02)	0.20 (0.02)	0.20 (0.02)	0.20 (0.02)	0.20 (0.02)
Course variance	ν_{0k}		0.03 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
Faculty variance				0.00 (0.00)		
Total variance	$e_{0ijkl} + \mu_{0jkl} + \nu_{0kl}$	0.46	0.48	0.48	0.48	0.48
Goodness of fit						
Deviance		4,038.73	4,012.24	4,012.24	4,011.73	4,010. 20.64
Model of reference			Model 1	Model 2	Model 2	Model 4
Chi-square fit improvement			$\chi^2_{(1)} = 26.49$	$\chi^2_{(1)} = 0$	$\chi^2_{(1)} = 0.51$	$\chi^2_{(1)} = 1.53$
P-value			$p < .001$	$p =$ n.s.	$p =$ n.s.	$p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is metacognition, measured 3 times (repeated measures $n = 2{,}090$; student $n = 1{,}047$; course of study $n = 13$; faculty $n = 2$). n.s. = non-significant *$p < .05$ **$p < .01$ ***$p < .001$

Table B.8.2

Results Multilevel Repeated Measures Analyses 'Metacognition' with Gender as Moderator

Effect	Parameter	Metacognition		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.64 (0.06)	3.69 (0.07)	3.68 (0.08)
Trend	γ_{100}	-0.19*** (0.01)	-0.20*** (0.02)	-0.20*** (0.03)
Intervention (= 1)	γ_{010}	0.02 (0.04)	0.04 (0.07)	0.03 (0.08)
Male (= 1)	γ_{020}	-0.15*** (0.04)	-0.30*** (0.07)	-0.29** (0.09)
Intervention*Trend	γ_{110}		-0.03 (0.03)	-0.03 (0.04)
Intervention*Male	γ_{030}		0.11 (0.07)	0.10 (0.12)
Trend*Male	γ_{120}		0.05 (0.03)	0.05 (0.04)
Intervention*Trend*Male	γ_{140}			0.01 (0.05)
		Random effects		
Repeated measures variance	e_{0ijk}	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)
Student variance	μ_{0jk}	0.19 (0.02)	0.19 (0.02)	0.19 (0.02)
Course variance	ν_{0k}	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.47	0.47	0.47
		Goodness of fit		
Deviance		3,996.05	3,988.59	3,988.58
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 7.46$	$\chi^2_{(1)} = 0.01$
P-value			$p = < .01$	$p =$ n.s.

Note. Idem to Table B.8.1

Table B.8.3

Results Multilevel Repeated Measures Analyses 'Metacognition' with Domain as Moderator

Effect	Parameter	Metacognition		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.38 (0.11)	3.21 (0.13)	3.17 (0.13)
Trend	γ_{100}	- 0.19*** (0.01)	-0.10 (0.03)	-0.08 (0.04)
Intervention (= 1)	γ_{010}	0.02 (0.04)	0.09 (0.09)	0.17 (0.12)
Teacher education (= 1)	γ_{020}	0.22 (0.12)	0.43 (0.14)	0.49 (0.15)
Intervention*Trend	γ_{110}		-0.03 (0.03)	-0.07 (0.05)
Intervention*Teacher ed.	γ_{030}		-0.03 (0.08)	-0.13 (0.14)
Trend*Teacher ed.	γ_{120}		-0.11 (0.03)	-0.14 (0.04)
Intervention*Trend*Teacher ed.	γ_{140}			0.06 (0.06)
		Random effects		
Repeated measures variance	e_{0ijk}	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)
Student variance	μ_{0jk}	0.20 (0.02)	0.20 (0.02)	0.20 (0.02)
Course variance	ν_{0k}	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.47	0.47	0.47
		Goodness of fit		
Deviance		4,008.63	3,993.67	3,992.71
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)}$ = 14.96	$\chi^2_{(1)}$ = 0.96
P-value			*p* = < .001	*p* = n.s.

Note. Idem to Table B.8.1

Table B.8.4

Results Multilevel Repeated Measures Analyses 'Metacognition' with Ethnicity as Moderator

Effect	Parameter	Metacognition		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.54 (0.06)	3.47 (0.07)	3.45 (0.07)
Trend	γ_{100}	-0.19*** (0.01)	-0.16 (0.02)	-0.14 (0.02)
Intervention (= 1)	γ_{010}	0.02 (0.04)	0.08 (0.06)	0.13 (0.07)
Ethnic minority (= 1)	γ_{020}	0.12** (0.04)	0.25** (0.08)	0.35*** (0.10)
Intervention*Trend	γ_{110}		-0.03 (0.03)	-0.06* (0.03)
Intervention*Ethnic min.	γ_{030}		0.00 (0.08)	-0.18 (0.13)
Trend*Ethnic min.	γ_{120}		-0.07* (0.03)	-0.13** (0.04)
Intervention*Trend*Ethnic min.	γ_{140}			0.10 (0.06)
		Random effects		
Repeated measures variance	e_{0ijk}	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)
Student variance	μ_{0jk}	0.19 (0.02)	0.19 (0.02)	0.20 (0.01)
Course variance	ν_{0k}	0.02 (0.01)	0.02 (0.01)	0.02 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.46	0.46	0.46
		Goodness of fit		
Deviance		4,002.95	3,995.43	3,992.58
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 7.52$	$\chi^2_{(1)} = 2.85$
P-value			$p = < .05$	$p = < .05$

Note. Idem to Table B.8.1

Table B.9.1

Results Multilevel Repeated Measures Analyses 'Attention'

Effect	Parameter	Attention				
		Model 1	Model 2	Model 3	Model 4	Model 5
				Fixed effects		
Intercept	γ_{0000}	3.65 (0.03)	3.70 (0.06)	3.60 (0.13)	3.57 (0.13)	3.57 (0.13)
Trend	γ_{1000}	-0.19*** (0.01)	-0.19*** (0.01)	-0.19*** (0.01)	-0.19*** (0.01)	-0.19*** (0.02)
Intervention (=1)	γ_{0100}				0.05 (0.04)	0.05 (0.06)
Intervention*Trend	γ_{1100}					-0.00 (0.02)
				Random effects		
Repeated measures variance	e_{0ijkl}	0.19 (0.01)	0.19 (0.01)	0.19 (0.01)	0.19 (0.01)	0.19 (0.01)
Student variance	μ_{0jkl}	0.31 (0.02)	0.28 (0.02)	0.28 (0.02)	0.28 (0.02)	0.28 (0.02)
Course variance	ν_{0kl}		0.03 (0.02)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Faculty variance	f_{0l}			0.03 (0.03)	0.03 (0.03)	0.03 (0.03)
Total variance	$e_{0ijkl} + \mu_{0jkl} + \nu_{0kl} + f_{0l}$	0.50	0.50	0.52	0.52	0.52
				Goodness of fit		
Deviance		3,877.48	3,836.80	3,833.77	3,832.20	3,832.20
Model of reference			Model 1	Model 2	Model 3	Model 4
Chi-square fit improvement			$\chi^2_{(1)}$ = 40.68	$\chi^2_{(1)}$ = 3.03	$\chi^2_{(1)}$ = 1.57	$\chi^2_{(1)}$ = 0.00
P-value			p = < .001	p = < .05	p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is attention, measured 3 times (repeated measures n = 2,086; student n = 1,050; course of study n = 13; faculty n = 2). n.s. = non-significant *p < .05 **p < .01 ***p < .001

Table B.9.2

Results Multilevel Repeated Measures Analyses 'Attention' with Gender as Moderator

Effect	Parameter	Attention		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{0000}	3.55 (0.14)	3.60 (0.14)	3.59 (0.14)
Trend	γ_{1000}	-0.19*** (0.01)	-0.21*** (0.02)	-0.21*** (0.02)
Intervention (= 1)	γ_{0100}	0.05 (0.04)	0.04 (0.07)	0.05 (0.08)
Male (= 1)	γ_{0200}	0.05 (0.04)	-0.06 (0.07)	-0.06 (0.08)
Intervention*Trend	γ_{1100}		-0.00 (0.02)	-0.00 (0.03)
Intervention*Male	γ_{0300}		0.02 (0.08)	0.01 (0.11)
Trend*Male	γ_{1200}		0.06 (0.02)	0.05 (0.03)
Intervention*Trend*Male	γ_{1400}			0.00 (0.05)
		Random effects		
Repeated measures variance	e_{0ijkl}	0.19 (0.01)	0.18 (0.01)	0.18 (0.01)
Student variance	μ_{0jkl}	0.28 (0.02)	0.28 (0.02)	0.28 (0.02)
Course variance	ν_{0kl}	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Faculty variance	f_{0l}	0.03 (0.03)	0.03 (0.03)	0.03 (0.03)
Total variance	$e_{0ijkl} + \mu_{0jkl} + \nu_{0kl} + f_{0l}$	0.52	0.51	0.51
		Goodness of fit		
Deviance		3,830.78	3,825.07	3,825.07
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 5.71$	$\chi^2_{(1)} = 0.00$
P-value				
			p = n.s.	p = n.s.

Note. Idem to Table B.9.1

Table B.9.3

Results Multilevel Repeated Measures Analyses 'Attention' with Domain as Moderator

Effect	Parameter	Attention		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.37 (0.10)	3.32 (0.11)	3.29 (0.12)
Trend	γ_{100}	-0.19*** (0.01)	-0.14*** (0.03)	-0.12*** (0.03)
Intervention (= 1)	γ_{010}	0.05 (0.04)	-0.03 (0.09)	0.04 (0.11)
Teacher education (= 1)	γ_{020}	0.36*** (0.10)	0.43*** (0.12)	0.48*** (0.13)
Intervention*Trend	γ_{110}		-0.00 (0.02)	-0.04 (0.05)
Intervention*Teacher ed.	γ_{030}		0.10 (0.09)	-0.00 (0.13)
Trend*Teacher education	γ_{120}		-0.06* (0.03)	-0.09 (0.04)
Intervention*Trend*Teacher	γ_{140}			0.06 (0.05)
		Random effects		
Repeated measures variance	e_{0ijk}	0.19 (0.01)	0.18 (0.01)	0.18 (0.01)
Student variance	μ_{0jk}	0.28 (0.02)	0.28 (0.02)	0.28 (0.02)
Course variance	ν_{0k}	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.48	0.47	0.47
		Goodness of fit		
Deviance		3,826.20	3,819.24	3,818.15
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 6.96$	$\chi^2_{(1)} = 1.09$
P-value			$p = < .05$	$p =$ n.s.

Note. Idem to Table B.9.1

Table B.9.4

Results Multilevel Repeated Measures Analyses 'Attention' with Ethnicity as Moderator

Effect	Parameter	Attention		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{0000}	3.54 (0.12)	3.50 (0.12)	3.49 (0.13)
Trend	γ_{1000}	-0.19*** (0.01)	-0.18*** (0.02)	-0.17*** (0.02)
Intervention (= 1)	γ_{0100}	0.05 (0.04)	0.09 (0.06)	0.11 (0.07)
Ethnic minority (= 1)	γ_{0200}	0.14** (0.04)	0.28** (0.08)	0.31** (0.09)
Intervention*Trend	γ_{1100}		-0.00 (0.02)	-0.01 (0.03)
Intervention*Ethnic min.	γ_{0300}		-0.14 (0.09)	-0.20 (0.13)
Trend*Ethnic min.	γ_{1200}		-0.04 (0.03)	-0.06 (0.04)
Intervention*Trend*Ethnic min.	γ_{1400}			0.04 (0.05)
		Random effects		
Repeated measures variance	e_{0ijkl}	0.19 (0.01)	0.19 (0.01)	0.18 (0.01)
Student variance	μ_{0jkl}	0.28 (0.02)	0.28 (0.02)	0.28 (0.02)
Course variance	ν_{0kl}	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)
Faculty variance	f_{0l}	0.02 (0.03)	0.02 (0.03)	0.02 (0.03)
Total variance	$e_{0ijkl} + \mu_{0jkl} + \nu_{0kl} + f_{0l}$	0.51	0.50	0.50
		Goodness of fit		
Deviance		3,822.22	3,817.51	3,817.03
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 4.71$	$\chi^2_{(1)} = 0.48$
P-value			p = n.s.	p = n.s.

Note. Idem to Table B.9.1

Table B.10.1

Results Multilevel Repeated Measures Analyses 'Resilience'

Effect	Parameter	Resilience				
		Model 1	Model 2	Model 3	Model 4	Model 5
Fixed effects						
Intercept	γ_{0000}	3.97 (0.02)	3.97 (0.04)	3.97 (0.04)	3.97 (0.04)	3.94 (0.04)
Trend	γ_{1000}	-0.06*** (0.01)	-0.06*** (0.01)	-0.06*** (0.01)	-0.06*** (0.01)	-0.04*** (0.01)
Intervention (=1)	γ_{0100}				-0.00 (0.03)	0.05 (0.05)
Intervention*Trend	γ_{1100}					-0.03 (0.02)
Random effects						
Repeated measures variance	e_{0ijkl}	0.14 (0.01)	0.14 (0.01)	0.14 (0.01)	0.14 (0.01)	0.14 (0.01)
Student variance	μ_{0jkl}	0.17 (0.01)	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)
Course variance	ν_{0kl}		0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Faculty variance	f_{0l}			0.00 (0.00)		
Total variance	$e_{0ijkl} + \mu_{0jkl} + \nu_{0kl} + f_{0l}$	0.31	0.31	0.31	0.31	0.31
Goodness of fit						
Deviance		3,013.35	2,987.23	2,987.23	2,987.22	2,985.00
Model of reference			Model 1	Model 2	Model 2	Model 4
Chi-square fit improvement			$\chi^2_{(1)} = 26.12$	$\chi^2_{(1)} = 0$	$\chi^2_{(1)} = 0.01$	$\chi^2_{(1)} = 2.22$
P-value			$p = < .001$	p = n.s.	p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is resilience, measured 3 times (repeated measures n = 2,085; student n = 1,046; course of study n = 13; faculty n = 2). n.s. = non-significant *p < .05 **p < .01 ***p < .001

Table B.10.2

Results Multilevel Repeated Measures Analyses 'Resilience' with Gender as Moderator

Effect	Parameter	Resilience		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.91 (0.04)	3.94 (0.05)	3.92 (0.05)
Trend	γ_{100}	-0.06*** (0.01)	-0.07*** (0.02)	-0.05** (0.02)
Intervention (= 1)	γ_{010}	-0.00 (0.03)	0.04 (0.06)	0.09 (0.06)
Male (= 1)	γ_{020}	-0.13*** (0.03)	0.01 (0.06)	0.06 (0.07)
Intervention*Trend	γ_{110}		-0.03 (0.02)	-0.06 (0.03)
Intervention*Male	γ_{030}		0.03 (0.06)	-0.07 (0.09)
Trend*Male	γ_{120}		0.06** (0.02)	0.03 (0.03)
Intervention*Trend*Male	γ_{140}			0.06 (0.04)
		Random effects		
Repeated measures variance	e_{0ijk}	0.14 (0.01)	0.14 (0.01)	0.14 (0.01)
Student variance	μ_{0jk}	0.15 (0.01)	0.15 (0.01)	0.15 (0.01)
Course variance	ν_{0k}	0.01 (0.00)	0.01 (0.01)	0.01 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.30	0.30	0.30
		Goodness of fit		
Deviance		2,970.45	2,958.82	2,956.72
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 11.63$	$\chi^2_{(1)} = 2.10$
P-value			$p = < .01$	$p =$ n.s.

Note. Idem to Table B.10.1

Table B.10.3

Results Multilevel Repeated Measures Analyses 'Resilience' with Domain as Moderator

Effect	Parameter	Resilience		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	4.02 (0.07)	3.91 (0.08)	3.91 (0.09)
Trend	γ_{100}	-0.06*** (0.01)	-0.01 (0.02)	-0.01 (0.03)
Intervention (= 1)	γ_{010}	-0.00 (0.03)	0.12 (0.07)	0.11 (0.09)
Teacher education (= 1)	γ_{020}	-0.06 (0.08)	0.07 (0.09)	0.06 (0.10)
Intervention*Trend	γ_{110}		-0.03 (0.02)	-0.02 (0.04)
Intervention*Teacher ed.	γ_{030}		-0.10 (0.07)	-0.09 (0.11)
Trend*Teacher ed.	γ_{120}		-0.05 (0.02)	-0.04 (0.03)
Intervention*Trend*Teacher ed.	γ_{140}			-0.01 (0.05)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.14 (0.01)	0.14 (0.01)	0.14 (0.01)
Student variance	μ_{0jk}	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)
Course variance	ν_{0k}	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.31	0.31	0.31
		Goodness of fit		
Deviance		2,986.61	2,978.03	2,978.00
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 8.58$	$\chi^2_{(1)} = 0.03$
P-value			$p = < .05$	$p =$ n.s.

Note. Idem to Table B.10.1

Table B.10.4

Results Multilevel Repeated Measures Analyses 'Resilience' with Ethnicity as Moderator

Effect	Parameter	Resilience		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.95 (0.04)	3.89 (0.05)	3.89 (0.05)
Trend	γ_{100}	-0.06*** (0.01)	-0.02 (0.02)	-0.02 (0.02)
Intervention (= 1)	γ_{010}	-0.00 (0.03)	0.03 (0.05)	0.04 (0.06)
Ethnic minority (= 1)	γ_{020}	0.08** (0.03)	0.19** (0.06)	0.20** (0.07)
Intervention*Trend	γ_{110}		-0.03 (0.02)	-0.03 (0.02)
Intervention*Ethnic min.	γ_{030}		0.05 (0.07)	0.03 (0.10)
Trend*Ethnic min.	γ_{120}		-0.08*** (0.02)	-0.08** (0.03)
Intervention*Trend*Ethnic min.	γ_{140}			0.01 (0.04)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.14 (0.01)	0.14 (0.01)	0.14 (0.01)
Student variance	μ_{0jk}	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)
Course variance	ν_{0k}	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.31	0.31	0.31
		Goodness of fit		
Deviance		2,981.79	2,967.66	2,967.60
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 14.13$	$\chi^2_{(1)} = 0.06$
P-value			p = < .001	p = n.s.

Note. Idem to Table B.10.1

Table B.11.1

Results Multilevel Repeated Measures Analyses 'Grit'

Effect	Parameter	Grit				
		Model 1	Model 2	Model 3	Model 4	Model 5
		Fixed effects				
Intercept	γ_{0000}	3.74 (0.02)	3.79 (0.04)	3.76 (0.03)	3.74 (0.04)	3.72 (0.04)
Trend	γ_{1000}	-0.13*** (0.01)	-0.12*** (0.01)	-0.13*** (0.01)	-0.13*** (0.01)	-0.12*** (0.01)
Intervention (= 1)	γ_{0100}				0.04 (0.03)	0.07 (0.05)
Intervention*Trend	γ_{1100}					-0.01 (0.02)
		Random effects				
Rep. meas. variance	e_{0ijkl}	0.12 (0.01)	0.12 (0.01)	0.12 (0.01)	0.12 (0.01)	0.12 (0.01)
Student variance	μ_{0jkl}	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)	0.15 (0.01)	0.15 (0.01)
Course variance	ν_{0kl}		0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
Faculty variance	f_{0k}			0.00 (0.00)		
Total variance	$e_{0ijk} + \mu_{0ij} + \nu_{0k} + f_{0k}$	0.28	0.29	0.29	0.28	0.28
		Goodness of fit				
Deviance		2,841.03	2,834.23	2,834.23	2,832.08	2,831.53
Model of reference			Model 1	Model 2	Model 2	Model 4
Chi-square fit improvement			$\chi^2_{(1)} = 6.80$	$\chi^2_{(1)} = 0$	$\chi^2_{(1)} = 2.15$	$\chi^2_{(1)} = 0.55$
P-value			$p = < .001$	$p =$ n.s.	$p =$ n.s.	$p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is grit, measured 3 times (repeated measures $n = 2{,}106$; student $n = 1{,}050$; course of study $n = 13$; faculty $n = 2$). n.s. = non-significant $*p < .05$ $**p < .01$ $***p < .001$

Table B.11.2

Results Multilevel Repeated Measures Analyses 'Grit' with Gender as Moderator

Effect	Parameter	Grit		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.80 (0.04)	3.82 (0.05)	3.81 (0.05)
Trend	γ_{100}	-0.13*** (0.01)	-0.14*** (0.02)	-0.13*** (0.02)
Intervention (= 1)	γ_{010}	0.04 (0.03)	0.05 (0.05)	0.07 (0.06)
Male (= 1)	γ_{020}	-0.14*** (0.03)	-0.21*** (0.05)	-0.19** (0.06)
Intervention*Trend	γ_{110}		-0.01 (0.02)	-0.02 (0.03)
Intervention*Male	γ_{030}		0.02 (0.06)	-0.02 (0.09)
Trend*Male	γ_{120}		0.04 (0.02)	0.02 (0.03)
Intervention*Trend*Male	γ_{140}			0.02 (0.04)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.12 (0.01)	0.12 (0.01)	0.12 (0.01)
Student variance	μ_{0jk}	0.15 (0.01)	0.15 (0.01)	0.15 (0.01)
Course variance	ν_{0k}	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.28	0.28	0.28
		Goodness of fit		
Deviance		2,811.86	2,807.91	2,807.56
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 3.95$	$\chi^2_{(1)} = 0.35$
P-value				
			p = n.s.	p = n.s.

Note. Idem to Table B.11.1

Table B.11.3

Results Multilevel Repeated Measures Analyses 'Grit' with Domain as Moderator

Effect	Parameter	Grit		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.66 (0.07)	3.61 (0.08)	3.53 (0.08)
Trend	γ_{100}	-0.13*** (0.01)	-0.10*** (0.02)	-0.06** (0.02)
Intervention (= 1)	γ_{010}	0.04 (0.03)	0.08 (0.07)	0.24** (0.09)
Teacher education (= 1)	γ_{020}	0.09 (0.07)	0.15 (0.09)	0.26** (0.09)
Intervention*Trend	γ_{110}		-0.01 (0.02)	-0.10* (0.04)
Intervention*Teacher ed.	γ_{030}		-0.02 (0.07)	-0.24** (0.10)
Trend*Teacher ed.	γ_{120}		-0.03 (0.02)	-0.09** (0.03)
Intervention*Trend*Teacher	γ_{140}			0.12** (0.04)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.12 (0.01)	0.12 (0.01)	0.12 (0.01)
Student variance	μ_{0jk}	0.15 (0.01)	0.15 (0.01)	0.15 (0.01)
Course variance	ν_{0k}	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.28	0.28	0.28
		Goodness of fit		
Deviance		2,830.48	2,828.01	2,819.83
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 2.47$	$\chi^2_{(1)} = 8.18$
P-value			p = n.s.	$p = < .01$

Note. Idem to Table B.11.1

Table B.11.4

Results Multilevel Repeated Measures Analyses 'Grit' with Ethnicity as Moderator

Effect	Parameter	Grit		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.71 (0.04)	3.66 (0.04)	3.65 (0.04)
Trend	γ_{100}	-0.13*** (0.01)	-0.11*** (0.02)	-0.10*** (0.02)
Intervention (= 1)	γ_{010}	0.04 (0.03)	0.08 (0.05)	0.12** (0.05)
Ethnic minority (= 1)	γ_{020}	0.10*** (0.03)	0.20*** (0.06)	0.27*** (0.07)
Intervention*Trend	γ_{110}		-0.01 (0.02)	-0.03 (0.02)
Intervention*Ethnic min.	γ_{030}		-0.06 (0.06)	-0.18 (0.10)
Trend*Ethnic min.	γ_{120}		-0.04 (0.02)	-0.08 (0.03)
Intervention*Trend*Ethnic min.	γ_{140}			0.07 (0.04)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.12 (0.01)	0.12 (0.01)	0.12 (0.01)
Student variance	μ_{0jk}	0.15 (0.01)	0.15 (0.01)	0.15 (0.01)
Course variance	ν_{0k}	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.28	0.28	0.28
		Goodness of fit		
Deviance		2,822.64	2,817.78	2,815.02
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 4.86$	$\chi^2_{(1)} = 2.76$
P-value			p = n.s.	p = n.s.

Note. Idem to Table B.11.1

Table B.12.1

Results Multilevel Repeated Measures Analyses 'Engagement'

Effect	Parameters of Model 4 and 5	Engagement				
		Model 1	Model 2	Model 3	Model 4	Model 5
Fixed effects						
Intercept	γ_{000}	3.42 (0.03)	3.47 (0.05)	3.45 (0.07)	3.45 (0.05)	3.45 (0.06)
Trend	γ_{100}	-0.11*** (0.01)	-0.11*** (0.01)	-0.11*** (0.01)	-0.11*** (0.01)	-0.11*** (0.02)
Intervention (= 1)	γ_{010}				0.04 (0.04)	0.05 (0.06)
Intervention*Trend	γ_{110}					-0.00 (0.03)
Random effects						
Rep. meas. variance	e_{0ijk}	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)
Student variance	μ_{0jk}	0.21 (0.02)	0.20 (0.01)	0.20 (0.02)	0.20 (0.02)	0.20 (0.02)
Course variance	ν_{0k}		0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Faculty variance				0.00 (0.01)		
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.46	0.47	0.47	0.47	0.47
Goodness of fit						
Deviance		4,026.88	4,000.95	4,000.95	3,999.63	3,999.62
Model of reference			Model 1	Model 2	Model 2	Model 4
Chi-square fit improvement			$\chi^2_{(1)} = 25.93$	$\chi^2_{(1)} = 0$	$\chi^2_{(1)} = 1.32$	$\chi^2_{(1)} = 0.01$
P-value			p = < .001	p = n.s.	p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is engagement, measured 3 times (repeated measures n = 2,085; student n = 1,046; course of study n = 13; faculty n = 2). n.s. = non-significant *p < .05 **p < .01 ***p <.001

Table B.12.2

Results Multilevel Repeated Measures Analyses 'Engagement' with Gender as Moderator

Effect	Parameter	Engagement		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.50 (0.06)	3.52 (0.07)	3.56 (0.07)
Trend	γ_{100}	-0.11*** (0.01)	-0.11*** (0.02)	-0.13*** (0.03)
Intervention (= 1)	γ_{010}	0.04 (0.04)	0.02 (0.07)	-0.05 (0.08)
Male (= 1)	γ_{020}	-0.11** (0.04)	-0.15 (0.07)	-0.23** (0.09)
Intervention*Trend	γ_{110}		-0.00 (0.03)	0.04 (0.04)
Intervention*Male	γ_{030}		0.06 (0.07)	0.20 (0.12)
Trend*Male	γ_{120}		0.01 (0.03)	0.05 (0.04)
Intervention*Trend*Male	γ_{140}			-0.08 (0.05)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)
Student variance	μ_{0jk}	0.19 (0.02)	0.19 (0.02)	0.19 (0.02)
Course variance	ν_{0k}	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.46	0.46	0.46
		Goodness of fit		
Deviance		3,991.32	3,990.61	3,988.31
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 0.71$	$\chi^2_{(1)} = 2.30$
P-value			p = n.s.	p = n.s.

Note. Idem to Table B.12.1

Table B.12.3

Results Multilevel Repeated Measures Analyses 'Engagement' with Domain as Moderator

Effect	Parameter	Engagement		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.27 (0.10)	3.27 (0.11)	3.19 (0.12)
Trend	γ_{100}	-0.11*** (0.01)	-0.12*** (0.03)	-0.08** (0.04)
Intervention (= 1)	γ_{010}	0.04 (0.04)	0.13 (0.09)	0.29* (0.12)
Teacher education (= 1)	γ_{020}	0.21* (0.11)	0.22 (0.12)	0.32** (0.13)
Intervention*Trend	γ_{110}		-0.01 (0.03)	-0.09 (0.05)
Intervention*Teacher ed.	γ_{030}		-0.11 (0.08)	-0.33 (0.14)
Trend*Teacher ed.	γ_{120}		0.02 (0.03)	-0.03 (0.04)
Intervention*Trend*Teacher ed.	γ_{140}			0.12* (0.06)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)
Student variance	μ_{0jk}	0.19 (0.02)	0.19 (0.02)	0.19 (0.02)
Course variance	ν_{0k}	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.45	0.45	0.45
		Goodness of fit		
Deviance		3,995.93	3,993.45	3,989.52
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 2.48$	$\chi^2_{(1)} = 3.93$
P-value			p = n.s.	$p = < .05$

Note. Idem to Table B.12.1

Table B.12.4

Results Multilevel Repeated Measures Analyses 'Engagement' with Ethnicity as Moderator

Effect	Parameter	Engagement		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	3.39 (0.05)	3.37 (0.06)	3.35 (0.06)
Trend	γ_{100}	-0.11*** (0.01)	-0.09*** (0.02)	-0.08*** (0.02)
Intervention (= 1)	γ_{010}	0.04 (0.04)	0.03 (0.06)	0.07 (0.07)
Ethnic minority (= 1)	γ_{020}	0.20*** (0.04)	0.29*** (0.08)	0.36*** (0.10)
Intervention*Trend	γ_{110}		-0.00 (0.03)	-0.03 (0.03)
Intervention*Ethnic min.	γ_{030}		-0.11 (0.08)	-0.10 (0.13)
Trend*Ethnic min.	γ_{120}		-0.06 (0.03)	-0.10 (0.04)
Intervention*Trend*Ethnic min.	γ_{140}			0.08 (0.06)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)
Student variance	μ_{0jk}	0.19 (0.02)	0.19 (0.02)	0.19 (0.02)
Course variance	ν_{0k}	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.46	0.46	0.45
		Goodness of fit		
Deviance		3,975.45	3,970.86	3,969.00
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 4.59$	$\chi^2_{(1)} = 1.86$
P-value			p = n.s.	p = n.s.

Note. Idem to Table B.12.1

Table B.13.1

Results Multilevel Repeated Measures Analyses 'Well-being'

Effect	Parameters of Model 4 and 5	Well-being				
		Model 1	Model 2	Model 3	Model 4	Model 5
		Fixed effects				
Intercept	γ_{000}	4.80 (0.04)	4.79 (0.07)	4.79 (0.06)	4.79 (0.07)	4.78 (0.07)
Trend	γ_{100}	-0.29*** (0.02)	-0.29*** (0.02)	-0.29*** (0.02)	-0.29*** (0.02)	-0.28*** (0.02)
Intervention (= 1)	γ_{010}				-0.01 (0.05)	0.03 (0.07)
Intervention*Trend	γ_{110}					-0.02 (0.03)
		Random effects				
Rep. meas. variance	e_{0ijk}	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)
Student variance	μ_{0jk}	0.37 (0.03)	0.33 (0.03)	0.33 (0.03)	0.33 (0.03)	0.33 (0.03)
Course variance	ν_{0k}		0.03 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
Faculty variance				0.00 (0.00)		
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.73	0.72	0.72	0.73	0.73
		Goodness of fit				
Deviance		4,901.86	4,855.05	4,855.05	4,855.03	4,854.69
Model of reference			Model 1	Model 2	Model 2	Model 4
Chi-square fit improvement			$\chi^2_{(1)} = 46.81$	$\chi^2_{(1)} = 0$	$\chi^2_{(1)} = 0.02$	$\chi^2_{(1)} = 0.34$
P-value			$p < .001$	p = n.s.	p = n.s.	p = n.s.

Note. Standard errors are in parentheses. Dependent variable is general psychological well-being, measured 3 times (repeated measures $n = 2{,}075$; student $n = 1{,}046$; course of study $n = 13$; faculty $n = 2$). n.s. = non-significant $^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table B.13.2

Results Multilevel Repeated Measures Analyses 'Well-being' with Gender as Moderator

Effect	Parameter	Well-being		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	4.60 (0.06)	4.67 (0.08)	4.65 (0.08)
Trend	γ_{100}	-0.29*** (0.02)	-0.31*** (0.03)	-0.30*** (0.03)
Intervention (= 1)	γ_{010}	0.00 (0.04)	0.00 (0.08)	0.05 (0.10)
Male (= 1)	γ_{020}	0.42*** (0.05)	0.23** (0.09)	0.28** (0.11)
Intervention*Trend	γ_{110}		-0.02 (0.03)	-0.05 (0.04)
Intervention*Male	γ_{030}		0.09 (0.09)	-0.02 (0.15)
Trend*Male	γ_{120}		0.08** (0.03)	-0.10 (0.04)
Intervention*Trend*Male	γ_{140}			0.06 (0.07)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)
Student variance	μ_{0jk}	0.30 (0.02)	0.30 (0.02)	0.30 (0.02)
Course variance	ν_{0k}	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.68	0.68	0.68
		Goodness of fit		
Deviance		4,779.86	4,772.29	4,771.39
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 7.57$	$\chi^2_{(1)} = 0.90$
P-value			$p = < .05$	$p =$ n.s.

Note. Idem to Table B.13.1

Table B.13.3

Results Multilevel Repeated Measures Analyses 'Well-being' with Domain as Moderator

Effect	Parameter	Well-being		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	4.83 (0.13)	4.76 (0.15)	4.75 (0.15)
Trend	γ_{100}	-0.29*** (0.02)	-0.25*** (0.03)	-0.24*** (0.04)
Intervention (= 1)	γ_{010}	0.01 (0.05)	0.01 (0.11)	0.04 (0.14)
Teacher ed. (= 1)	γ_{020}	-0.05 (0.14)	0.03 (0.16)	0.05 (0.17)
Intervention*Trend	γ_{110}		-0.02 (0.03)	-0.03 (0.06)
Intervention*Teacher ed.	γ_{030}		0.09 (0.09)	-0.02 (0.17)
Trend*Teacher ed.	γ_{120}		-0.05 (0.04)	-0.06 (0.05)
Intervention*Trend*Teacher ed.	γ_{140}			0.02 (0.07)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)
Student variance	μ_{0jk}	0.33 (0.03)	0.33 (0.03)	0.33 (0.03)
Course variance	ν_{0k}	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.72	0.72	0.72
		Goodness of fit		
Deviance		4,854.92	4,853.08	4,852.99
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 1.84$	$\chi^2_{(1)} = 0.09$
P-value			p = n.s.	p = n.s.

Note. Idem to Table B.13.1

Table B.13.4

Results Multilevel Repeated Measures Analyses 'Well-being' with Ethnicity as Moderator

Effect	Parameter	Well-being		
		Model 1	Model 2	Model 3
		Fixed effects		
Intercept	γ_{000}	4.82 (0.07)	4.77 (0.10)	4.73 (0.08)
Trend	γ_{100}	-0.29*** (0.02)	-0.26*** (0.03)	-0.23*** (0.03)
Intervention (= 1)	γ_{010}	-0.01 (0.05)	-0.00 (0.10)	0.08 (0.09)
Ethnic minority (= 1)	γ_{020}	-0.08 (0.05)	0.03 (0.10)	0.19 (0.12)
Intervention*Trend	γ_{110}		-0.02 (0.03)	-0.06 (0.04)
Intervention*Ethnic minority	γ_{030}		0.10* (0.04)	-0.21 (0.16)
Trend*Ethnic minority	γ_{120}		-0.09* (0.04)	-0.18 (0.05)
Intervention*Trend*Ethnic minority	γ_{140}			0.17** (0.07)
		Random effects		
Rep. meas. variance	e_{0ijk}	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)
Student variance	μ_{0jk}	0.33 (0.03)	0.33 (0.03)	0.34 (0.03)
Course variance	ν_{0k}	0.03 (0.02)	0.03 (0.01)	0.03 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.72	0.72	0.73
		Goodness of fit		
Deviance		4,852.72	4,845.34	4,839.76
Model of reference			Model 1	Model 2
Chi-square fit improvement			$\chi^2_{(3)} = 7.38$	$\chi^2_{(1)} = 5.58$
P-value			p = n.s.	$p = < .05$

Note. Idem to Table B.13.1

Nederlandstalige Samenvatting

Hoger onderwijs is een grotendeels publiek goed waar alle burgers aan bijdragen. Vanuit de overheid wordt daarom zorgvuldig gemonitord of het hoger onderwijsinstellingen lukt zoveel mogelijk studenten binnen de geplande tijd te diplomeren. Vanuit het perspectief van docenten ligt het meer voor de hand om aandacht te besteden aan wat de studenten leren. Vanuit het perspectief van studenten is de studie vaak van groot belang, maar niet het enige belangrijke levensdomein tijdens de studietijd. Deze verschillende perspectieven zijn alle drie relevant en beïnvloeden elkaar. Een student kan bijvoorbeeld door een bijbaan of ziekte niet aan studeren toekomen en vertragen.

Dit proefschrift stelt voor dat onderwijsonderzoekers interventies onderzoeken en uitkomstmaten gebruiken die deze perspectieven combineren. Dit omvattende perspectief kan ook wel 'academic thriving' of 'academisch floreren' worden genoemd. Het staat voor de mate waarin het studenten tijdens de studietijd zowel in de studie als in andere levensdomeinen goed vergaat (bijv. een bijbaan, gezondheid, familie).

Hoofdstuk 1 van het proefschrift beschrijft hoe onderzoek naar academisch floreren bij kan dragen aan 'evidence-based (EB) onderwijs'. EB onderwijs staat voor de mate waarin onderwijzers op zoek gaan naar en correct gebruik maken van het beste, beschikbare, wetenschappelijk bewijs in hun onderwijscontext en de mate waarin relevant nieuw wetenschappelijk bewijs wordt gegenereerd ter verbetering van onderwijs. EB onderwijs is in opmars en speelt een steeds belangrijkere rol in onderwijsbeleid, maar roept ook kritische reacties op. Dit hoofdstuk categoriseert en evalueert de bezwaren tegen EB onderwijs.

De eerste categorie kritiek op EB onderwijs is gericht op de onderliggende aannames over kennis. Zo uiten verschillende wetenschappers kritiek op de veronderstelling dat je met een gerandomiseerd experiment (RCT) aan kan tonen dat iets 'werkt' en dat het daarom ook in andere contexten zal werken. Lokale factoren en de manier waarop een interventie wordt uitgevoerd en ontvangen, kunnen een substantiële rol spelen en moeten ook bestudeerd worden. Door het mechanisme waarmee een interventie werkt te onderzoeken, kan bovendien geleerd worden wat er voor nodig is om de betreffende interventie in verschillende contexten werkend te krijgen. Meer fundamentele kritiek gaat voorbij aan het feit dat sociale wetenschap probabilistisch is, RCT's bieden uiteindelijk de minst slechte manier om te schatten wat iets veroorzaakt.

De tweede categorie kritiek is economisch van aard. Het uitvoeren van grootschalige RCT's is kostbaar en leidt in 40% van de gevallen tot de conclusie dat op basis van dit betreffend experiment niet vast te stellen is of het wel of niet werkt. Door betere monitoring van de implementatie en meer grootschalige experimenten kunnen deze problemen worden verholpen. Het is alleen niet aantrekkelijk voor een schoolleider om te investeren in een experiment waaruit mogelijk blijkt dat de geïnvesteerde tijd en moeite een negatief of geen significant effect opleverden. Dit maakt het nodig effectonderzoek met overheidsbeleid te ondersteunen.

De derde categorie kritiek is normatief van aard. EB onderwijs zou volgens sommigen voorbijgaan aan de normativiteit van onderwijs. Leren is altijd gericht op een bepaald doel. Dit is een terecht punt. Onderwijsonderzoekers moeten, net als leraren, zich altijd bewust zijn van hun doelen en hier zo transparant mogelijk over zijn. Voor zover dit nog niet genoeg gebeurt, zou dit meer centraal moeten komen te staan in EB onderwijs. Dat leren gericht kan zijn op verschillende doelen betekent echter niet dat het niet te onderzoeken valt. Zodra er enigszins consensus is over wat van belang is, is dit in een leerdoel of in een construct te operationaliseren. Dat kan wiskunde zijn, maar ook welzijn.

Samenvattend levert de kritiek op EB onderwijs geen voldoende rechtvaardiging om het af te wijzen. De kritiek bevat echter wel aanbevelingen die onderwijsonderzoekers ter harte kunnen nemen. RCT's moeten methodisch van hoog niveau zijn om een zinvolle schatting van causale verbanden te kunnen bieden. De uitvoering en lokale contextfactoren moeten zorgvuldig worden gemonitord en aanvullende (kwalitatieve) studies naar de mogelijke mechanismes zijn van grote waarde. Onderwijs kan een complexe context zijn voor onderzoek omdat er sprake is van veel onvoorspelbare factoren. Het is echter ook een context met specifieke voordelen en kansen voor onderzoek.

Onderwijsonderzoekers zouden meer gebruik mogen maken van de longitudinale data en evaluatiedata die in potentie beschikbaar zijn. Onderwijsinstellingen toetsen de voortgang van studenten en deze data is steeds meer digitaal beschikbaar. In dit licht bezien zou het van grote waarde zijn als de vele manieren waarop onderwijs geëvalueerd wordt meer voldoen aan wetenschappelijke criteria, zodat ze wetenschappelijk meer kunnen worden benut. Ten slotte interacteren onderwijsdoelstellingen met andere levensdomeinen. Voor studiegerelateerde problemen die verweven zijn met andere levensdomeinen kan meer gebruik gemaakt worden van multidisciplinaire interventies en gecombineerde uitkomstmaten.

Hoofdstuk 2 presenteert de resultaten van een grootschalig veldexperiment (RCT) dat de effecten van een goal-setting opdracht op het welzijn en de prestaties van studenten van de lerarenopleidingen, pabo en ondernemerschap in kaart bracht. De meeste studies naar het effect van goal setting rapporteerden dat studenten die de opdrachten kregen er zowel qua prestaties als welzijn op vooruit gingen maar een studie vond geen enkel effect. Het veldexperiment in dit hoofdstuk moest meer uitsluitsel bieden door zowel de effecten als het mechanisme en de implementatie grondig te onderzoeken. De onderzochte goal-setting opdracht bestond uit twee delen. In het eerste deel beschreven de studenten met reflectieve schrijfoefeningen de verschillende domeinen van hun gewenste toekomstige leven. In het tweede deel moesten ze op basis hiervan doelen formuleren en prioriteren, en per doel subdoelen en een strategie uitwerken. De studenten die de interventie ontvingen behaalden significant meer studiepunten dan studenten die een controleopdracht kregen en vielen 6 procentpunt minder vaak uit. De twee groepen studenten scoorden niet verschillend op welzijn. De interventie leek dit niet te verbeteren maar ook niet te schaden. De gevonden effecten golden onafhankelijk van domein (onderwijs of economie), geslacht, etniciteit of vooropleiding. Met behulp van herhaalde metingen werd onderzocht of de effecten gemedieerd werden door zelfregulerend leren, weerbaarheid, volharding en engagement. De experiment- en controlegroep verschilden op geen van de onderzochte constructen. De manier waarop een interventie wordt uitgevoerd en ontvangen in een context kan van belang zijn in onderwijsonderzoek. Deze studie monitorde daarom op verschillende manieren hoe dit ging. De studenten schreven substantieel minder woorden dan in studies die grotere effecten rapporteerden en deden niet allemaal serieus mee aan de opdrachten. Hierdoor ligt het in de verwachting dat nog meer effect kan hebben wanneer met een hogere mate van succesvolle implementatie. Het is goed mogelijk dat verschillen in implementatie de tegenstrijdige resultaten van eerdere studies kunnen verklaren. De interventie bereikte ondanks de grootschalige en niet perfecte toepassing een relatief groot effect. Bovendien zijn de opdrachten goedkoop en schaalbaar. Het is daarom voor vergelijkbare opleidingen aan te raden deze opdrachten (zorgvuldig) in het curriculum te integreren.

Hoofdstuk 3 biedt een conceptueel vervolg op het experiment in hoofdstuk 2. Uit focusgroepen met studenten waarmee het experiment uit hoofdstuk 2 werd geëvalueerd kwam naar voren dat de studenten behoefte hadden aan persoonlijke follow-up. Deze follow-up zou multidisciplinair van aard moeten zijn om in te kunnen gaan op de verschillende en met elkaar verbonden doelen en behoeften die studenten formuleren: van studievaardigheden tot therapie. Dit hoofdstuk vult de bevindingen op het vlak van goal-setting interventies aan met bevindingen uit twee andere wetenschappelijke domeinen die een passende vorm van follow-up kunnen bieden. Ontwikkelingen op het vlak van artificiële intelligentie hebben geleid tot chatbots die persoonlijke gesprekken kunnen voeren en door kunnen vragen. Binnen de psychologie en psychiatrie zijn er steeds effectievere online vormen van therapie ontwikkeld. Recente meta-analyses laten zien dat de cognitieve gedragstherapie die deze interventies bieden, al net zo goed werkt als hulp van een professional. De online therapie en chatbots tonen aan persoonlijk contact te kunnen bieden en effectief te helpen bij lichte vormen van depressie en faalangst. Therapie heeft echter een stigma en bereikt veel studenten niet of te laat. AI-gedreven chatbots zijn tot nu toe met dit doel alleen door middel van kleine steekproeven getest. Dit hoofdstuk stelt daarom voor dat onderzoekers combinaties van deze interventies testen en onderzoeken. Een goal-setting interventie kan een grote groep studenten via het curriculum preventief bereiken zonder stigma. Door deze interventie te integreren in een chatbot kan de chatbot studenten indien nodig stimuleren meer te schrijven. Dezelfde chatbot kan vervolgens persoonlijke follow-up bieden. Door zowel naar de doelstellingen te vragen als naar hoe het gaat, kan de chatbot laagdrempelig cognitieve gedragstherapie of andere ondersteuning op maat bieden.

Hoofdstuk 4 gaat over de relatie tussen studieprestaties en werken tijdens de studie. De overgrote meerderheid van de voltijdstudenten heeft tegenwoordig een bijbaan wanneer ze aan de studie beginnen. Op de lerarenopleidingen krijgen bovendien veel studenten nog tijdens de studie een betaalde baan voor de klas aangeboden. In de literatuur is er onenigheid over of bijbanen overwegend slecht zijn omdat ze concurreren met studietijd, of dat ze overwegend positief zijn vanwege de vaardigheden die je er door opdoet. Tot nog toe werd bij het onderzoeken van de effecten van werk op studievoortgang in het onderwijsdomein geen onderscheid gemaakt tussen stages, betaalde banen buiten en betaalde banen binnen het onderwijs. Bovendien is onbekend of de invloed van werken naast de studie gedurende de vier jaar van de studie verandert. Deze studie onderzocht het effect op studievoortgang van zowel onbetaalde stage-overuren, als betaalde banen binnen en buiten het onderwijs. Dit deden we met behulp van multilevel groeimodellen waarin studievoortgang van 132 aankomende leraren op 25 meetmomenten over vier jaar werd gekoppeld aan de hoeveelheid tijd die ze gemiddeld per week aan verschillende soorten werk besteedden. Vanaf het derde jaar verruilden de meeste studenten hun baan buiten het onderwijs voor een binnen het onderwijs. De meeste studenten besteedden meer tijd aan hun stages dan volgens de studie nodig was of waarvoor ze betaald werden, dit nam geleidelijk aan toe gedurende de opleiding. Betaald werk in het onderwijs gedurende het derde en vierde studiejaar was een significante voorspeller van groei in studiepunten. In het eerste en derde semester hing ongeveer een dag per week betaald werk buiten het onderwijs samen met de meeste studievoortgang, vanaf meer dan twee dagen per week werd dit verband negatief. Dit onderzoek falsifieert de aanname dat betaalde banen in het onderwijs tot studievertraging leiden. Betaalde, domeinrelevante banen lijken de studenten in staat te stellen andere bijbanen op te zeggen of af te bouwen. Indien men bang is dat betaald werk met studietijd concurreert, zou het daarom wellicht verstandiger zijn om de stagevergoeding te verhogen dan om betaald werk in het onderwijs te verbieden. Wetenschappelijk voegt deze studie twee variabelen toe aan verklarende modellen over bijbanen en studievoortgang: niet alleen het soort werk, maar ook of dit betaald wordt en de fase van de studie blijken van belang.

De verschillende hoofdstukken van dit proefschrift zijn los te lezen maar bouwen ook op elkaar voort. Hoofdstuk 1 en 3 beschrijven hoe cijfermatig studiesucces samenhangt met andere relevante factoren en stellen manieren voor waarop dit kan worden onderzocht. Hoofdstukken 2 en 4 tonen de resultaten van onderzoek waarin deze principes werden toegepast. Ik hoop dat zowel de inhoudelijke inzichten als de methodische aanpak een bijdrage leveren aan de discussie over studiesucces en uitval in het hoger onderwijs. Uiteindelijk zijn zowel de perspectieven van studenten, docenten en beleidsmakers nodig om studenten academisch tot bloei te laten komen.

About the author

Izaak Dekker was born on December 3, 1986, on a houseboat in the centre of Amsterdam. After finishing high school he moved to Rotterdam to study Philosophy and Economics at the Erasmus University Rotterdam. Izaak first combined his study with a job as editorial assistant at Economische Statistische Berichten and later with jobs as a philosophy teacher at primary schools in Rotterdam and Amsterdam. Upon completion of his Master degree in Philosophy, Izaak accepted a job as a lecturer at the Rotterdam University of Applied Sciences (RUAS) and a job as philosophy teacher at a school for secondary education. During his first years as a teacher, and after obtaining a postgraduate diploma in didactics at the Vrije Universiteit Amsterdam, Izaak became aware of the critical importance of professional development of teachers in higher education. In 2013 he became strategic advisor to the board of the RUAS, a position he combined with teaching. Within 2 years he became one of the senior advisors and headed a program targeted at improving academic performance of the RUAS students. Together with the chairman he published an essay in which they underscored the importance of a pedagogical and didactical context that allows students to succeed. In line with this focus, the program Izaak led involved projects that renewed the strategic HRM and HRD policies of the university, resulting in new recruitment, induction and development policies. During his time as a board advisor, Izaak noticed how few of the policies or interventions in higher education were thoroughly evaluated. This induced him to write a PhD proposal aimed at studying effects of a promising goal-setting intervention on academic achievement and student well-being. In September 2018 Izaak started, alongside his job as teacher for the Master Programme 'Learning and innovation', the part-time PhD programme of the Rotterdam School of Management, funded by a promotion voucher from RUAS.

Author's Portfolio

Peer-reviewed publications

Rozendaal J. & Dekker, I. (2021). Leren onderzoeken aan de MLI. *Tijdschrift voor Lerarenopleiders, 42*(1).

Dekker I., De Jong E. M., Schippers M. C., De Bruijn-Smolders M., Alexiou A. & Giesbers B. (2020). Optimizing students' mental health and academic performance: AI-enhanced life crafting. *Frontiers in Psychology*, *11*:1063. https://doi.org/10.3389/fpsyg.2020.01063

Working papers

Dekker, I., Chong, C.F., Schippers, M. C. & Van Schooten, E. J. (2021). The right job pays: Effects of work on the study progress of pre-service teachers. *SSRN*, http://dx.doi.org/10.2139/ssrn.3848452

Dekker, I., Schippers, M. C., & Van Schooten, E. J. (2021). Reflective goal-setting improves academic performance in teacher and business education: A large-scale field experiment. *SSRN*, http://dx.doi.org/10.2139/ssrn.3778544

Peer-reviewed conference paper presentations

Dekker, I., Chong, C.F., Schippers, M. C. & Van Schooten, E. J. (2021, September 9-11). *The right job pays: Effects of work on the study progress of pre-service teachers* [Paper presentation]. Association for Teacher Education in Europe Annual Conference, Online.

Dekker, I., Schippers, M. C., & Van Schooten, E. J. (2021, August 23-27). *Reflective goal-setting improves academic performance in teacher and business education: A large-scale field experiment* [Paper presentation]. European Association for Research on Learning and Instruction Annual Conference, Online.

Dekker I., De Jong E. M., Schippers M. C., De Bruijn-Smolders M., Alexiou A. & Giesbers B. (2021, August 23-27). *Optimizing students' mental health and academic performance: AI-enhanced life crafting* [Paper presentation]. European Association for Research on Learning and Instruction annual conference, Online.

Book reviews

Dekker, I. (2021). Het minst onbetrouwbaar [Review of the book *Taming RCTs in education*, by K. Morrison]. *Thema Hoger* Onderwijs, *28*(3), Instondo.

Dekker, I. (2021). Ministerschap met beperkingen [Review of the book *Ministerie van verbeelding*, by J. Bussemaker]. *Thema Hoger* Onderwijs, *28*(2), Instondo.

Dekker, I. (2021). Beter leren kiezen [Review of the book *Een leven lang kiezen*, by E. Meens]. *Thema Hoger Onderwijs. 28*(1), Instondo.

Dekker, I. (2020). De nieuwe student [Book review of *De gelukkigste tijd van je leven?*, by T. Andreoli]. *Thema Hoger Onderwijs. 27*(5), Instondo.

Dekker, I. (2020). Slimmer duiden [Book review of *Statistisch handboek studiedata*, by I. Van der Staaij et al.]. *Thema Hoger Onderwijs. 27*(2), Instondo.

Dekker, I. (2019). Teleurstellende toekomst voorspellingen [Book review of *Possible selves and higher education*, by H. Henderson et al.]. Boekbespreking. *Thema Hoger Onderwijs. 26*(3), Instondo.

Dekker, I. (2019). De programmaficatie van onderwijsonderzoek [Book review of *European educational research (re)constructed*, by M. Zapp et al.]. *Thema Hoger Onderwijs. 26*(2), Instondo.

Dekker, I. (2019). Het nut van nadenken over de toekomst [Book review of *The psychology of thinking about the future*, by G. Oettingen et al.]. *Thema Hoger Onderwijs. 26*(1), Instondo

Dekker, I. (2018). Geen derde weg, wel een interessant pleidooi [Book review of *De terugkeer van het lesgeven*, by G. Biesta]. *Thema Hoger Onderwijs*, *25*(5), Instondo.

Dekker, I. (2018). Mooie perspectieven [Book review of *Studiesucces in het hoger onderwijs*, by F. Glastra and D. Van Middelkoop]. *Thema Hoger Onderwijs. 25*(5), Instondo.

Dekker, I. (2018). Biesta voor bestuurders [Book review of *Educational goods*, by Brighouse et al.]. *Thema Hoger Onderwijs*, *25*(4), Instondo.

Dekker, I. (2018). Bijles als symptoom van segregatie [Book review of *De bijlesgeneratie*, by L. Elffers]. *Thema Hoger Onderwijs*, *25*(3), Instondo.

Dekker, I. (2017). Waarom wil je voor de klas? [Book review of *Global perspectives on teacher motivation*, by H. M. G. Watt et al.]. *Thema Hoger Onderwijs*, *24*(5), Instondo.

Dekker, I. (2017). Welke toekomst willen we? [Book review of *Haalt de universiteit 2040?*, by B. Van der Zwaan]. *Thema Hoger Onderwijs*, *24*(1), Instondo.

Dekker, I. (2016). De docent als waarheidszoekende vriend [Book review of *Streven naar beter: Nietzsche als gids voor het hbo*, by H. Joosten]. *Thema Hoger Onderwijs*. *23*(4). Instondo.

Essays

Dekker, I. & Adema, D. (2021). Deltaplan kansengelijkheid. *Didactiefonline*. https://didactiefonline.nl/blog/blonz/deltaplan-kansengelijkheid

Dekker, I. (2021). 'Groenpluk' blijkt mythe. *Didactiefonline*. https://didactiefonline.nl/artikel/groenpluk-blijkt-mythe

Dekker, I. (2018). Barbaren en sofisten: Een kritiek op een oppervlakkige toekomstvisie. *DeFusie*. http://defusie.net/barbaren-en-sofisten/

Dekker, I. & Vos, I. (2017). Gemeenschappelijke waar(he)den in tijden van verdeeldheid. *DeFusie*. http://defusie.net/terug-naar-toekomstige-liberaal/

Berding, J. & Dekker, I. (2017). *Onderwijsfilosofen in actie: Zeven lessen over visies en idealen in en voor de klas*. https://www.scienceguide.nl/wp-content/uploads/2017/11/onderwijsfilosofen_in_actie__voor_op_website_.pdf

Dekker, I. & Vos, I. (2017). Je hebt wel een opinie. *DeFusie*. http://defusie.net/hebt-wel-opinie/

Bormans, R. & Dekker. I. (2017). De grote opdracht voor het mbo. *Didactiefonline*. https://didactiefonline.nl/blog/blonz/de-grote-opdracht-voor-het-mbo

Dekker, I. & Vos, I. (2017). Op zoek naar discipline. *DeFusie*. http://defusie.net/op-zoek-naar-discipline/

Dekker, I. (2017). De teloorgang van het publieke belang. *Scienceguide*. http://www.scienceguide.nl/201707/de-teloorgang-van-het-publieke-belang.aspx

Bormans, R. & Dekker, I. (2016). *Samen leven in de moderne samenleving*. https://www.hogeschoolrotterdam.nl/globalassets/documenten/hogeschool/over-ons/essay-samen-leven-in-de-moderne-samenleving.pdf

Bajwa, M., Braam, E. van, Bormans, R. & Dekker, I. (2015). *Kwaliteit in de klas.* https://scienceguide.nl/media/1933220/_kwaliteit_in_de_klas__-_essay_ron_bormans__maaike_bajwa__erwin_van_braam_en_izaak_dekker_-_hogeschool_rotterdam.pdf

Dekker, I. (2014). De virtuoze leraar. *Scienceguide*. http://www.scienceguide.nl/201405/de-virtuoze-leraar.aspx

Acknowledgements

To my supervisors

With her enormous drive and creativity, Michaéla Schippers pushed me to rewrite drafts that I was initially content with and opened chances and options for fruitful innovative collaborations. Thank you Michaéla for your supervision! Ellen Klatter supported my proposal and research ambitions from the start. I am thankful for her ever so emphatic guidance and constructive feedback. I had little to no statistical background when I started writing my PhD proposal. Erik van Schooten took me under his wings. Erik recommended seminal works in methodology, and gave me many private lessons about CFA's, multilevel modelling, and the underlying methodological assumptions and principles. Not only did he kickstart my statistical development, but he also taught me about the importance of scientific rigor and integrity. I am greatly indebted to him and thoroughly enjoyed our collaboration. Together the three of you formed a great team: Michaéla's creativity and ambition pushed me to greater heights, Ellen supported me and offered constructive solutions, and Erik challenged me to uphold the highest standards of scientific rigor and supported my statistical learning process.

To my colleagues

Thank you Els de Bock, Erwin van Braam, Ron Bormans, Frans Spierings, and Annemarie van Efferink for their instrumental role in enabling this PhD project as administrators. Because of your unwavering support I was able to spend two days every week on this dissertation during the past years. Thank you Clementine Degener, Monique de Bruijn-Smolders and Bettina de Jong for being the professional and dedicated fellow researchers you are. Thank you Justin Slabbekoorn, Sanne Wielenga, Marika Kouvelis, Jeanine Hoogerland, Janine Slootmaekers, Arian van Staa, Anjali Raghunath, Jan Wouter van Cappelle, Mehmet Sahan, Emiel de Ruiter, Ruben Hoogendoorn, Suzanne van Dijk, Dian van Arkel, Annet Brugel, Emmie van den Boon, Annemiek Grootendorst, Fariel Soliana, Michiel Helsen, Nina Adriaanse, Rachid Jemaoui, Ishany Gaffar, Henriett Duijs, Marlies Ronner, Janneke Verloop, Tatjana van der Sluijs, Ellian van Strien, Niels Westerlaken, Juliette Hoogeveen, Anneke Delhaas, Jeroen Cok, Claudia Gomes, Bert van Wijk, Kees-Jan van Klaveren, Ahmet Olgun, Jeffry

Sleddens, Chanté Zuidgeest, Fleur Doolaard, Aad Schippers, August Papadopoulos en Lianne Verhoef for your contribution to the execution of the studies! Thank you to all the colleagues at the Research Centre for Urban Education. Thank you to all the other colleagues at the Rotterdam University of Applied Sciences who were involved with a part of this research or helped me to take my mind off of it every now and then.

To my fellow PhD candidates

The RSM part-time PhD program contains obligatory courses and a few electives that you follow together with a cohort of fellow part-time PhD candidates. From the start I felt like the odd-one out because my subject area is education instead of business administration. Luckily the fellow students and the staff endured me nonetheless and made it completely worthwhile all the way through advanced econometrics. A special thanks to Duncan Rooders for being my academic buddy, and all the fellow part-time PhD students for the support and PhD related memes.

Outside of the RSM confines I met some of my favourite new academic friends who shared a fanatical interest in educational research and offered excellent feedback on papers. During the only physical conference I attended I met the stellar Laurie Delnoij, thank you Laurie! Hannah Bijlsma, and Esther van Dijk, we met at the end of our PhD projects, thank you for offering feedback during the final phase.

To my friends

Thank you Duco Adema for pushing me to consider doing a PhD project and for the good talks and dinners during the journey. Thank you Thijs van Mens for the annoyingly useful advice. Thank you to the gentlemen of the IJshockeyvereniging Amstel Tijgers Hockey Club Amsterdam Night Owls for keeping me in shape. 'Mens sana in corpore sano': a healthy mind exists in a healthy body. Thank you Tina Rahimy, Marjolein van Oosterbos, en Thijs Schippers for your psychological guidance.

To my family

Dearest of all, thank you Violet for being the best wife ever. I could always talk to you about whatever exciting things I was working on or whatever frustration came along. Sometimes my obsession with research topics made me less attentive and present. You were understanding but also helped me to balance the different important domains of life, allowing me to thrive. Thank you Pina and Sal for being such good sleepers, and bringing such joy to our lives.

RSM PT PhD Series

1. Duijm, P. On the Cyclical Nature of Finance: The role and impact of financial institutions, Promotor(s): Prof. D. Schoenmaker & Prof. W.B. Wagner, 1, **https://repub.eur.nl/pub/120767**

2. Maas, S.A. In the moment of giving: Essays on contemporary forms of private and corporate philanthropy. Promotors: Prof. L.C.P.M. Meijs & Prof. J.P. Cornelissen. **https://repub.eur.nl/pub/124976**

3. Langenbusch, C. A lot to lose. Organizational identity and emotions in institutional contexts. Promotors: Prof. J.P. Cornelissen & Prof. G. Jacobs. **https://repub.eur.nl/pub/125099**

4. Zanten, J.A.P. van, Business in the Age of Sustainable Development, Promotor: Prof. R. van Tulder, Co-promotor: Dr. F. Wijen **https://repub.eur.nl/pub/135674**

5. Dekker, I., Academic Thriving: Optimising Student Development with Evidence-Based Higher Education. Promotor: Prof. M.C. Schippers **https://repub.eur.nl/pub/**

Academic thriving stands for a combination of academic outcomes as well as success in other relevant domains, such as well-being or finding the right job. What causes students to thrive academically? The studies in this dissertation contributed to this question with the use of experimental, interdisciplinary and longitudinal studies, and a critical theoretical examination of the arguments against evidence-based education.

Izaak Dekker (1986) studied philosophy before he started a career as a teacher in primary, secondary, and tertiary education. Izaak combined his teaching jobs with a job as a policy advisor for the Rotterdam University of Applied Sciences (RUAS). Izaak combined a job as teacher educator with his work on this PhD project, which was funded with a research grant from the RUAS.

This PhD thesis has sprung from the Part-time PhD Programme at the Rotterdam School of Management, Erasmus University (RSM). Part-time PhD candidates conduct research against the highest academic standards on topics with real-world application value, thereby contributing to the positive impact of RSM research on business and other societal stakeholders.

This programme allows candidates to develop their academic and research skills while they work. During the programme, candidates are trained in research methods, use RSM's research facilities and databases, participate in international conferences, and are supervised by research-active faculty.

RSM is one of Europe's top business schools with a strong reputation for academic research. It aims to develop business leaders immersed in international careers, who can become a force for positive change by carrying a critical, creative, caring, and collaborative mindset into a sustainable future.

RSM PhD Series
Research in Management

Rotterdam School of Management
Erasmus University
Mandeville (T) Building
Burgemeester Oudlaan 50
3062 PA Rotterdam
The Netherlands

P.O. Box 1738
3000 DR Rotterdam
The Netherlands
+ 31 10 408 1182
info@eur.nl
www.eur.nl