SRLA: Self-Regulated Learning Analytics

by

Heeryung Choi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2022

Doctoral Committee:

Assistant Professor Christopher Brooks, Chair
Assistant Professor Andrew Krumm
Research Professor Stephanie D. Teasley
Professor Philip H. Winne

Heeryung Choi
heeryung@umich.edu
ORCID iD: 0000-0001-8955-8905

© Heeryung Choi 2022

ACKNOWLEDGMENTS

I consider myself one of the luckiest PhD students as I have fully enjoyed the last six years of PhD journey with no regret. This was only possible because of support from friends, family, and faculty.

First of all, I have met a wonderful committee who deeply cares about my development as a scholar. I truly appreciate their time and effort. Stephanie Teasley is my role model as a female scholar. I always find something to learn from Phil Winne who is passionate, humorous, considerate, and inspiring – all at the same time. Andrew Krumm opened my eyes on how to better position myself as an interdisciplinary researcher, which is exciting but also challenging.

I cannot ask for a better advisor than Christopher Brooks. He is not only a great academic advisor but also an irreplaceable life mentor. A senior PhD student has once advised me to find an advisor with whom I can share difficulties as well as successes. I truly believe I have found the right advisor.

I also would like to thank my family. Clearly, a 13-hour difference did not stop them from having weekly calls to tease me for being 'such a graduate student.' Because of their firm support, I could pursue my dream for the last six years abroad.

I am grateful to Allan Martell, Allison Tyler, T. Charles Yun, Daphne Chang, Tawfiq Ammari, Lia Bozarth, Sam Carton, Soohye Yeom, Suin You, Sungjin Nam, and Yujin Lee (listed in an alphabetical order). Your support was crucial for me to navigate my everyday life in the US without feeling lost, detached, or isolated. I am also truly thankful to etc lab people. They were the best labmates I could ask for. They brightened so many of my days.

I appreciate Warren Li for all the goofy 'hap hap' time where we walked around downtown and had endless tea-drunk conversations. One of the most valuable things I have got out of this PhD program is the friendship with him. I also wish all the best for the rest of his journey as a PhD student.

Last but not least, I appreciate Ryan Burton for every single moment with him. Because of his endless love, support and patience, I could not only understand but also embrace who I am, and hence, I could become a better person. He is the biggest reason why I did not give up and tried to stay strong even when I was as vulnerable as a candle light facing the wind.

Thank you, all! I am finally done!

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

Learning analytics researchers have been diligently integrating trace data to study Self-Regulated Learning (SRL). Compared to traditionally used survey data, trace data, such as log or clickstream data designed and interpreted to understand a certain SRL construct, are considered to be more effective in capturing dynamic SRL as fine-grained events. Yet, trace data is not completely free from validity issues, since researchers' understanding of contexts and target constructs heavily affect the validity of design and interpretation of trace data [4, 102, 173]. Rather, researchers can adopt survey and trace data to complement each other [19, 148]. They can also compare survey and trace data on the same construct to deepen the understanding of what each type of data captures.

The aim of this dissertation is to understand the different nature of self-reported survey and trace data and to adopt context-specific indicators for particular SRL constructs. This dissertation is composed of three studies, each of which employed both survey and trace data to answer specific SRL-related questions. In Chapter 3, the first study demonstrates the importance of reflection in facilitating learning from hints through integrating complementary surveys and log data. The second study in Chapter 4 contrasts the alignment between achievement goal theory and trace data, and the misalignment between the theory and survey responses. The third study presented in Chapter 5 investigates causes of misalignments and builds theoretical interpretation from data.

This dissertation contributes to the field of learning analytics and SRL in multiple ways. First of all, the dissertation shows that effort in posing more valid methodological approaches could contribute to theories. In Chapters 4 and 5, I revealed the huge difference between learners' goal statement before learning and goal-relevant behaviors during learning, and questioned some of the previous findings that heavily relied upon survey data. Furthermore, it emphasizes the importance of thoroughly understanding targeted constructs in specific contexts. Only then, can we select a strongly valid methodological approach.

CHAPTER 1

Introduction

Measurement is one of the main components of learning analytics [149]. With continuous improvement in educational technology, learning analytics researchers have been pushing the methodological boundary by integrating more and more trace data into their studies. Trace data are clickstream or logs designed to capture a certain construct using devices that learners interact with during learning. The methodological integration is particularly meaningful to Self-Regulated Learning (SRL) researchers. This is not only due to general benefits of technology such as allowing researchers to easily collect large amount of data, but also because there are several criticisms about surveys.

For decades, knowledge of SRL constructs has been predominantly built with self-reported instruments, especially surveys [13, 23, 31, 42, 51, 59, 63, 75, 76, 98, 131, 142, 177, 185]. Multiple researchers have questioned the validity of using survey data in studying SRL and have raised a concern over detachment from dynamic SRL theories [153, 185]. A common practice of surveying learners to study SRL is prompting them with questions multiple times before, during, or after learning. This is not fine-grained enough to capture the dynamic nature of SRL constructs, which can frequently change depending on time and context [60, 170, 188]. Another limitation is that surveys heavily rely on various assumptions about learners that are often not guaranteed. For instance, validity of using survey data could be negatively affected if learners misinterpret survey questions or fail to respond to questions honestly and mindfully [6, 79, 164, 170]. Furthermore, many surveys ask learners to aggregate their experiences. These responses often do not accurately represent an actual SRL-relevant behavior during learning in a particular context [170].

Trace data, on the other hand, suffer less from potential theoretical detachment. Trace data are fine-grained enough to report learners' SRL as context-specific events. Furthermore, trace data allow automatic and less obtrusive data collection which does not require learners' attention. Due to these benefits, previous learning analytics work has widely applied trace data to measure several SRL constructs including the impact of scaffolds, achievement goals, and different SRL phases [19, 148, 153, 185]. Yet, trace data can also suffer from their own validity issues. Trace data is inherently equivocal; that is, a piece of log data could represent several constructs. Furthermore, the validity of trace data is affected by how researchers understand target constructs and translate

that understanding to educational system data design. If the researchers' understanding of the learning context or the trace data design process is flawed, it is likely that the data will fit the constructs poorly and, hence validity will be low.

Addressing the validity issues of self-reported data and trace data is especially critical to SRL researchers in the field of learning analytics. Learning analytics builds on traditional research methods, such as surveys, and seeks to enhance these with digital data such as trace data [57]. Understanding how to appropriately employ these methods together to advance learning-related knowledge is under one of the main interests of learning analytics researchers. Furthermore, SRL has been one of the core topics in learning analytics, which has expanded understanding of how to support achievement of learners across a wide range of proficiency [139]. Thus, SRL knowledge built on validly captured constructs could contribute to understanding and supporting learning experiences. As a learning analytics researcher studying SRL, I aim to answer the following question through this dissertation:

- How can the different nature of **self-reported surveys** and **trace data** be understood in order to adopt **more appropriate indicators** to capture specific **SRL constructs**?

While there are many ways to answer this question posed above regarding the validity issues in using SRL measurements, I have taken two major approaches in this dissertation. One approach is adopting survey and trace data together to measure different constructs and to complement each other [19, 122, 148]. By taking the benefits of each measure, researchers can enhance not only the validity of measurements but also the validity of their study findings. **This complement approach** would be a practical option for researchers whose main goal is to advance knowledge on learners' SRL. This approach is explored in Chapter 3 of this dissertation, where the the study integrated self-reported surveys and log data to measure the effect of reflection prompts on learning with hints. Specifically, self-reported surveys were used to measure learners' aggregated affect on overall learning experiences with the prompts, and log data were employed to capture learners' task submission behaviors.

Another approach is **the comparison approach** which compares survey and trace data on the same construct to find potential discrepancies between the two [64, 153, 185]. This approach is appropriate for researchers who aim to investigate when and why weak validity of using measurement is observed so that methodological characteristics can be considered. The studies in Chapters 4 and 5 applied this approach. In Chapter 4, I examined alignment between an achievement goal theory with survey and trace datasets using confirmatory factor analysis. In Chapter 5, the survey and trace datasets on achievement goals were investigated further in order to understand what information these datasets captured and how different they are from each other when building theory.

This dissertation contributes to the better methodological understanding of how to capture

SRL-relevant constructs, particularly in the context of studying effects of reflection processes and achievement goals. The dissertation also demonstrates that methodological understanding can contribute to theoretical understanding. **The complement approach** builds SRL knowledge on previous findings. In Chapter 3, I employed the complement approach to efficiently measure learners' affect and behaviors in a large-scale online classroom. Through the combination of surveys and traces, I revealed the positive delayed effects of reflection prompts on performance and perceived learning. The finding also suggested a more parsimonious intervention design to elicit reflection processes and help learners to improve learning gain.

On the other hand, **the comparison approach** questions current SRL research instruments and suggest novel ways to expand those. In Chapter 4, I found that trace datasets demonstrated good fits with the goal complex theory while survey datasets did not. This study result is particularly striking as achievement goals have been predominantly measured by surveys. Chapter 4 proposed the possibility of temporal goal changes of mastery learners in an authentic learning environment to avoid course failures.

In Chapter 5, which also posed **the comparison approach**, I found that learners' self-reported goals stated before learning did not translate into the goal-relevant behaviors during learning. This finding questions interpretations of previous work which has more strongly linked self-reported goal constructs with behaviors. Chapter 5 also suggested that instructors might want to keep their eyes on self-reported mastery learners as they could turn out to be less motivated or performance-oriented.

Taken together, this dissertation serves as a call to the learning analytics community to carefully understand how SRL constructs should be measured. Through more SRL investigations with complement and comparison approaches, learning analytics can benefit from both self-reported surveys and trace data to shed light on and build relevant theory.

CHAPTER 2

Self-regulated Learning and Measurement

2.1 Models and their implications for measurement designs

Self-regulated learning (SRL) is the process where a learner monitors and controls metacognition, cognition, motivation, affect, and contextual factors to achieve goals [22, 62, 126, 176, 187]. SRL includes numerous goal-oriented actions from building strategies to tackle a given task to evaluating learners' own learning experiences. For instance, learners with more productive self-regulation skills would carefully lay out their goals, build plans, and constantly self-monitor. Decade-long research in SRL has expanded understanding of how to support learners' achievement with a wide range of proficiency [139]. By understanding SRL models and each learner's different SRL skill level, a researcher can design a better intervention for learners with different goals.

There are six primary SRL models [123]. Zimmerman [187] proposes the Cyclical model that is composed of three phases: forethought, performance, and reflection. The model is considered dynamic as, for instance, learners start to self-regulate with goal setting and planning i.e., the forethought phase and then segue to the performance phase to implement the plans, followed by evaluating the degree of successful implementation and progress toward their goal, which is the reflection phase. Measures based on this model are those that can detect and captures that dynamic process. Another feature of the Cyclical model is a focus on how a person constructs, controls and adjusts to environments. Because of the importance of the environment for learners and of their learning behavior to control environments in the model, the Cyclical model could be useful in studying the role of learning context in SRL research. The emphasis on learners' environments would not be limited to physical surroundings such as teachers' presence but also stressors like poverty and stereotype threat [159].

Pintrich's model [126] is also considered to be dynamic. It is composed of four phases: (a) forethought, planning, and activation, (b) monitoring, (c) control, and (d) reaction and reflection [123]. The model particularly focuses on both motivational and cognitive components of an SRL, primarily mastery and performance achievement goal orientations [126]. While mastery goals refer

to efforts to increase learners' own competence and understanding, performance goals are focused on learners' assessment of their own ability and self-worth such as besting other people, receiving public recognition, and avoiding seeking help in order not to appear incompetent [96].

As with Pintrich's focus on mastery and performance goals, Boekaerts' model [22] distinguishes goals for gaining knowledge from those self-focused goals [128]. Depending on whether a learners' primary interpretation is task-focused or self-focused, goal setting leads to goal striving that is either problem-focused (i.e., positive goals) or emotion-focused (i.e., negative goals). When learners aim at knowledge gain, they are considered to be problem-focused; when learners are focused on maintaining or restoring mental well-being they are emotion-focused. One unique feature of Boekaerts' model is a recommended balance of task-focused and self-focused goals whereas Pintrich [126] advocated mastery goals over performance goals.

Winne's model [172, 174, 176] specifies a schema of information processing that occurs at every SRL phase: Conditions, Operations, Products, Evaluations, and Standards (COPES) [61, 176]. Each component of the COPES schema describes how information is related to the task at each of four SRL phases – scanning conditions, planning and setting goals, executing plans, and considering major adaptations to working on tasks — and how that information is processed. For instance, conditions of COPES are defined as (1) internal contexts such as self-efficacy of a learner, affect, motivation and (2) external contexts including task difficulty, availability of peers for help, noise in a classroom. These conditions can positively or negatively affect learners' information processing and thus also their performance [170].

Efklides' Metacognitive and Affective Model of Self-Regulated Learning (MASRL) [47] explains SRL at two levels. One is a macro level represented by personal characteristics such as motivation, self-concept, and metacognitive skills. The other is a micro level which is task-specific such as ongoing feeling and thinking during task processing.

All of these models seemingly agree with the importance of contexts in understanding SRL, although each model has a their own definitions of what a primary context are: environments in Zimmerman's model, motivation in Pintrich's model, emotion in Boekaerts' model, conditions in Winne's model, and personal and task-specific traits in Efklides's model. A learner's SRL practice is defined in these specific contexts and these contexts are key to understand how to support their SRL. The IF-THEN-ELSE model of Winne [170] offers a framework for structural understanding of SRL in contexts. For example, *IF* a learner developed a phobia from an unsuccessful exam experience, *THEN* the learner might not be able to self-regulate their affect and quit the next exam. *ELSE*, they might try to find a tactic for their mental well-being during the exam. In this example, a methodological approach that could not detect the learners' context, i.e., the phobia, would miss crucial information that is significant for understanding learners' experiences and potentially implementing successful interventions to improve their learning experiences.

Despite the general agreement on the importance of contextual information in understanding SRL, how to appropriately adopt an SRL measurement to capture SRL in contexts has still been a major question. In the next section, I discuss the concept of construct validity, which is one criterion that helps researchers determine whether an SRL measurement is adopted appropriately in a particular study. Furthermore, the major criticism on currently well-received measurements in SRL studies was also discussed.

2.2 SRL measurements

2.2.1 Construct validity

Construct validity indicates how well an interpretation of a measurement is justified in showing the presence or a degree of a construct [102]. Therefore, construct validity is a key concern when considering whether a measure appropriately reflects a variable or factor in theories or models. Construct validity is also frequently pursued by correlating different measures of the same construct to investigate if there is any correspondence between those measures [4, 35, 36, 44, 64, 114, 153, 175, 185]. What is critical for the present discussion is the conditional nature of measurement validity. To address design, evaluate, and interpret of general research measures, Messick [102] emphasized two important features of construct validity: (a) that validity is not dichotomous, "unlikely to be zero or near zero [102, p.147]" and (b) that validity is specific to the context in which it was obtained. That is, the validity of the same measurement is not a fixed property across studies with different research questions and study settings.

The perspective that validity is not stable across studies naturally leads to a question of how generalizable findings of examining validity would be. In fact, the goal of these studies is not to generalize findings regarding validity of interpretation of each measurement across all SRL studies. To generalize findings involves knowing facets composing measurement scores as well as collecting measurement data from various contexts [107]. With that being said, the present studies, particularly the second and the third studies in Chapters 4 and 5, aimed to contribute to the field by adding an context-specific investigation of the validity issue to accumulate the empirical evidence toward a generalizable conclusion.

2.2.2 SRL measurement classification and criticism

Among various self-reported measures, surveys have been the most frequently used approach for studying SRL. It is cost-effective both to produce survey instruments and to obtain data from a large number of participants [6, 179]. Unlike other methods specialized in measuring cognitive,

metacognitive, affective, or motivational processes, self-report surveys are also generally capable of providing evidence regarding any of those constructs including learners' inner contexts such as affect, belief, knowledge, or perceptions.

Yet, surveys have been criticized for limitation in capturing dynamic SRL and contexts. SRL theory states that SRL changes between and during tasks in response to internal and external contexts [60, 188]. Since each SRL action is made with respect to a certain context, it is crucial to report both context and reaction in order to fully understand the SRL of learners in details. Yet, survey respondents are often asked to summarize their general experiences relatively detached from context rather than explain their specific experiences in a particular context. For example, survey instruments often include key phrases such as 'most of the time' or 'typically' to encourage respondents to aggregate their experiences and knowledge and then elicit responses across different contexts [170, 188]. Winne [170] asserts that this could cause theoretical detachment from dynamic SRL models which change throughout contexts.

Another problem of surveys asking learners to aggregate their experiences or actions is that the measures rely on learners' ability to interpret their experiences and decide what to report. To generate answers to survey questions, learners may selectively summarize experience which may provide a skewed understanding of past events. In other words, learners subjectively interpret or 'sample' their past experiences in the process of articulating their cognitive events [170]. Furthermore, if respondents are asked retrospectively to think about their previous behaviors that may have happened a while ago, it increases the chances of incomplete recall or memory distortion [164]. Regarding using a pre-survey which does not require respondents' recall, Azevedo [6] pointed out the potential discrepancy between learners' perception of how they would behave a priori and actual actions occurs during learning. Karabenick et al. [79] also raised the issue of respondents' interpretations of survey items, particularly when items are designed to operationalize abstract concepts such as those used to assess self-regulated learning.

On the other hand, in terms of trace measures such as clickstream designed to capture a certain SRL phase or construct, the data collection process is not intrusive and therefore learners are barely interrupted during a learning episode. Nevertheless, a trace measure is not completely free from the the issues of the constructive validity.

Researchers' perspectives when interpreting trace data could be a source of low construct validity. Kovanovic et al. [86] pointed out that the concept of time-on-task, which has been prevalently used in educational research, has been defined in multiple ways with little agreement on the definition. Their investigation shows differences in definition can lead to statistically different results. Similarly, Jo et al. [78] raised the issue that there is variance in operationally defining how to derive a learner's time management skill; researchers have used different variables such as total login time, login frequency, and login intervals. It is important to recognize that the variety in how

a variable is defined is not necessarily a problem for SRL research. A key issue when defining variables and interpreting data is consideration of contexts. Kovanovic et al. [86] demonstrated that no single definition of the time-on-task variable offers the best fit for courses on different subjects. Their results indicated that using measures based on the temporal data (e.g., how long a user viewed a quiz material) were beneficial for fully-online courses, while simple count measures (e.g., how many times a user logged in to the LMS) offered better fit for different contexts. Therefore, contexts should be thoroughly considered to decide which definition of the variable worked better.

Furthermore, representing learners' internal context such as emotional reaction through behavioral data has been still considered difficult even for well-built trace measures particularly if a platform is not developed with the purpose of capturing such data. Trace data can provide higher accuracy by excluding learners' perceptions that may influence survey responses. However, measuring the intentions or perceptions of the learning experience may be difficult through trace data which are not designed to detect such constructs [18]. For example, Aleven and Koedinger [1] could gather granular data which gave them insight on learners' help-seeking strategy usages on the ITS throughout time. Yet, they could not identify exact reasons behind learners' hesitation to look for hints, and instead, they concluded their study by suggesting several potential explanations on internal contexts of learners. Depending on which research questions researchers would like to answer, losing this information on learners' internal contexts could significantly negatively impact the quality of researchers' understanding of SRL.

2.3 How to address the validity issues of SRL measurements

One approach to address the validity issue is to compensate for weaknesses of each measure by using both trace measures and self-reported measures to report constructs. Karabenick and Zusho [80] suggested that there is a benefit to using multiple sources of data instead of relying on one type of measurement. As an example of benefiting from the strengths of each measure, Paans et al. [122] used both the think-aloud protocol and log data to understand time-varying SRL and navigation behavior in hypermedia learning environments. Participants in their study were 5th-grade children who could find it difficult to perform a think-aloud protocol while working on their tasks. Therefore, log data were adopted to record behavior that may have not been verbalized using a think-aloud protocol. The authors also investigated a correlation between those different types of data and mapped learners' navigation behaviors recorded through log data and metacognitive and cognitive processes measured through the think-aloud protocol. Bernacki et al. [19] also adopted multiple methods to study how much self-efficacy was related to learners' problem-solving performance. Learners' self-efficacy was reported through self-prompted utterances, while learners' performance and learning processes were tracked through log file data.

Another approach to address the issue is to scrutinize the validity issues of these measures in specific contexts by adopting them to report the same constructs and comparing their outputs. A few previous studies found that self-reported data and trace data did not provide the aligned outputs on measuring the same constructs. Hadwin et al. [64] used log data of learning behaviors and surveys responses to the Motivated Strategies for Learning Questionnaire (MSLQ) [125]. Two sets of student profiles were built, based on each type of data, and students were subsequently clustered into those profiles. Not only did profiles based on log data provide data with more details, there was also a discrepancy between students' self-reported responses on how they studied and the log data indicating their actual learning behavior. Zhou and Winne [185] adopted both self-reported surveys and log data to understand learners' goal orientation and found that log data, which is trace measures, were a better predictor of learners' performance and thus supported achievement goal theory better than self-reported measures. Considering that achievement goal has been predominantly studied through self-reported measures, this is a striking result. In their findings, there was nearly no correspondence between goals derived from self-reported and trace data. Another example is Susac et al. [153] who compared eye-tracking data and survey data to understand learners' strategy usage during algebraic equation rearrangement. In the study, eye-tracking trace data showed that a few learners checked correct answers given on the same page while working on the problem sets. Yet, almost no learner reported on the survey that they used such a strategy. From the results, the authors speculated that the accuracy of their metacognition of the learners was not high enough to fully track their strategy usages.

While these previous works raised the question on the validity of researchers' interpretations of SRL measurements, more in-depth investigations are necessary to conclude how to properly use each measurement in an SRL study. For example, there should be further research on the contexts and reasons of such discrepancies between self-reported and trace measures reported. The previous work [64, 153, 185] has proposed that possible reasons could be the lower accuracy of self-reported data in capturing actions due to inaccurate recall, dishonest responses motivated by concerns of social presentation, or time delay between self-reported responses and actions. Yet, none of these researchers have formally tested these hypotheses. Furthermore, they did not shed light either on (1) potential weaknesses of trace data that might have been responsible for the discrepancy or (2) contextual factors that could cause one or both measurements to be less successful in providing results with high validity.

2.4 Current research

In the previous section, two types of approaches were suggested to address the validity issues of self-reported surveys and trace measures. One is to complement the weaknesses of both measures

when they are used together to capture different constructs. Another approach is to compare the levels of agreement between the two measures, when studying the same construct. Both types of approaches are necessary to understand the validity of interpreting SRL measurements. The complement approach is a practical option for researchers to continue to answer SRL-related research questions with decent construct validity of interpreting measurements. The comparison approach is a path for researchers to optimize their research methods by deciding whether self-reported surveys or trace measures are more applicable to their study.

The following chapters present three studies posing each of these approaches to the validity issues of SRL measurements. The first study integrated self-reported survey and log data to measure short-term and long-term effects of reflection prompts on learners' satisfaction, perceived learning, and task performance. The second study examined trace data on goal-relevant behaviors and responses to achievement goal surveys to ensure consistency with theoretical models. Finally, in the third study, a data-driven investigation was conducted on the same datasets from the second study to expand the theory mainly built on survey data.

To avoid common misunderstanding that statistical significance i.e., p-value$<$0.05 firmly divides the rejection of affirmation of the null hypotheses (See the statement of American Statistical Association (ASA) [167]), the following three studies adopted recommendations from ASA [168] such as (1) reporting all findings without favoring statistically significant results, (2) sharing p-value in continuous format (e.g., p-value$=$0.023 instead of p-value$<$0.05), (3) presenting more evidence such as effect size, multiple model fit criteria, or contextual information for transparency and replicability, and most importantly (4) abandoning the phrase 'statistically significant.'

CHAPTER 3

Complementing SRL measurements

3.1 Introduction

3.1.1 Reflection prompts and hints

Providing reflection prompts to learners has been a widely adopted instructional practice aimed at improving learners' academic performance. Reflection is generally defined as a process of expanding and deepening one's understanding by critically analyzing what has been learned and how [24, 41, 93, 108]. It scaffolds knowledge acquisition by facilitating the development of cognitive structures and making them available for problem-solving [21, 160]. Reflection also leads learners to deliberately review their learning experiences and learn from them. Hence it tends to increase academic performance [24, 41, 93, 108, 121]. Reflection has been also employed in the field of Computer-Supported Collaborative Learning (CSCL), as it has been revealed to contribute to building more constructive answers as a group [133, 134]. Due to these clear benefits of reflection, researchers have been designing various kinds of interventions aimed at encouraging learners to engage in reflection during different phases of the learning process. An often-used intervention is metacognitive prompts. Over more than two decades, researchers have confirmed the effect of prompts on stimulating various metacognitive processes including reflection [32, 83, 94, 95].

Given the role of reflection in knowledge acquisition and academic performance, reflection prompts might be a useful metacognitive support encouraging meaningful learning from hints and improving learning gain. Hints are commonly used to provide cognitive support explaining how to solve a given problem [132] or triggering a particular cognitive process [116]. Hints could be provided in multiple formats. For example, in CSCL work, different parts of hints were given to each learner to encourage learners' collective thinking [43, 155]. Yet, hints are often shown to be ineffective in increasing learning gain despite their designed purpose to support learning. For example, Zhou et al. [186] did not find any evidence that hints could increase adult learners' correct first attempt on MOOC assignment. One reason for this is a lack of learners' mindful interactions with hints [2, 132]. In general, learners do not deliberately activate metacognitive processes such

as reflection when they learn [17, 25, 32, 95]. Furthermore, in their study of an Intelligent Tutoring System (ITS), Aleven et al. [2] showed that nearly half of undesirable interactions with hints are caused by learners aiming to find answers from hints without trying to thoroughly understand them (e.g., mindlessly clicking through hints). Follow-up work by Roll et al. [132] found that guiding learners away from these behaviors did not make a meaningful difference in learning gains, suggesting that a solution might be instead to encourage mindful interactions with hints. In their study [109] in Khan Academy with Spanish undergraduates, Muñoz-Merino et al. found that hint abusers tended to submit incorrect answers in less than 10 seconds. The authors identified this behavior tendency as the evidence of a lack of reflection processes while engaging with questions and hints. Taken together, providing reflection prompts along with hints might be the missing piece in encouraging learners to mindfully interact with and learn from hints.

The combined effect of hints and reflection prompts on task performance is still underexplored. Berthold et al. [20] investigated a combination of cognitive and metacognitive prompts as means of promoting reflection and monitoring. Yet, their cognitive prompts were different from hints in that they were designed to activate organization and elaboration strategies instead of directly explaining domain knowledge that learners might lack. Marwan et al. [101] showed that hints meaningfully increased immediate programming performance, but only when they were accompanied by self-explanation prompts. However, the self-explanation prompts did not seem to elicit reflection processes and instead were (1) designed primarily for general critical thinking rather than reflection and (2) given before letting learners apply hints to solve a problem. In fact, Marwan et al. [101] reported that during the post-study interview, one of the study participants suggested presenting self-explanation prompts after the task so that they could reflect upon the whole process including how they understood and used hints. It is clear that more investigations are necessary to comprehend the effect of reflection prompts on learners' interaction with hints and to examine the effect of such interactions on task performance.

This work addresses such a need and examines how hints and reflection prompts support programming learning, taking into consideration the duration of task complexity and prompt effects. To understand how providing learners with hints and reflection prompts relates to performance, we conducted an experiment combining three scaffolds: (1) hints alone, (2) reflection prompts alone, and (3) both hints and reflection prompts. We measured both immediate and delayed effects on transfer task performance, perceived learning, and course enjoyment. Our findings show that combining reflection prompts and hints benefits delayed transfer task performance, but does not affect immediate performance. Our findings show that (1) reflection prompts alongside hints can support mindful interaction with hints, and (2) such mindful interaction with hints impacts delayed performance. These findings have design implications supporting the need to include reflection prompts with hints in the learning environments.

3.1.2 Previous studies on design factors

Previous studies suggest that it is important to consider factors that might affect the effect of reflection prompts on learning outcomes [7, 8, 9, 10, 25, 38, 82, 88, 93, 95, 141]. Previous research has shown that reflection prompts could increase performance most effectively when applied to knowledge transfer tasks, that is, tasks requiring learners to apply adopted knowledge to new contexts [10, 25, 38, 82]. Learners within the same intervention did not show meaningful improvement in performance on simpler tasks such as recall or knowledge comprehension tasks [7, 8, 9, 93, 95, 141]. For instance, three studies by Bannert and colleagues presented the compelling evidence for an increase in transfer task scores with moderate effect sizes (d=0.55, d=0.58, d=0.44) [7, 8, 9]. Schworm and Renkl [141] also used transfer tasks and found an increase in learners' task performance after providing self-explanation prompts which evoked reflections along with other metacognitive processes. Lew and Schmidt [93] suggested that addressing reflection prompts assisted with the synthesis of new and prior knowledge, while reflection prompts alone did not meaningfully affect learners' performance on the knowledge acquisition tasks. While Krause and Stark [88] did not find any impact of reflection prompts on task performance, they found that learners made more progress on their tasks, with a large effect size (d=0.8), only when working on relatively complex problems. There was no compelling evidence for a difference in progress on simpler problems. Lin and Lehman [95] observed meaningful increases in score only on more complex problems, which led learners to far transfer their conceptual understanding to dissimilar contexts. These findings clearly present a relationship between the complexity of tasks and the benefits of reflection prompts. Reflection prompts are highly likely to improve performance on transfer tasks or other tasks with higher complexity, whereas they do not affect performance on less complex tasks, such as recall or knowledge comprehension tasks.

Some of the previous work has also tried to reveal a duration of the effect of reflection prompts, with Bannert et al. [9] being one of few studies that examined their delayed effect. They found that learners who received reflective prompts performed better on their delayed transfer tasks compared to the control group. In contrast, Jeong et al. [77], who studied reflection prompts in the context of concept map drawing tasks, found no effect of metacognitive prompts on their delayed task performance. Studies that reported effects measured by post-tests immediately after learning sessions also showed conflicting results [7, 8, 88, 95, 160]. For example, Krause and Stark [88] presented learners with reflection prompts asking them to justify the choice of learning strategies, and found no significant effect on learners' performance at 0.05 level. Similarly, van den Boom et al. [160] showed no meaningful effect of reflection prompts on the immediate post-test score. Yet, Bannert and colleagues [7, 8] found the compelling evidence for positive effects of reflection prompts on immediate post-test scores. These mixed results show that the duration of the effect might be an important factor to consider in reflection prompt study but it requires more investiga-

tion. The effect of reflection prompts might appear either immediately or after a delay. To account for the potential effect of reflection prompts as well as to understand the duration of the impact from the prompts, it is necessary to measure both immediate and delayed effects while controlling for the task complexity.

Finally, it is important to design reflection prompts in a way to balance likeability and effectiveness. Likeability is highly influential in learners deciding to engage with given interventions. One component of likeability leading learners to make more mindful engagement with prompts would be perceived learning. Multiple previous studies have shown that perceived learning is a component of metacognitive monitoring and evaluation which could affect learners' willingness to engage with a certain intervention [11, 46, 163]. Another factor would be how much learners enjoyed their interactions with prompts. Previous studies showed that low enjoyment and high annoyance are factors that can lead learners to low compliance with prompts [8, 20]. Taken together, if learners do not consider that interactions with reflection prompts bring enough perceived learning and enjoyment, learners would easily lose interest toward the interventions and fail to take benefit of them. To design reflection prompts which could make impacts at learning practice, it would be particularly helpful to examine designs in authentic learning environments where learners are more genuinely motivated stakeholders compared to a controlled lab study.

3.2 Research questions

In this study, we aimed to understand the immediate and delayed effects of different combinations of hints and reflection prompts on task performance. The task complexity was controlled and thus only the transfer task was used. Furthermore, we designed a field study instead of a lab study, to collect the data in an authentic learning environment. We asked the following research questions:

- RQ1. What is the immediate effect of the combined hint and reflective prompt intervention on transfer task performance?
- RQ2. What is the delayed effect of the combined hint and reflective prompt intervention on transfer task performance?
- RQ3. What is the immediate effect of the combined hint and reflective prompt intervention on meaningful interaction with the task?
- RQ4. What is the delayed effect of the combined hint and reflective prompt intervention on meaningful interaction with the task?
- RQ5. What is the immediate effect of the combined hint and reflective prompt intervention on perceived learning?

- RQ6. What is the immediate effect of the combined hint and reflective prompt intervention on enjoyment on learning?

We posed RQ1 and RQ2 to investigate if there is any difference between the immediate (RQ1) and delayed effects (RQ2) of the interventions on transfer task performance. RQ3 and RQ4 were added to understand learners' immediate and delayed behavioral changes in solving programming tasks measured through the number of task submissions. In this particular learning environment, learners were allowed to submit their tasks as many times as they wanted and could check their scores for each submission. If learners in a certain condition submitted tasks more only to achieve the same or lower scores than other conditions, it would be the evidence of learners' less meaningful engagement with interventions and tasks. Furthermore, particularly within the hint condition, it also might mean learners tried to game the system by carelessly guessing answers and checking if their guesses were correct through abusing hints. RQ5 and RQ6 were posed to answer the effect of the interventions on perceived learning and enjoyment after using reflection prompts, which are components of the likeability of the interventions..

This study addressed these questions through a three-condition experiment: (1) the hint intervention was provided while a learner was working on a task i.e., the hint condition, (2) a reflection prompt intervention would appear when a learner finished tasks i.e., the reflection condition, and (3) both the hint and a reflection prompt intervention present the programming task delivered at times described above i.e., the hint-reflection condition. The hint condition was a control condition, while the hint-reflection condition was the main treatment condition. The reflection condition was additionally included to see if reflection prompts alone could impact the listed learning outcomes, when compared to the hint condition.

3.3 Pilot study

3.3.1 Study context

Prior to the main field experiment, a pilot study was conducted to ensure the appropriate design of hints and reflection prompts. The primary design decisions to be made were (1) when to present reflection prompts - every time before or after the hint was presented or only after a learner completed the task, (2) which concepts or code elements to explain through hints, (3) which type of reflection prompts was effective in eliciting reflection, and (4) if learners could understand task questions, reflection prompt questions, and hints. For the pilot study, nineteen learners who had already taken this course were recruited.

3.3.2 Study procedure

A set of reflection prompts were designed based on previous literature [37, 40, 74, 183]. The authors and the course instructor reviewed and discussed the prompt designs to ensure that the prompts can activate reflection processes. During the pilot study, the first author explained the benefits of reflection prompts to the participants, to prevent potential annoyance and unwillingness from interacting with the prompts [8]. Learners were then asked to think aloud while working on tasks with both hint and reflection prompt interventions. Learners were provided with a reflection prompt before and after every hint request. Tasks were followed by an individual interview.

3.3.3 Instruments

3.3.3.1 Follow-up interviews

The first author asked common questions and follow-up questions to clarify some of the statements made during each participant's think-aloud session. Common questions were as follows:

- Think about how you interacted with the prompts. Did prompts help you check your understanding with hints?
- Was it easy to understand what prompts asked you to think about?
- Think about how you interacted with hints while working on the Jupyter Notebook programming tasks. Did hints address what you wanted to know to solve your problem?
- Was it easy to understand the suggestions that hints gave?

Follow-up questions varied and were determined by some specific statements the participants said while self-reporting (think-aloud) on the task.

3.3.4 Pilot study results

Overall, pilot study participants did not comply with reflection prompt instructions when they saw prompts either before or after a hint. The reflection prompt given before each hint was designed to encourage them to reflect on what they already knew and what they were not sure of. However, 13 out of 19 participants ignored these reflection prompts and proceeded to see hints. During the follow-up interview, one participant (P2) said that they did not have enough motivation to engage with a reflection prompt when they were frustrated, and were eager to see a hint to resolve the frustration as soon as possible. Another participant (P12) added that using reflection prompts before seeing hints did not add anything to their understanding. One participant (P15) even described

the prompt as 'an extra obstacle to the selection of hints.' Similarly, 9 out of 19 participants also considered reflection prompts given after every hint annoying and did not comply with prompts asking them to activate reflection.

The participants' annoyance due to the high frequency of prompting along with low compliance with prompts is not at all surprising considering findings from previous studies [8, 20]. Yet, it was concerning that most participants did not engage with reflection prompts even though the benefits of engaging with reflection prompts were explained to them, to motivate them for such engagements [8]. To reduce learners' annoyance and to draw their attention to reflection prompts, for the main study we decreased the frequency of prompting to only once when learners completed their tasks. In the modified design, reflective prompts aimed to activate learners' reflection process over the entire tasks, including but not limited to their use of hints. Marwan et al. [101] suggested such a design to encourage reflections instead of eliciting them after each hint.

Based on learner feedback, we also included an additional hint to address a problematic area in the assignment. Otherwise, learners agreed that the assignment questions and reflection prompts were clear and understandable.

3.4 Method

3.4.1 Study context

A field study was conducted where learners received different combinations of hints and reflection prompts interventions while working on a programming transfer task. There were two tasks: (1) the task measured the immediate effect of the interventions and (2) the task measured the delayed effect of the interventions. Interventions (i.e., hints and reflection prompts) accompanied immediate tasks, but not the delayed tasks.

We used G*Power [55] to decide the estimated sample size, which showed that there should be more than 159 participants total i.e., at least 53 per condition for the medium effect size (Cohen's $f = 0.25$) with an alpha of .05 and power of .80 for one-way ANOVA tests. Since most previous studies reported their results with medium to large effect sizes, we determined to use the medium effect size based on the Cohen [33].

Data were collected during two iterations of the same introductory data science course in the fully online Master's degree program of University of Michigan, School of Information. The course materials included lecture videos, readings, and programming assignments. Iteration 1 took place first which was followed by iteration 2 in the next semester. Each iteration was four weeks long and had a different sample and different number of enrolled learners. The instructor team was ethically obliged to provide a similar learning experience to all learners who enrolled in the course

for credits in the same term. Hence, all learners enrolled in the first iteration were assigned in the reflection condition. All learners who enrolled in the second iteration of the course were split between the hint-reflection condition and the hint condition. This was based on discussions with the instructor who expected that giving only a portion of learners hints would discourage learners who did not get hints and even feel they were not fairly treated compared to other students in the same class.

The total number of learners who enrolled in the two iterations of the course was 432 before removing learners who did not complete even a single task. While each of the course materials was recommended in certain weeks, some learners worked ahead and finished materials earlier. That is, while they followed all the course materials in order as planned, they might have finished courses earlier than other students. IRB oversight was obtained through University of Michigan study ID HUM00151900.

3.4.2 Instruments

3.4.2.1 Questionnaire on metacognitive skills

To confirm that learners in different conditions did not meaningfully differ in self-reported metacognitive skills including reflection before engaging with the interventions, a questionnaire on metacognitive skills was conducted at the beginning of the course. Six question items were adopted from the Metacognitive Self-regulation section of the Motivated Strategies for Learning Questionnaire (MSLQ) [125] and slightly reworded to follow the context of the course.

3.4.2.2 Tasks and assignments

To measure learners' performance on immediate and delayed knowledge transfer tasks, four weekly assignments were prepared in Jupyter Notebook, a web-based programming environment [84] used in programming courses. All assignments consisted of Python programming tasks that were automatically graded by a system embedded on the Coursera platform upon submission. All tasks were designed and reviewed by the course staff and the first author to confirm that they measured what learners were supposed to learn each week. Conceptually, each assignment was built upon the previous assignments. Therefore, it was expected that learners who did not perform well on earlier assignments would perform poorly on the subsequent assignments. Learners had to reach 80% of the full credit per assignment to pass the course; they were allowed to submit assignments as many times as they wanted prior to a weekly deadline.

The first two assignments, assignments 1 and 2, consisted of three separate tasks each and included an intervention. These tasks were used to measure the immediate effect of the intervention on learners' academic performance. Since each of these tasks was provided separately, learners

who were provided with reflection prompts could activate their reflection per task. That is, these learners saw reflection prompts at least three times per assignment since they had to submit three different tasks for each of assignments 1 and 2. The last two assignments, assignments 3 and 4, included more than one task per assignment and did not offer any interventions to learners in any of the conditions. These assignments were used to measure the delayed effect of interventions received for assignments 1 and 2 on learners' academic performance.

3.4.2.3 Hints

Learners who were offered hints (i.e., the hint-reflection condition and the hint condition) could engage with the hint intervention by clicking the 'Show Hint' button while working on a task. When learners clicked the button, a pop-up with a list of summaries of available hints was displayed (Figure 3.1 (a), (b)).

When learners chose a hint and clicked the 'Next' button, they could see the full text of the chosen hint on the next pop-up. This full hint was inscribed below the associated task cell (Figure 3.1 (c)), so that learners could easily look up hints while working on a task even after closing the hint pop-up.

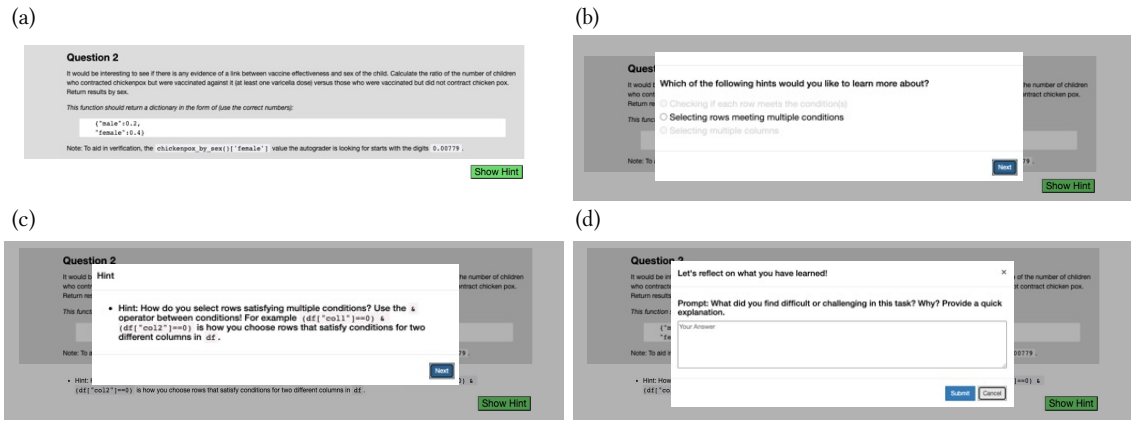

Figure 3.1: (a) Each task has a 'Show Hint' button below a question. (b) When a learner clicks the 'Show Hint' button, a pop-up appears and shows a list of available hints. Some hints are grayed out here since they have already been chosen by the learner in previous interactions. (c) The full text of the chosen hint is shown on a pop-up and also inscribed below the associated question so that a learner can see hints even after closing the pop-up. (d) When a learner clicks the submit button, a reflection pop-up appears.

3.4.2.4 Reflection prompts

Learners assigned to the conditions with a reflection prompt intervention (i.e., the reflection condition and the hint-reflection condition) were given a reflection prompt when they clicked the submit button to submit their task and receive auto-graded credit (Figure 3.1 (d)).

All reflection prompts were designed as directed reflection prompts, i.e. prompts that offer specific instructions such as 'stop and think about what you misunderstood before seeing hints.' We opted for the directed prompts based on the pilot study participants' preference for them over the generic prompts that encourage learners' reflection by simply asking 'stop and think.' The distinction was proposed by Davis [37] and no agreement currently exists on which of the two are more beneficial for the effective reflection process [37, 74, 87, 92]. Hence, upon task submissions, a learner would be provided with a reflection prompt statement randomly chosen from the following list:

- What steps did you take when solving the problem? Why? Provide a short justification for each step.
- What did you find difficult or challenging in this task? Why? Provide a quick explanation.
- What was the main thing you learned by completing this task?

3.4.2.5 Questionnaire on perceived learning and enjoyment

A questionnaire was also distributed to learners to measure perceived learning and enjoyment over the learning experience with the given intervention. The questionnaire had four 7-point Likert scale questions for perceived learning and three 7-point Likert scale questions for enjoyment, where both of the question sets asked how much a learner agreed with each statement. Statements were adopted from Barzilai and Blau [11] and revised for this study as presented in Table3.1.

3.4.3 Study procedure

At the beginning of the first week, learners were asked to take the questionnaire on metacognitive skills. Learners assigned to the hint condition and the hint-reflection condition could use the hints while working on tasks included in assignments 1 and 2. Learners in the reflection condition or the hint-reflection condition were also asked to report their reflection process as a written response, upon submitting their tasks for assignments 1 and 2. Having completed week 1 and week 2 materials and prior to embarking on week 3, learners were required to take a questionnaire on perceived learning and enjoyment. The course was self-paced with fixed dates of assignment deadlines.

Construct measured	Questionnaire items
Perceived learning	• The weekly assignments helped me learn more about the topic (e.g., regular expression) • I learned new things from the weekly assignments. • The weekly assignments helped me remember the things I learned. • The weekly assignments helped me apply the things I learned to other problems.
Enjoyment	• I enjoyed the weekly assignments. • I had fun working on weekly assignments. • Working on the weekly assignments was pleasant.

Table 3.1: Question items for measuring perceived learning and enjoyment

3.4.4 Data analyses

For data cleaning purposes, we removed incomplete submissions (e.g., questionnaires which were only partially filled out) while including learner data as long as they made at least one complete submission of a task or an assignment. For example, there were learners who received the grace period for a portion of assignments due to their personal circumstances and their tasks were manually graded with different criteria. In this case, we still retained their other submissions and questionnaire response data instead of dropping the entire submission data of the learner.

To address RQ1 and RQ2 related to immediate and delayed effects of interventions on the transfer tasks, we conducted a mixed-effect model analysis in R using the lme4 package [12]. A mixed-effect model was adopted to account for scores of multiple assignment submissions per learner. Data of submitted assignment scores were split into two sets and respectively used to analyze the immediate effect (data of week 1 and week 2 assignments) and a delayed effect (data of week 3 and week 4 assignments). That is, there were two mixed effect models: one for the immediate task and the other for the delayed task. The experimental condition and assignment label for the week that the assignment was due (i.e., week 1, week 2, week 3, and week 4) were used as the fixed effects. For example, in the mixed-effect model for the immediate task, there were two assignment values, i.e. week 1 or week 2, included as a covariate. Learner IDs were used as random effects in the model. Regarding RQ3 and RQ4, a one-way Analysis Of Variance (ANOVA) was conducted over the number of submissions for measuring the immediate effect on performance (RQ3) and for measuring the delayed effect on performance (RQ4). For RQ5 and RQ6, we ran a one-way ANOVA over the questionnaire responses on perceived learning (RQ5) and enjoyment (RQ6).

The normality of the residuals of the assignment score data was visually inspected and con-

firmed by a quantile-quantile (QQ) plot. Considering that the submission count data were highly right-skewed, we applied reciprocal transformations to the immediate task submission count and logarithmic transformation to the delayed task submission count data. The data transformation reduced the skewness values respectively to 0.46 and -0.39 which are in the generally accepted range of normal distribution. Even though the perceived learning data and enjoyment data were also non-normally distributed and all showed high skewness values, statistical tests such as ANOVA and regression analysis are still statistically valid options for the non-normally distributed Likert-scale data [67, 117].

3.5 Results

3.5.1 Data overview

The total number of learners in the dataset used for analysis was 354 with the following breakdown across the conditions: 70 participants in the hint condition, 222 participants in the reflection condition, and 62 participants in the hint-reflection condition. Overall, the number of participants meets the target sampling size.

The overall means of the submission score of immediate tasks and delayed tasks were 53.26 out of 100 ($SD = 49.89$) and 44.02 out of 100 ($SD = 34.71$). Means and standard deviations of the submission scores per condition are given in Table 3.2. It is important to note that each of assignments 1 and 2 was split into each task which makes them into six tasks (1-1, 1-2, 1-3, 2-1, 2-2, and 2-3) and therefore task score was either 0 or 100, which explains large standard deviations. On the other hand, assignments 3 and 4 were composed of multiple tasks.

Average submission counts of immediate tasks and delayed tasks were respectively 2.08 ($SD = 2.30$) and 16.79 ($SD = 13.95$). Means and standard deviations of the number of submissions per condition are presented in Table 3.3.

Learners' average enjoyment score was reported as 17.88 out of 21 ($SD = 3.26$) and the average score of perceived learning was 25.01 out of 28 ($SD = 2.87$). Means and standard deviations for the scores per condition are shown in Table 3.4.

An ANOVA was run on the questionnaire on metacognitive skills and it confirmed that the learners' prior metacognitive skills were not statistically different across conditions at the 0.05 level.

Assignment/task labels	Hint	Reflection	Hint-reflection
Task 1-1	67.48 (46.99)	73.83 (44.01)	74.02 (44.03)
Task 1-2	66.46 (47.36)	71.36 (45.26)	82.08 (38.54)
Task 1-3	34.60 (47.66)	36.82 (48.26)	40.38 (49.18)
Task 2-1	63.23 (48.38)	60.34 (48.97)	69.77 (46.11)
Task 2-2	53.30 (50.03)	58.85 (49.26)	68.91 (46.48)
Task 2-3	40.09 (49.11)	36.58 (48.20)	44.86 (49.87)
Assignment 3	40.32 (35.61)	47.43 (33.42)	48.18 (34.87)
Assignment 4	41.24 (35.38)	37.74 (35.34)	48.07 (34.43)

Note. Mean (standard deviation)

Table 3.2: Means and standard deviations of assignment scores per condition

Assignment/task labels	Hint	Reflection	Hint-reflection
Task 1-1	1.85 (1.30)	1.57 (1.01)	1.58 (1.22)
Task 1-2	1.85 (1.32)	1.54 (1.27)	1.32 (0.63)
Task 1-3	2.98 (2.68)	2.89 (3.00)	2.66 (2.17)
Task 2-1	1.80 (1.60)	1.85 (1.56)	1.61 (1.43)
Task 2-2	2.11 (1.65)	1.84 (1.99)	1.48 (0.99)
Task 2-3	2.72 (3.05)	3.00 (4.36)	2.31 (1.66)
Assignment 3	21.56 (17.38)	24.18 (16.44)	17.63 (12.20)
Assignment 4	8.69 (5.44)	12.22 (8.04)	7.00 (4.98)

Note. Mean (standard deviation)

Table 3.3: Means and standard deviations of assignment submission counts per condition

	Hint	Reflection	Hint-reflection
Perceived learning	23.87 (4.06)	25.27 (2.38)	25.43 (2.56)
Enjoyment	17.14 (3.86)	18.00 (3.06)	18.28 (3.13)

Note. Mean (standard deviation)

Table 3.4: Means and standard deviations of perceived learning and enjoyment per condition

3.5.2 Study results

RQ1 and RQ2 asked whether hint and reflective prompt interventions had the immediate effect (RQ1) and the delayed effect (RQ2) on the performance of a transfer task. These results analyzed through mixed-effect modeling are reported in line with the guidelines suggested by Brown [29].

Analysis to address RQ1 revealed that there was no compelling evidence showing a difference across the conditions in the performance of the transfer tasks that measured the immediate effect of the interventions ($\chi^2 = 2.83$, $p = 0.24$). Assignment label (i.e., week 1 and week 2) also did not

make statistically different performance at the 0.05 level ($\chi^2 = 1.64$, $p = 0.19$) (Table 3.5). Yet, data analysis addressing RQ2 showed that there was a compelling evidence for a difference across the conditions in the effect on the performance of the delayed transfer tasks ($\chi^2 = 12.14$, $p = 0.002$). Assignment labels (i.e., week 3 and week 4) also affected task performance ($\chi^2 = 103.08$, $p = 2.2e - 16$) (Table 3.6). Task score of the hint-reflection group was on average an estimated 10.26 points higher than the hint group ($\hat{\beta} = 10.26$, $SE = 2.91$, $t = 3.51$). Furthermore, a mean of week 4 assignment scores was 7.41 points lower than a mean of week 3 assignment scores ($\hat{\beta} = 6.75$, $SE = 2.21$, $t = 3.04$). Tukey's post-hoc test confirmed that when results are averaged over the levels of assignment label factor[1], the hint-reflection condition gained higher task scores than the hint-reflection condition with a small effect size ($\hat{\beta} = -9.06$, $p = 0.004$, Cohen's $f = 0.17$). The task score of the reflection group was higher than the hint group ($\hat{\beta} = -1.64$, $SE = 2.14$, $t = -0.76$). Yet, the post-hoc test showed no compelling evidence for a difference between the reflection condition and the hint condition ($p = 0.72$). Finally, a difference was observed between reflection condition and the hint-reflection condition with a small effect size ($\hat{\beta} = -7.41$, $p = 0.003$, Cohen's $f = 0.17$).

Predictors	Estimates	[Lower CI, Upper CI]	p-value
Fixed Effects			
(Intercept)	56.71	[49.95, 63.47]	1.00e-58
Reflection	4.50	[-3.08, 12.09]	0.24
Hint-reflection	8.07	[-1.85, 17.99]	0.11
Assignment week 2	-1.02	[-8.24, 6.20]	0.78
Reflection $*$ assignment week 2	-1.06	[-9.20, 7.08]	0.79
Hint-reflection $*$ assignment week 2	-1.96	[-12.78, 8.86]	0.72
Random Effects			
σ^2	2171.55		
$\tau_{00Learner\_id}$	257.89		
ICC	0.11		
$N_{Learner\_id}$	315.00		
Observations	3748.00		
Marginal R^2/Conditional R^2	0.002/0.108		

Note 1. Lower CI = lower bound of 95% confidence interval, Upper CI = upper bound of 95% confidence interval

Note 2. ICC: Intraclass correlation coefficient

Table 3.5: A summary table of the mixed effect model for RQ1

[1]Suppose d1 is the difference between condition 1 and condition 2 for assignment week 3 and d2 is the difference between condition 1 and condition 2 for assignment week 4. Then the result reported would be (d1 + d2)/2.

Predictors	Estimates	[Lower CI, Upper CI]	p-value
Fixed Effects			
(Intercept)	44.58	[40.68, 48.48]	1.55e-108
Reflection	6.75	[2.41, 11.10]	2.31e-03
Hint-reflection	10.26	[4.54, 15.98]	4.37e-04
Assignment week 4	0.48	[-3.31, 4.27]	0.80
Reflection $*$ assignment week 4	-10.22	[-14.35, -6.09]	1.28e-06
Hint-reflection $*$ assignment week 4	-2.41	[-8.15, 3.32]	0.40
Random Effects			
σ^2	1068.52		
$\tau_{00Learner\_id}$	136.89		
ICC	0.11		
$N_{Learner\_id}$	315.00		
Observations	10090.00		
Marginal R^2/Conditional R^2	0.018/0.129		

Note 1. Lower CI = lower bound of 95% confidence interval, Upper CI = upper bound of 95% confidence interval
Note 2. ICC: Intraclass correlation coefficient

Table 3.6: A summary table of the mixed effect model for RQ2

RQ3 and RQ4 asked how each of the interventions affected the behavioral pattern of interacting with tasks measured by the number of immediate and delayed task submissions. An ANOVA for RQ3 showed that there was a difference across conditions in the number of immediate task submissions with less than a small effect size ($F(2, 1852) = 4.06$, $p = 0.01$, Cohen's $f = 0.07$) (Table 3.7). The follow-up Tukey post-hoc test also did not show compelling evidence for any difference. For effect on the number of delayed task submissions (RQ4), there was a difference across conditions in the number of delayed assignment submissions with small effect size ($F(2, 617) = 6.80$, $p = 0.001$, Cohen's $f = 0.15$) (Table 3.8). The follow-up Tukey post-hoc test revealed that the reflection group submitted more than both the hint-reflection group ($p = 0.0008$, Cohen's $f = 0.15$). There was no statistical difference at 0.05 level in the submission count between (1) the hint group and the hint-reflection group and between (2) the hint group and the reflection group.

RQ5 asked if the interventions affected perceived learning over assignments 1 and 2. An ANOVA revealed a difference across conditions with small effect size ($F(2, 312) = 4.61$, $p = 0.001$, Cohen's $f = 0.20$) (Table 3.9). Tukey's post-hoc test results supported that perceived learning in the hint-reflection condition was higher than the hint condition with a small effect size ($p = 0.04$, Cohen's $f = 0.17$). Furthermore, reflection condition also reported higher perceived learning than the hint condition ($p = 0.009$, Cohen's $f = 0.14$). There was no compelling evidence

Predictors	Estimates	[Lower CI, Upper CI]	t value	p-value
(Intercept)	-0.69	[-0.73, -0.65]	-36.52	0.00
Reflection	-0.04	[-0.09, -0.00]	-2.12	0.03
Hint-reflection	-0.07	[-0.12, -0.02]	-2.78	0.01
Observations	1855.00			
R^2	0.004			
Adjusted R^2	0.003			

Note 1. Lower CI = lower bound of 95% confidence interval, Upper CI = upper bound of 95% confidence interval

Note2. Robust standard error was calculated through HC3 option in jtool R package summ function [97].

Table 3.7: A summary table of the regression model for RQ3

Predictors	Estimates	[Lower CI, Upper CI]	t value	p-value
(Intercept)	15.79	[12.90, 18.69]	10.71	0.00
Reflection	1.64	[-1.56, 4.84]	1.01	0.31
Hint-reflection	-4.00	[-7.58, -0.42]	-2.19	0.03
Observations	620.00			
R^2	0.02			
Adjusted R^2	0.01			

Note 1. Lower CI = lower bound of 95% confidence interval, Upper CI = upper bound of 95% confidence interval

Note2. Robust standard error was calculated through HC3 option in jtool R package summ function [97].

Table 3.8: A summary table of the regression model for RQ4

for a difference between the hint-reflection condition and the reflection condition ($p = 0.99$).

Predictors	Estimates	[Lower CI, Upper CI]	t value	p-value
(Intercept)	23.92	[22.66, 25.19]	37.17	0.00
Reflection	1.32	[0.02, 2.63]	1.99	0.05
Hint-reflection	1.38	[-0.08, 2.85]	1.86	0.06
Observations	315.00			
R^2	0.03			
Adjusted R^2	0.02			

Note 1. Lower CI = lower bound of 95% confidence interval, Upper CI = upper bound of 95% confidence interval

Note2. Robust standard error was calculated through HC3 option in jtool R package summ function [97].

Table 3.9: A summary table of the regression model for RQ5

Lastly, RQ6 asked how much learners enjoyed assignments 1 and 2 (Table 3.10). An ANOVA showed that there was no compelling evidence for a difference in enjoyment across conditions ($p = 0.10$).

Predictors	Estimates	[Lower CI, Upper CI]	t value	p-value
(Intercept)	16.86	[15.70, 18.03]	28.45	0.00
Reflection	1.08	[-0.16, 2.32]	1.72	0.09
Hint-reflection	1.10	[-0.42, 2.61]	1.43	0.15
Observations	315.00			
R^2	0.01			
Adjusted R^2	0.01			

Note 1. Lower CI = lower bound of 95% confidence interval, Upper CI = upper bound of 95% confidence interval
Note2. Robust standard error was calculated through HC3 option in jtool R package summ function [97].

Table 3.10: A summary table of the regression model for RQ6

3.6 Discussion and future work

This field study deployed in an online programming course did not show substantial differences between hints, reflection, and hints-reflection conditions in terms of effects of interventions on (1) immediate effect of interventions on the immediate task performance and (2) enjoyment. However, it provided evidence for the delayed effect on delayed transfer tasks performance when hints and reflection prompts are combined. This suggests that hints could have a delayed effect on achievement only when they are followed by reflection prompts.

Combined with the findings of the submission count, the effect of interventions on the delayed task performance becomes clearer. Compared to those in the hint-only condition, learners exposed to both hints and reflection prompts scored higher, whereas maintaining a similar count of assignment submissions. These results can be interpreted in several ways. Considering that each task was conceptually built on the previous tasks, the hint-reflection intervention can be understood as beneficial in maintaining the knowledge obtained from the previous tasks activated through reflection processes. Maintaining knowledge could enable learners in the hint-reflection condition to perform better by making fewer incorrect submissions. The other interpretation, compatible with the former, is that the hint-reflection intervention led learners to interact with learning materials and tasks in a more meaningful manner. They might have built such interaction patterns, instead of ineffective interactions with hints (e.g., making 'wild' guesses) while working on the immediate tasks and might have maintained the patterns. It is also important to acknowledge that the reflection

group made more submissions than the hint-reflection group only to achieve a similar score on the delayed task. This might suggest that reflection prompts without hints or other cognitive support could not fill learners' knowledge gaps. That is, the combination of both reflection prompts, as metacognitive support, and hints, as cognitive support, could effectively improve learners' delayed performance.

The results have implications for how to encourage learners to learn from hints. The first implication is the importance of interventions eliciting reflection processes in learning with hints. Previous studies have been consistent in showing that hints are generally ineffective in increasing learner performance [2, 132]. Their findings showed that more than half of learners adopt a mindless and passive approach to hints without directing themselves to learn from the hints [2]. The reflection prompts with hints might effectively address the issue by encouraging learners to learn from hints, by reviewing how they have applied hints while solving the given problems and what they learned from the hints. Furthermore, our results have shown that the perceived learning in both conditions with reflection prompts was higher than in the hint condition, while there was no compelling evidence for a difference in enjoyment across conditions. This suggests that the current design of prompts with hints does not affect learners' annoyance with intervention design.

Another implication for design stemming from our results is that a reflection prompt does not need to follow every hint presented to the learner. Most previous studies, even without hints, designed a system which displayed reflection prompts more than once or even showed reflection prompts after every decision making [7, 8, 9, 88, 160]. While this might have worked in a lab setting, this study showed (a) that in an authentic online classroom, learners are highly likely not to comply with reflection prompts due to annoyance, if prompts are shown before or after every hint and (b) that presenting a reflection prompt once at the end of each task could meaningfully increase task performance.

The importance of the reflection process on learning from hints as demonstrated in the current study suggests that future work can further examine the effects of other metacognitive processes on learners' use of hints. Considering the overall importance of metacognitive processes in learning [90, 151, 156, 178], the reflection would not be the only metacognitive process encouraging mindful learning. It would be interesting to see if hints affect academic performance when they are accompanied by metacognitive prompts or other interventions that elicit different metacognitive processes. With such findings, more generalizable conclusions on the importance of metacognitive processes could be drawn for the design of hint interventions.

In this study, we did not analyze the responses of individual learners which could have indicated the quality of reflection processes. For example, Engelmann et al. [54] did not find a meaningful effect of self-created metacognitive prompts including prompts asking for reflection on performance, and their analysis showed that how learners utilized prompts changed the impact of the prompts.

Accordingly, another natural extension of the current study would be to measure the quality of learners' use of prompts and examining its impact on the size of the effects of reflection prompts provided with hints. Furthermore, while the study findings are aligned with [9] who showed the benefit of reflection prompts on the delayed transfer task, we acknowledge that the current study findings have reduced power. While the study sample size was set for medium or large effects, the findings did not reach the threshold of medium effect size according to Cohen [33].

3.7 Conclusion

This work investigated the immediate and delayed effects of reflection prompts and hints on the performance in transfer tasks, perceived learning, and enjoyment in the domain of programming education. We have demonstrated that (a) none of the combinations of metacognitive and cognitive support affected immediate task performance, and (b) the combination of reflection prompts and hints meaningfully affected delayed performance on a transfer task. This study poses critical design implications on how to design a learning environment that can lead to deeper learning from hints and increased task performance.

CHAPTER 4

Confirming SRL theory through data

4.1 Introduction

Designing and validating trace indicators of the inner state of mind is challenging yet significantly rewarding to the SRL field. Despite the limitations of surveys in capturing various SRL constructs which were discussed in Chapter 2, less work has been done in examining the possibility of trace indicators in capturing learners' state of mind such as achievement goals. This is likely due to the difficulty of measuring learners' inner state of mind through behavioral data.

These difficulties should not discourage researchers from balancing the limitations of surveys with trace indicators. Designing and adopting trace indicators could contribute substantially to expand and deepen the understanding of internal SRL constructs. For example, internal constructs are often sensitive to other contexts such as task difficulty or peer pressure, and could frequently change accordingly. Less obtrusive and fine-grained trace data could be helpful in detecting dynamic changes without consistently interrupting the learning process to prompt learner with questions. Researchers could also explore different insights by adopting trace measures that can capture goal-relevant behaviors in real-time through relying less on learners' memory and honesty. Zhou and Winne [185] have also reported empirical evidence questioning the validity of surveys specifically and advocating the potential use of trace indicators in studying achievement goals.

The present study was conducted and brought several contributions. First, this study expanded the achievement goal theory through capturing behavioral evidence demonstrating how learners exercise their achievement goals in authentic learning situations. Contrasting between behavioral evidence and survey responses, this study explored beyond the findings of its previous studies which solely relied on survey responses of learners' expectation of achievement goals. Secondly, this study adds a case of developing trace data and evaluating their fit with goal complex theory [49, 53, 158]. The process of developing trace indicators should be well-thought-out considering that the researchers' imprudent interpretation of indicators is the main source of low construct validity [78, 86]. A design template [3, 58] was adopted as a tool to approach a high validity of

trace indicators. To evaluate if and how data collected with the surveys and trace indicators fit the achievement goal theory, a confirmatory factor analysis was conducted.

4.2 Related Works

4.2.1 Goal complex theory

While there are multiple versions of achievement goal theories, the theoretical base of the present study is the goal complex theory. Goal complex theory [49, 53, 158] suggests that there are at least two dimensions with which we can understand academic goals: *the what* dimension and *the why* dimension. *The what* dimension is further composed of two categories: *mastery* and performance. On one end of a mastery-performance dimension, there are learners aiming for self-improvement judged by an intra-personal standard i.e., mastery. On the other end, there are learners aiming to outperform others with a normative standard such as grade i.e., performance [51, 99, 144]. *The why* dimension is composed of autonomous and controlling motivations which arise in self-determination theory [39, 135]. Autonomous motivation drives learners to align personal values or satisfaction with outputs from accomplishing tasks. People in this category, for example, are interested in outperforming others for the sake of a higher grade or feeling of achievement without any strong desire to demonstrate competence. Controlling motivation is driven due to external pressures or tangible rewards. For example, learners with this motivation want to demonstrate their ability and then get recognition from others such as peers, family, or instructors [39].

Goal complex theory is particularly useful in clearly designing indicators of performance goals which has a history of confusing definitions. Goal orientation theory, one of the earlier achievement goal theories, defined the motivation behind the performance goals as a desire to demonstrate one's competence and gain recognition [5, 45, 115]. Yet, a review paper by Hulleman et al. [73] showed that some studies either combined or replaced the original definition of performance goals with a more general motivation: to outperform others [48, 52, 53, 144]. The desire to demonstrate one's competence could be understood to reflect a social comparison component that learners want to stand out compared to others. On the other hand, the desire to outperform others does not necessarily include that component. As long as one can attain competence to a satisfactory extent, that is enough. These two definitions of performance goals led researchers not only to debates on which definition should be dropped [26] but also to confusion over previous studies which used the same term 'performance goals' with different definitions.

Goal complex theory, combined with self-determination theory, explains why learners would aim to outperform others. Controlling motivation comes from internal or external pressures such as fear of punishment or desires of reward. This type of motivation represents the initial definition

of performance goals: to show off one's competence and gain recognition for that. On the other hand, learners with autonomous motivation would perform tasks for inherent satisfaction and joy in accomplishing tasks or seeing their core values aligned with task accomplishment. Thus, autonomous motivation describes people who would still like to outperform others even without controlling motivation. This categorization of the motivations behind performance goals has also been supported by multiple empirical studies including Urdan and Mestas [158]. The authors showed through interviews that learners' main reasons for pursuing performance goals were autonomous and controlling motivations.

4.2.2 Self-determination theory and Self-concordance

Goal complex theory clarified the definition of performance goals by adding the dimension of autonomous and controlling motivations from self-determination theory. Self-determination theory [39, 135] also provokes another important question to achievement goal theorists who have been modeling relationships between goals and trying to understand how multiple goal pursuit happens: Can performance goals with each of these motivations be pursued together?

Empirical evidence of multiple previous studies gives positive response to this question. For instance, Koestner et al. [85] conducted three studies on how these two motivations were associated with goal progress and the use of goal implementation plans. In their studies, the authors found that autonomous and controlling motivations were not strongly related to each other. In fact, the negative correlation between autonomous and controlled motivation that Sheldon and Elliot [146] found was also weak. Ratelle et al. [130] also found that more than 90 percent of students in their two different samples endorsed a combination of equivalent levels of autonomous motivations and controlled motivations. In summary, the previous empirical findings [85, 130] suggested that autonomous and controlling motivations are not opposite poles on a continuum and thus learners can simultaneously endorse both goals despite the original interpretation of Sheldon and Elliot [146].

4.2.3 Survey and trace data

Learners' achievement goals have been predominantly measured through self-reported surveys. Surveys have been commonly used to prospectively and retrospectively measure learners' achievement goals on their tasks in numerous studies across domains from sports to psychology [13, 23, 31, 42, 51, 59, 63, 75, 76, 98, 131, 142, 177, 185]. There are multiple frequently-used surveys such as the Achievement Goal Questionnaire (AGQ) [50], Achievement Goal Questionnaire-Revised (AGQ-R) [51], and Patterns of Adaptive Learning Scales (PALS) [103, 104].

One limitation is that survey data are not fine-grained enough to capture learners' context-specific goals. Several studies tried to narrow learning experience to specific contexts through adding words to survey questions such as 'this semester' or 'this class' [56, 152]. Yet, these approaches are still not fine-grained enough to capture contexts that could heavily impact learners' goals such as the moment when a learner encounters an exam question that is unexpectedly difficult. To overcome this issue, researchers would have to develop a large number of questions which address numerous important contexts. Even after researchers develop this survey, it is likely that such a time-consuming survey would overwhelm learners and lead them to make less mindful responses.

An inherent drawback of surveys is that they require learners' attention to obtain honest and accurate descriptions of their goals or motivation. Winne [170] pointed out that researchers do not know how learners selectively choose or 'sample' what to report during self-report measures. Survey respondents have to generate their responses through recall, while trace data can directly and automatically record learners' dynamic behavior in a specific context in a less obtrusive way. Survey respondents have to predict or retrieve goal-relevant information from diverse experiences and then compute a statistic representative to shape their response to answer a given question. This process is not only cognitively demanding but also could be easily biased or distorted for various reasons, including misrecall or concerns of social presentation i.e., social desirability. Furthermore, Karabenick et al. [79] also raised the concern that respondents' interpretations of survey items might not be aligned with what researchers intended.

Despite these drawbacks, there has been little research exploring instruments other than surveys to measure achievement goals. Zhou and Winne [185] conducted one of the few studies which investigated the potential of trace data in measuring achievement goals. The authors adopted both prospective survey and trace data to understand learners' goals and found weak to no correlation between them. The authors also found that trace data had stronger correlations with learners' achievement test scores than survey data, indicating that trace data were a better predictor for test performance. Considering that goal orientation has been predominantly studied through surveys, these are striking results. The authors have suspected the lower accuracy of survey data in capturing achievement goals was due to respondents' inaccurate recall, dishonest responses, or time delay between self-reported responses and actions.

One limitation of Zhou and Winnie [185] is that the authors did not conduct a further investigation on whether trace data only validly captured achievement goals. While trace data showed a strong correlation with learners' achievement, a correlation does not guarantee high validity of the trace data in capturing what the data are supposed to measure. While trace data are comparably free of the drawbacks of survey data mentioned above, trace data, similar to other event data, can be noisy without meticulous definition of trace data; log data or clickstream data often include

signals which are not what researchers are interested. For example, a simple click on a learning resource could mean learners' engagement, attention, interest, enjoyment, or a combination of these. Extracting clickstream data which only represent a certain construct may be extremely difficult for a researcher. Thus, interpreting a clear signal from trace data is often challenging [14, 89, 171]. This issue of trace data were also found in Zhou and Winne [185], and hence there were multiple chances for misinterpretation of trace data in their study. Tasks given to learners in their study were reading online articles and taking on a post-test. While reading, learners could click hyperlinks to additional materials or tag content to study the given articles better for their post-tests. Engagement with hyperlinks or tags were used as behavioral indicators associated with learners' achievement goals in the study. For instance, if a learner clicked a hyperlink titled "Find out more information about this" while reading an article, the behavior was interpreted as an manifestation of a mastery-approach goal. On the other hand, if a learner clicked a hyperlink labeled "Avoid misunderstanding about this," it was considered as a representation of performance-avoidance goal. Yet, the behavioral indicators could have measured not only achievement goals but also other constructs as well. For instance, some learners might have been simply curious about the new technical feature, or learners might have avoided using particular features if they did not feel comfortable with incorporating this new technology in their learning, instead of a lack of mastery or performance motivation. These concerns are especially relevant in this lab study, as learners might have not had enough time to explore the given technical features and then to naturally integrate the features in their learning.

One framework which could reduce this concern of trace data inaccurately representing the targeted construct is the Evidence-Centered Design (ECD). The Evidence-centered design is created to design features of assessments to provoke particular knowledge or skills and to have learners speak or behave in a way that provides evidence about the knowledge or skills [14, 105, 106]. ECD broadly defines the assessments ranging from traditional exams composed of multiple-choice questions to instructional opportunities to capture learners' development progress [14]. In particular, the domain modeling stage of the framework is useful to identify potential under- or over-representation of targeted constructs and to design more accurate indicators. In particular, a researcher can use 'a design pattern' which is a diagram tool to scrutinize design plans of assessments [3, 58]. With the tool, researchers can represent how the design of an assessments (1) obtains data evidence about the targeted knowledge and skills of learners, (2) supports claims based on those data evidence, and (3) considers possible counter-claim.

It is also important to investigate how much each indicator of both surveys and trace data contribute to measuring achievement goals. In particular, examining trace indicators and comparing them against survey data could shape more rigorous approaches to collect well-defined data of context-specific goal-relevant behaviors with less noise. Confirmatory Factor Analysis (CFA) is a

frequently adopted analytical approach to estimate if and how much each indicator (i.e., observed data) contributes to shape factors (i.e., latent variables) [28, 65]. While multiple achievement goal studies have conducted CFA, these previous studies focused on validating a fit between the survey dataset and a given achievement goal theory using different contexts. For example, AGQ and AGQ-R were built based on data of American college students and were examined and evaluated with several populations: American students who were in short medical training, African American middle and high school students, Argentinean university students, and Italian primary and secondary school students [34, 68, 129, 136]. However, in all of these examples, only the survey data was used to conduct CFA.

The approach in this chapter is to apply CFA to both survey and trace data to explore how the two forms of evidence relate to theoretical constructs in an authentic learning environment. This is crucial to understand the potential gap between learners' expected goals before learning and their actual goal-relevant behaviors during learning in the real-world setting.

4.3 Research Questions

The aim of this study is to investigate the construct validity of interpretations grounded in self-reported survey data and trace data in a study following goal complex theory. This study is conducted in the context of an online credit-bearing course instead of a controlled lab environment. Research questions are as follows:

- RQ1. How well do self-reported survey data fit the goal complex model?
- RQ2. How well do trace data fit the goal complex model?

To answer RQ1 and RQ2, survey and trace data were collected, and then CFAs were conducted separately on the collected survey and trace datasets. Both data were collected from iterations of an online course for a master's degree at the University of Michigan, School of Information. Considering that there are already multiple well-received surveys, the ECD was only applied to design trace data indicators. By investigating the fit between datasets and CFA solution structured upon the goal complex model, evidence for the construct validity of each measurement was investigated. IRB oversight was obtained through the University of Michigan study ID HUM00203534.

4.4 Methods

4.4.1 Study context

Survey and trace data were collected during two iterations of the same introductory Python course for an online Master's degree program for applied data science offered by the University of Michigan, School of Information. This course was one credit unit and learners needed 34-credit units to complete the degree. The course was 4-week long and each iteration took place in September, 2021 (151 enrollments) and January, 2022 (98 enrollments) on Coursera. The course was one of the first technical courses that learners had to take for their degree. Learners communicated with peers and teaching staff through Slack, a communication platform providing chat rooms, and and most of assignments were provided on Jupyter Notebook, an integrated development environment.

The course was different from both residential college courses and Massive Open Online Courses (MOOCs). Compared to traditional residential college students, learners in this course were part-time students and were often more diverse in ages, background knowledge, occupations, and levels of education, as they came back to earn the degree a few years to decades after their last degree. These learners were also expected to be high in retention since they paid for the degree program to gain the access to the courses.

Learners earned credit through four weekly mandatory assignments each of which was worth 25% of the full credit. Learners could submit mandatory assignments as many times as they wanted, and their submissions were graded through an automated code grading system 'auto-grader' in a few seconds to few minutes. Learners could also obtain additional credit by completing bonus assignments. The top letter grade, $A+$, was only awarded to learners who earned 100% from mandatory assignments and additionally completed at least one bonus assignments. Figure 4.1 shows when these assignments were given to learners, and furthermore, when survey and trace data were collected.

4.4.2 Measurements and indicators

Two types of data were collected throughout the course: AGQ-R and motivation survey data [51, 162], and trace data. The survey data were collected to identify learners' expectations on what their self-reported achievement goals would be at two points of the course. Surveys were presented to learners at the beginning of weeks 1 and 3 to capture possible goal changes between the first half and the second half of the course.

As there are no widely agreed designs of trace measures for achievement goals, the domain modeling step of the ECD was applied to design course materials generating trace data which were used to shape behavioral indicators of learners' achievement goals. This involved the researcher

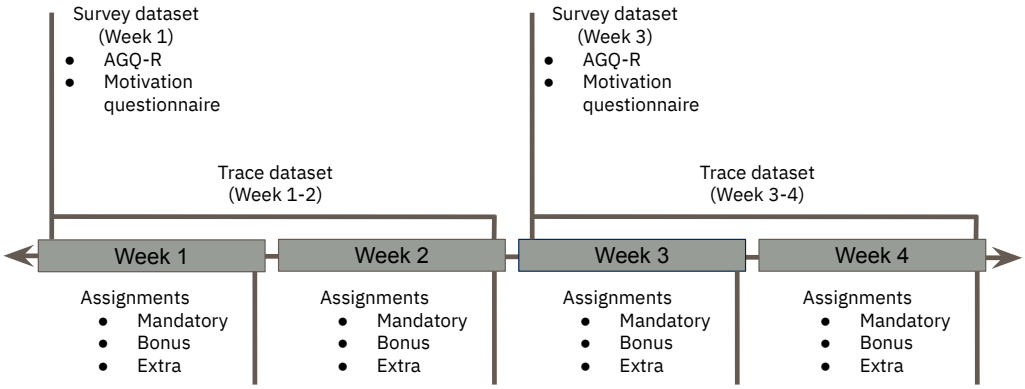

Figure 4.1: A course timeline showing when surveys and assignments were provided to learners. Trace data were collected throughout the course and were broken into two datasets based on when the data were collected.

and the instructor modifying the curriculum to include differentiated materials that theoretically attract learners with different goals. Interactions that learners made with these materials generated log data throughout the course and were used to generate behavioral indicators associated with achievement goals. For each course material designed for the study, a design pattern table [3, 58, 106] showing rationales and supporting data were included. The following subsections described each instrument used for data collection.

4.4.2.1 Self-reported: Achievement Goal Questionnaire-Revised (AGQ-R)

At the beginning of weeks 1 and 3, all learners were asked to answer the AGQ-R [51] to prospectively self-report if they expected themselves to pursue mastery goals or performance goals. While the survey was originally designed by Elliot and Murayama [51] to differentiate 2×2 goal orientations (i.e., mastery-approach, mastery-avoidance, performance-approach, and performance-avoidance), the survey has also been adopted to measure mastery-performance dimension only to identify goals based on the goal complex theory [144]. There were twelve 5-point Likert-scale questions. The AGQ-R used in the study is attached in appendix B.

4.4.2.2 Self-reported: Motivation questionnaire

Along with the AGQ-R, a questionnaire developed by Vansteenkiste et al. [162] based on Sheldon and Kasser [147] was presented to learners in weeks 1 and 3 to measure learners' autonomous and controlling motivations. Each of the four 5-point Likert-scale items measured the external, introjected, identified, and intrinsic motivations for performance goal pursuit. The external and the introjected questions represented the controlling motivation and the identified and the intrinsic questions indicated the autonomous motivation. The motivation questionnaire revised for the present study is attached in appendix B.

4.4.2.3 Trace: Tip-of-the-week email and Jupyter Notebook

The instructor uploaded a Jupyter Notebook explaining a tip on how to write more efficient Python code in week 2 and week 4. Immediately after an upload, learners received a notification email. While these Jupyter Notebook materials were designed to advance learners' understanding of Python programming for data science, comprehending these materials was neither necessary nor directly relevant to their mandatory weekly assignments, and tip-of-the-week Jupyter Notebooks were designed specifically to attract mastery learners. These materials were clearly marked as optional, and did not give any tangible rewards or benefits for short-term performance such as bonus credit but offered opportunities for the long-term development of Python skills. There were two indicators generated from learners' interaction with tip-of-the-week emails and Jupyter Notebooks: (1) the number of times that they opened notification emails and (2) the number of log data generated while learners using a tip-of-the-week Jupyter Notebook (Table 4.1).

4.4.2.4 Trace: Bonus assignments

Optional bonus weekly assignments provided learners more opportunities to practice skills they picked up from lecture videos. These assignments were expected to attract performance-oriented learners since learners could earn additional credits if they submitted the correct answers. For week 1 bonus assignment, a tool built by Wu et al. [181] was used to provide learners practice questions on regular expressions. The rest of weekly bonus assignments were provided on the Jupyter Notebook. Learners could submit assignments multiple times and every submission was graded through the autograder.

When learners made submissions, the system asked learners if instructors may share their answers with other instructors and learners. This sharing preference question was designed to identify controlling motivations from autonomous motivations. Learners could choose between three options below:

- I do not want my answers shared with others after the assignment deadline.

- You may share my answers with others but only anonymously after the assignment deadline.
- You may share my answers with others but only with credit after the assignment deadline.

Learners who chose to share their answers only with credit could be understood as having a stronger controlling motivation who would like to get recognition from instructors, faculty, and other learners (Table 4.2).

4.4.2.5 Trace: Extra assignments

In addition to the bonus assignments, extra weekly assignments were also provided. Similar to the mandatory assignments and bonus assignments, learners were also able to submit extra assignments as many times as possible and each submission was graded through the autograder. As bonus assignments, learners were provided with two formats of extra assignments: a set of regular expression practice questions in week 1 and Jupyter Notebook assignments in weeks 2, 3, and 4. Unlike the bonus assignments, the extra assignments were designed to teach learners advanced concepts or skills beyond what lecture videos covered without giving additional credits. Thus, the extra assignments were expected to attract mastery-oriented learners (Table 4.3).

4.4.3 Data analysis

As the first step of data analysis, learners who did not submit any of the surveys were dropped from the final dataset during data cleaning (e.g., no responses to the AGQ-R surveys given in week 3). Then, for the main data analysis, data were split into four survey datasets and two trace datasets as follows: the week 1 survey dataset, the week 3 survey dataset, the week 1 no-avoidance survey dataset, the week 3 no-avoidance survey dataset, the week 1-2 trace dataset, and the week 3-4 trace dataset.

4.4.3.1 Survey dataset

One approach to measuring the mastery-performance dimension of achievement goals from the perspective of the goal complex theory is to combine approach items and avoidance items [161]. The first two survey datasets followed this approach. The week 1 survey dataset was composed of all twelve items from the AGQ-R and four items from the motivation survey which were collected in the first week of the course. The second dataset was composed of the same items collected in the third week. In these two datasets, responses to three mastery-approach and three mastery-avoidance items were used as six indicators of an individuals' mastery goal pursuit, and three performance-approach and three performance-avoidance questions were included as six indicators of their performance goal pursuit.

Attribute	Definition	Argument Component
Summary	The more learners (1) actively interacted with a tip-of-the-week Jupyter Notebook and/or (2) opened a notification email of it, the more the learners were learning toward the mastery achievement goal.	-
Focal Knowledge, skills, and abilities	Interacting with course materials with mastery achievement goals.	Claim
Rationale	A tip-of-the-week Jupyter Notebook was an optional activity without any benefits to short-term performance (e.g., mandatory weekly assignment scores). Actively interacting with the emails and/or Jupyter Notebook showed that learners willingly spent time and effort to develop their Python skills for long-term learning outcomes.	Warrant
Additional Knowledge, skills, and abilities	Learners could have opened such notification emails and Jupyter Notebooks simply to check contents (e.g., curiosity). Yet, in this case, the number of them opening the emails and interacting with the Jupyter Notebook would not be high.	Alternative explanation
Potential observation	Learners' interactions such as • opening a notification email from their email inbox. • adding a cell on the Jupyter Notebook. • executing a cell successfully. • executing a cell and receiving an error message. • removing a cell. • changing contents in a cell.	Data
Potential work product	Log data generated every time when • learners actively interacted with a Jupyter Notebook.. • learners opened a tip-of-the-week notification email. The log data were used to form continuous variables per learner showing the number of interactions with each tip-of-the-week Jupyter Notebook and with each notification email.	Data

Table 4.1: A design pattern [3, 58] of tip-of-the-week emails and Jupyter Notebooks

Attribute	Definition	Argument Component
Summary	Through submitting an optional weekly bonus assignment, learners expressed their performance goals. If learners chose to share their assignments with credit to peers and faculty members, the sharing preference showed learners' controlling motivations.	-
Focal Knowledge, skills, and abilities	Interacting with course materials with performance-controlling or -autonomous goal.	Claim
Rationale	Through submitting a bonus assignment, learners showed their interest in earning bonus credits and outperforming other learners. Considering that it does not offer opportunities to practice new skills, it would not be hugely attractive to mastery learners. A preference or sharing answers with credit showed learners' interest to be recognized by faculty members and peers (controlling motivations). Other sharing preferences would show that they were not interested in earning recognition (autonomous motivations).	Warrant
Additional Knowledge, skills, and abilities	Learners could have engaged with the bonus assignments simply to check contents (e.g., curiosity). To focus on learners who showed strong interest in earning credits, the indicator was designed to include only submissions. This indicator is also incapable of detecting learners who sought recognition or other types of rewards from people other than faculty and peers (e.g., parents, colleagues at work).	Alternative explanation.
Potential observation	Bonus assignment submissions with one of the following sharing preferences: • No permission to share their answers. • Permission to share them anonymously. • Permission to share them with credit.	Data
Potential work product	Log data generated per each submission of a bonus assignment where they also expressed their sharing preferences. The log data were used to shape three binary variables per submission showing sharing preferences (e.g., if permission for sharing the week 2 bonus assignment of a particular learner was (1) not given, (2) given only if it was anonymously shared, or (3) given only if it was shared with credit).	Data

Table 4.2: A design pattern [3, 58] of bonus assignments

Attribute	Definition	Argument Component
Summary	Through submitting an optional weekly extra assignment, learners showed their mastery goals.	-
Focal Knowledge, skills, and abilities	Interacting with course materials with performance-controlling or -autonomous goal.	Claim
Rationale	Through submitting an extra assignment, learners showed their interest in developing their long-term skills and knowledge which are not directly relevant to short-term performance. Considering that this optional bonus assignment does not provide any additional credit or other benefits to short-term performance such as higher scores on weekly mandatory assignments or the final letter grade, extra assignments would not be hugely attractive to performance learners.	Warrant
Additional Knowledge, skills, and abilities	Learners could have engaged with the bonus assignments simply to check contents (e.g., curiosity). To focus on learners who committed to developing their skills, the indicator was designed to include only submissions, not any other engagements (e.g., opening any of assignments and writing code).	Alternative explanation
Potential observation	Extra assignment submission status.	Data
Potential work product	Log data was generated every time learners made a submission of an extra assignment. The log data were used to form binary variables each of which showing learners' submission status (e.g., a binary variable on if a learner submitted a week 3 extra assignment or not)	Data

Table 4.3: A design pattern [3, 58] of extra assignments

On the other hand, there were other studies [15, 16, 127, 144] where an individual's mastery goal and performance goal was respectively measured only through mastery-approach items and performance-approach items without any mastery-avoidance and performance-avoidance items. Thus, the last two survey datasets were only composed of ten items: the approach items from the AGQ-R as well as the four motivation items. In addition to these four survey datasets, there were two trace datasets each of which is based on the first two weeks or the last two weeks of the 4-week long course.

4.4.3.2 CFA

A CFA was conducted independently on each of these six datasets. A three-factor model as shown in Figure 4.2 was chosen as most empirical evidence from previous work has shown that autonomous motivation and controlling motivation could be modeled together [85, 130]. When a CFA solution was not identified, Exploratory Factor Analysis (EFA) was conducted to identify a solution. Schmitt [137] recommended this follow-up EFA to explore other possible solutions when the expected theory does not fit the data. In this particular study, such a follow-up EFA can provide insights on how different (1) the goal complex model and (2) associations between observed are. CFAs and EFAs were conducted using Mplus 8.6 [113] and MplusAutomation R package [66]. All the programming code used in this study is attached in Appendix A.

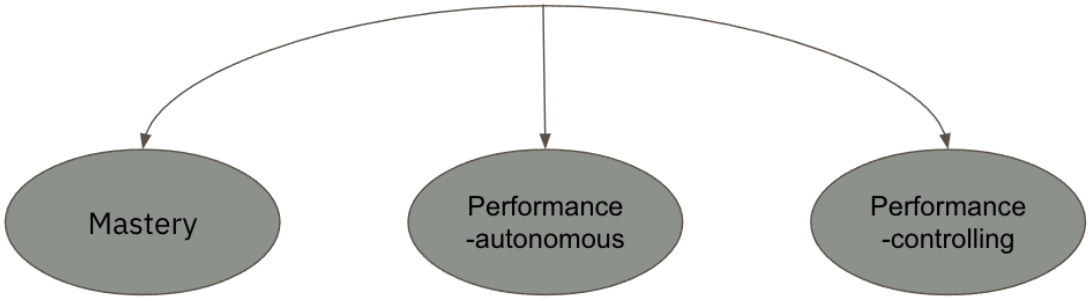

Figure 4.2: A three-factor CFA model with the premise that autonomous motivation and controlling motivation can be simultaneously endorsed together

While there were various suggestions on how to report and understand the fit of a model, the present study follow the recommendations from Brown and colleagues [27, 28] in terms of reporting and comparing a fit of each solution: reporting not only (1) goodness-of-fit indices evaluating a global fit between a dataset and a model but also (2) modification indices which show fits of localized areas and (3) the theoretical interpretability of the parameter estimates.

4.4.3.3 Quantifying fit significance

Goodness-of-fit indices, χ^2 and χ^2 p-value (absolute fit), Root Mean Square Error of Approximation (RMSEA) (parsimony correction), Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI) (comparative fit) were used as criteria to assess overall fits between each dataset and CFA solutions based on the goal complex model. These fit indices assess either how far an examined model is from a model perfectly following data (RMSEA, χ^2, and χ^2 p-value) or from the worst model (CFI and TLI) [182].

Modification indices represent to what extent χ^2 values could be decreased if a particular path between an indicator and a factor is added. A modification index equal to or greater than 3.84 has been commonly used as a generally agreed cutoff point which corresponds to 1 degree of freedom at $p < 0.05$. To avoid misusing the $p < 0.05$ as a dichotomous cutoff of statistical significance, the entire modification indices are reported as continuous values. Furthermore, instead of simply checking the existence of any values higher than 3.84, the overall number or the distribution of the values were checked.

Only when a solution shows good overall goodness-of-fit and modification indices, were parameter estimates of the solution evaluated. In this third step of interpreting parameter estimates, factors and indicators were inspected for standardized factor correlation no greater than 1.00, negative factor variances, and negative indicator error variances. Furthermore, it was evaluated if factor loadings were theoretically interpretable.

It is important to explain why the present study did not solely rely on particular cutoff points of goodness-of-fit indices to indicate an acceptable fit between each dataset and a model. Hu and Bentler [72] originally suggested cutoff values for each fit index, such as 0.06 or below for RMSEA, with a note of caution that these values should not be considered as a universal rule across different contexts including sample sizes and estimators. Yet, their suggestions have been wrongly adopted as 'Golden rules' to judge acceptable fits without consideration of such contexts, which is similar to how the p-value became misunderstood as a dichotomous cutoff concluding true or false of null hypotheses (See the statement of American Statistical Association (ASA) on p-value [167]). Multiple other papers have also questioned the reliability of a universal cutoff point for those fit indices without considering which estimator was used, a degree of freedom, and a sample size [30, 81, 100, 154]. Regarding this issue, Steiger [150] recommended that researchers report confidence intervals of a fit index since confidence intervals are associated with a statistical rationale of sample size; the bigger the sample size is, the more precise the confidence intervals become. Brown and colleague [27, 28] proposed to report multiple fit indices since each of them provides different information about model fits, from absolute fit to comparative fit, and therefore reporting them together provides a more conservative and reliable way to assess fits. Thus, the CFA models in this study were estimated with multiple goodness-of-fit indices including confidence intervals of

RMSEA as well as modification indices for localized fits, and theoretical interpretability instead of heavily relying on a couple of cutoff points of goodness-of-fit indices.

Considering that both survey data and trace data included categorical or binary factor indicators, the Weighted Least Square Mean and Variance adjusted (WLSMV) estimator was used to compute estimates such as factor loadings [28, 112]. Goals of a CFA model is to obtain these estimates of each indicator and latent variable which produce a predicted variance-covariance matrix that is similar to the input variance-covariance matrix as much as possible. To achieve this goal, estimators improve an initial set of parameter estimates (i.e., starting values in Mplus [113]) throughout iterative processes, and thus choosing the right estimator is important to obtain a set of accurate estimates.

4.5 Results

4.5.1 Data overview

The total of 193 learners were considered after removing those who missed answering any surveys conducted in in weeks 1 and 3. This decision was made since an absence of complete responses to a survey from a learner as missing data could lower the statistical power of analyses. Among 193 learners, 136 learners were from the first iteration and 57 learners were from the second iteration of the course. Table 4.5 shows a descriptive statistic summary of the week 1 survey and the week 3 survey data collected.

Table 4.6 displays a summary of descriptive statistics for survey and trace data collected. Trace indicators which showed observations lower than three were removed from final datasets to avoid CFA solutions not being identified. In this removal process, the following indicators were removed: (1) indicators representing learners' disagreement to share their bonus assignments for week 1, week 3, and week 4 and (2) indicators representing learners' submissions of extra assignments for week 2 and week 4. Trace indicators included in the final dataset are shown in Table 4.4.

Source	Description	Theoretical Relationship	Specific Indicator Variable
Tip-of-the-week email	(continuous) How many times learners opened a tip-of-the-week notification email in the Nth week	Mastery	email2_count email4_count
Tip-of-the-week Jupyter Notebook	(continuous) How many log data learners generated while interacting with a tip-of-the-week Jupyter Notebook in the Nth week	Mastery	notebook2_count notebook4_count
Bonus assignment	(binary) If learners submitted the Nth-week bonus assignment and did not want to share their answers with faculty or other students	Performance-autonomous	bonus2_no_sharing
Bonus assignment	(binary) If learners submitted the Nth-week bonus assignment and wanted to anonymously share their answers	Performance-autonomous	bonus1_anonymous bonus2_anonymous bonus3_anonymous bonus4_anonymous
Bonus assignment	(binary) If learners submitted the Nth-week bonus assignment and wanted to share their answer with credit	Performance-controlling	bonus1_credit bonus2_credit bonus3_credit bonus4_credit
Extra assignment	(binary) If learners made at least one submission for the extra weekly assignment of the Nth week, which did not give bonus points	Mastery	extra1_submit extra3_submit

Table 4.4: The trace indicators included in the final datasets

		Mastery						Performance									
		Approach			Avoidance			Approach			Avoidance			Autonomous		Controlling	
		mp1	mp2	mp3	mv1	mv2	mv3	pp1	pp2	pp3	pv1	pv2	pv3	at1	at2	ct1	ct2
W1	$\bar{x}$	4.709	4.750	4.715	4.062	4.502	4.036	3.932	3.859	3.625	3.616	3.668	3.651	3.659	4.026	2.363	3.322
	σ	0.576	0.511	0.527	1.116	0.757	1.089	1.015	0.984	1.020	1.162	1.057	1.091	1.078	1.055	1.136	1.219
W3	$\bar{x}$	4.362	4.326	4.497	3.927	4.227	3.989	3.673	3.652	3.388	3.455	3.518	3.479	3.647	3.963	2.364	3.354
	σ	0.920	0.885	0.743	1.170	0.912	1.065	1.056	1.084	1.065	1.207	1.141	1.214	1.132	1.084	1.154	1.219

Note. $\bar{x}$ = mean, σ = standard deviation.

Table 4.5: A descriptive statistic summary of survey data

		Mastery				Performance-autonomous				Performance-controlling	
		email2_count	notebook2_count	extra1_submit	extra2_submit	bonus1_no_sharing	bonus1_anonymous	bonus2_no_sharing	bonus2_anonymous	bonus1_credit	bonus2_credit
W1	$\bar{x}$	1.673	6.683	0.974	1.000	0.005	0.181	0.015	0.062	0.160	0.041
	σ	1.668	18.545	0.159	0.000	0.071	0.386	0.124	0.242	0.368	0.199
		email4_count	notebook4_count	extra3_submit	extra4_submit	bonus3_no_sharing	bonus3_anonymous	bonus4_no_sharing	bonus4_anonymous	bonus3_credit	bonus4_credit
W3	$\bar{x}$	1.414	2.041	0.927	0.994	0.010	0.108	0.010	0.093	0.145	0.093
	σ	1.452	7.300	0.260	0.071	0.101	0.312	0.101	0.291	0.353	0.291

Note. $\bar{x}$ = mean, σ = standard deviation.

Table 4.6: A descriptive statistic summary of trace data. Variable names and types in the columns are equal to the variable names in the row index of Table 4.4.

4.5.2 Findings

Two RQs were posed for the present study as follows:

- RQ1. How well do self-reported survey data fit the goal complex model?
- RQ2. How well do trace data fit the goal complex model?

Regarding the RQ1, all CFA solutions except one based on week 1 no-avoidance survey datasets were identified in the first iteration of CFAs. Therefore, a follow-up EFA was conducted [137] (Table 4.8). In terms of RQ2, both solutions built on each trace dataset were identified.

4.5.2.1 RQ1: Investigation on survey-based solutions

As the first step to assess fits between the goal complex model and each dataset, the goodness-of-fit indices were computed (Table 4.7). In general, the survey-based solutions showed worse fits than trace-based solutions. χ^2 values of survey-based solutions were larger than those of trace-based solutions. χ^2 p-values also showed that the survey data were detectably different from the data perfectly fitting the model. Both RMSEA values and confidence intervals of survey-based solutions were also reasonably far away from the conventional threshold of 0.06.

For the unidentified solution for the week 1 no-avoidance survey dataset, EFA was conducted to further understand the associations between the survey data and three factors (Table 4.8). Geomin rotation was used to compute factor loadings. Factor loadings showed that the major difference between the goal complex model and the observed associations was how items were related to performance goals. Factor 2 had strong positive associations with all performance-approach items and performance-autonomous items. On the other hand, factor 3 which was positively related with both performance-controlling items did not show any association with performance-approach items. That is, none of the performance-approach items contributed to shape performance-controlling factor.

Model	Parameters	χ^2	df	χ^2 p-value	RMSEA	[lower CI, upper CI]	CFI	TLI
Survey (week 1)	85	266.303	95	0.000	0.097	[0.083, 0.110]	0.977	0.971
Survey (week 3)	88	354.192	95	0.000	0.119	[0.106, 0.132]	0.966	0.957
Survey, no avoidance (week1)	NA	NA	NA	NA	NA	NA	NA	NA
Survey, no avoidance (week3)	55	86.253	29	0.000	0.101	[0.077, 0.126]	0.995	0.993
Trace (week1-2)	21	23.955	17	0.120	0.046	[0.000, 0.086]	0.881	0.805
Trace (week3-4)	19	12.733	11	0.311	0.029	[0.000, 0.084]	0.997	0.995

Note 1. Lower CI = lower bound of 90% confidence interval, Upper CI = upper bound of 90% confidence interval.

Note 2. Fit indices for the solution for the week 1 no-avoidance survey dataset were reported as NA since the solution was not identified.

Table 4.7: The goodness-of-fit indices of solutions for each dataset

Indicators	Factor 1	Factor 2	Factor 3
mastery-approach1	0.582 *	0.118	−0.034
mastery-approach2	0.886 *	0.000	0.004
mastery-approach3	0.904 *	−0.006	0.054
performance-approach1	−0.002	0.889 *	−0.072
performance-approach2	0.013	0.971 *	0.001
performance-approach3	−0.019	0.902 *	0.081
performance-autonomous1	0.107	0.465 *	−0.058
performance-autonomous2	0.022	0.682 *	0.053
performance-controlling1	0.006	−0.008	0.773 *
performance-controlling2	−0.003	0.256 *	0.551 *

Note. Values significant at 5% level appeared with *.

Table 4.8: EFA loadings of the week1 no-avoidance survey data solution. Geomin rotation was used to compute loadings.

Modification indices were statistical criteria also used in the study to evaluate how much fit could be improved when a particular connection (1) between an indicator and a factor or (2) between indicators is newly established. A modification index equal to or greater than 3.84, which corresponds to 1 degree of freedom at $p < 0.05$, was used as a criteria to tell if a index is meaningfully large. The localized fits measured by modification indices were aligned with the findings of the goodness-of-fit values.

For survey datasets, there was no modification index computed between indicators, which showed that indicators did not show any correlation which could have improved a fit but was not included in the models (See Appendix C for Table C.1, Table C.2, and Table C.3 showing modification indices). Survey-based solutions generally reported poor fits. Every survey-based model showed several modification indices between factors and indicators larger than 3.84: 34.6% of indices in the week 1 survey model, 26.9% of indices in the week 3 survey model, and 23.5% of indices in the week 3 no-avoidance model. Not only the number of these large values of survey-based solutions but also their values themselves were generally large. Thus, modification indices re-confirmed the overall poor fit between the survey datasets and the goal complex model. Due to these results, theoretical interpretability was not investigated.

4.5.2.2 RQ2: Investigation on trace-based solutions

The trace-based solutions reported generally acceptable goodness-of-fit values. RMSEA values, and confidence intervals were also lower than or near 0.06. Yet, the solution for the week 1-2 trace data reported the lowest CFI and TLI values (Table 4.7).

In terms of localized fit estimation shown by modification indices, the week 1-2 trace dataset

only showed 10.7% of indices which were larger than 3.84 cutoff. The week 3-4 trace dataset did not report any numbers larger than 3.84.

Only trace-based solutions were examined for their interpretability since only these showed acceptable global and localized fits. Both solutions showed neither standardized factor correlation greater than 1.000 nor negative indicator error variances. Since all the factor variances were fixed at 1.000, there were also no negative factor variances.

Theoretical interpretability of factor loadings from the week 1-2 trace datasets is mixed (Table 4.9). Except for the negative factor loading between the mastery goal factor and the extra2_submit indicator, which shows that mastery learners tended not to submit extra assignments in week 2, other indicators showed strong associations with each factor toward the expected direction. Yet, trace indicators for performance-autonomous and performance-controlling showed mixed results. Performance-autonomous reported negative factor loadings for learners' preference over anonymously sharing their bonus assignment answers, which was opposite to what was expected. Furthermore, a factor loading of performance-controlling with the bonus1_credit indicator was not strong, and the bonus2_credit indicator was negatively associated with the factor which was different from the original expectation built upon the goal complex theory.

On the other hand, the theoretical interpretability of factor loadings of the week 3-4 trace datasets generally followed the goal complex theory (Table 4.10). Mastery goal showed positive factor loadings for notebook4_count email4_count, and extra3_submit, which all showed mastery learners' engagement with materials designed for them. Performance-autonomous goal factor showed positive loadings of bonus3_anonymous and bonus4_anonymous which show that they do not mind not getting credit for sharing their answers. The performance-controlling factor was positively loaded with bonus3_credit and bonus4_credit. That is, the factor was explained by learners' tendency of submitting bonus assignments and sharing answers of the assignments with credit, which are indicators designed to capture the performance-controlling goal factor depicted in the goal complex theory.

4.6 Discussion and future work

Results showed that only the week 3-4 trace-based solution showed a good global, local, and theoretical fit between the theoretical goal complex model [51] and the observed data. While week 1-2 trace-based solution also reported a reasonable global and local fit, factor loadings were not easily explainable with the goal complex theory. All four survey-based solutions did not show acceptable global and local fits according to the commonly used cutoff criteria [182].

Surveys have been traditionally and commonly used to measure achievement goals across multiple achievement goal theories, however, this study shows poor fit for all four survey-based solu-

Indicators	Mastery	Performance -autonomous	Performance -controlling
notebook2_count	1.062 (0.004)		
email2_count	0.354 (0.008)		
extra1_no_submit	-0.234 (0.041)		
bonus1_no_sharing		0.278 (0.038)	
bonus1_anonymous		-0.429 (0.001)	
bonus2_anonymous		-0.441 (0.000)	
bonus1_credit			0.364 (0.217)
bonus2_credit			-0.690 (0.000)

Note. Factor loadings (p-value).

Table 4.9: Standardized estimates of the solution for week 1-2 trace data. Standardization was conducted through STDYX standardization computed by Mplus.

Indicators	Mastery	Performance -autonomous	Performance -controlling
notebook4_count	0.174 (0.027)		
email4_count	0.213 (0.006)		
extra3_no_submit	0.787 (0.000)		
bonus3_anonymous		0.995 (0.000)	
bonus4_anonymous		0.935 (0.000)	
bonus3_credit			1.097 (0.000)
bonus4_credit			0.840 (0.000)

Note. Factor loadings (p-value).

Table 4.10: Standardized estimates of the solution for week 3-4 trace data. Standardization was conducted through STDYX standardization computed by Mplus.

tions. It is important to note that these findings are bound to the three-factor model used in this study. There is a possibility that surveys might have shown better results with different models such as a four-factor or a two-level model. The trace-based solutions, especially one based on the week 3-4 trace dataset, showed not only good fit but also strong theoretical interpretability.

There is also significant evidence to support the argument that the AGQ-R and motivation survey did not sufficiently measure learners' achievement goals of learners in this study context. Modification indices of survey-based solutions showed that unexpected connections between each goal factor and survey item could increase fits, and these connections are seemingly random from the perspective of the goal complex theory. For example, a fit of the solution for the week 1 survey dataset could be improved if the following connections are made: (1) mastery goal factor and performance-controlling item indicator, (2) performance-autonomous goal factor and mastery-approach, mastery-avoidance, and performance-controlling item indicators, and (3) performance-

controlling goal factor and mastery-approach and mastery-avoidance item indicators. These connection suggestions are not congruent with the goal complex theory.

There could be multiple reasons for these unexpected poor fits of survey-based solutions. One could be due to the low accuracy of learners' responses to the prospective survey in terms of predicting their future goal-relevant behaviors. In this study, learners had to answer what their goals were going to be for the next few weeks, which would be more difficult than predicting their goals for a one-time task. In addition, because the course was the first of many in the master's program, participants did not have a robust body of expectations. Thus, their predictions might be less precise; as the difficulty of assignment tasks, the time commitment required, and the strategies to use may be unclear to the learners. Another compatible explanation is that learners did not give candid responses due to social desirability. The survey was introduced by instructors and therefore learners might have been concerned that the instructor would judge them based on their responses to the surveys. Therefore, they might have endorsed most of the questions to show 'the best version of themselves.'

The week 1-2 trace-based solution showed a negative correlation between the mastery goal factor and the extra1_submit, which contradicts the expected positive correlation. This does not necessarily force a rejection of the goal complex theory, but instead might suggest expansion of the theory to include the effect of time pressure on mastery learners. For instance, Beck and Schmidt [13] have found that time pressure led more learners to state their goals as performance-oriented and led fewer learners toward mastery-oriented goals. When the students perceived that they were under time constraints (e.g., near the final exam), strategies that might have facilitated long-term skill development yet required more time invested were utilized less. Instead, learners might have prioritized strategies which could increase their short-term performance and hence help them pass the course or earn good grades. This might explain why mastery learners in this study tended not to submit extra assignments; These mastery learners could have temporarily chosen to be performance-focused by not engaging with optional advanced materials to save time for themselves to safely pass the course. In this study, optional extra assignments shared weekly deadlines with mandatory assignments and learners could have prioritized mandatory assignments over completing the extra assignments to pass the course. The tip-of-the-week Jupyter Notebooks, which were also designed to attract mastery-oriented learners with opportunities to learn advanced knowledge, did show a high positive loading to the mastery goal factor, perhaps because they were comparably less burdensome tasks. Learners did not have to work on any exercises or questions but simply read the notebook and run cells as needed.

This finding has two implications. First, if the aim is to design tasks attractive to mastery learners, it may be important to give learners flexibility in time to explore the task. Giving them deadline-free experiences such as the tip-of-the-week Jupyter Notebook used in this study would be

one option to encourage mastery learners. To generalize this finding, it will be important to conduct replication studies on different contexts or with different deadline-free experience designs.

Secondly, caution must be exercised as researchers try to infer mastery learners' context-specific behaviors using surveys measuring achievement goals, specifically with the AGQ-R [51] and motivation survey [162]. In an online or residential classroom, it would be rare to see purely mastery-oriented learners who do not mind failing the course as long as they can develop their knowledge and skills. Even if learners lean towards mastery achievement goals, they might still show some performance-oriented behaviors to avoid risks such as failing their course or having trouble with earning their degree. If mastery learners occasionally adopt performance-oriented strategies and these goals are brief and more temporally situated, it is hard to capture through survey questions measuring learners' general behaviors. To extend understanding of learners' temporal goal changes or multiple-goal pursuit, researchers should closely follow learners' goal changes more frequently, to which log data or clickstream could be useful.

The solution for the week 1-2 trace dataset also showed loadings representing associations between performance-autonomous and performance-controlling goal factors and their indicators are not easily explainable with the goal complex theory. Most of the learners in the performance-autonomous either did not submit bonus assignments or chose to share assignments with credit. Furthermore, the performance-controlling learner group in weeks 1-2 was distinguished from other learners by either not submitting week 2 bonus assignments or preferring to anonymously share assignments. These factor loadings of performance-controlling and performance-autonomous are contradictory to the goal complex theory. This comparably low theoretical interpretability is sharply contrasted with the factor loadings of the week 3-4 trace dataset. The solutions for the two datasets shares similar types of indicators, but their main differences are when these datasets were collected. Learners might have required some time to become familiar with various course materials and evaluate their skills so that they could set up and follows realistic goals. Thus, during the first couple of weeks, their goals might not have been stable. To confirm this possible effect of time factor, future study is necessary.

CHAPTER 5

Generating SRL theory through data

5.1 Introduction

The previous chapter examined how well survey data and trace data indicated online learners' achievement goals. Field study results with the data composed of responses to the Achievement Goal Questionnaire-Revised (AGQ-R) [51] and to the survey measuring controlling-autonomous motivations [162]. Results showed that Confirmatory Factor Analysis (CFA) solutions based on the goal complex model did not fit survey data globally and locally. On the other hand, solutions based on the same model showed better global and local fits when using trace data. In particular, trace data collected in the last two weeks of a course (i.e., weeks 3 and 4), after learners may have had time to experience and acclimate to the course and its circumstances, showed strong theoretical alignment with the goal complex model tested.

A natural next step would be investigating causes of the differences through scrutinizing how well each dataset indicates information relevant to achievement goals. In such an investigation, the focus would be analyzing data with less reliance on a previous theory. Through this data-driven approach, several types of insights might be drawn ranging from how to modify the survey and trace indicators to how to further expand and articulate the goal complex theory.

5.1.1 Variable-centered approach versus person-centered approach

Studies in learning analytics often apply one of two methodological approaches – variable-centered or person-centered – to data analysis, each of which answers different types of research questions. The variable-centered approach focuses on identifying associations among variables which can characterize the entire dataset. It identifies predictor-outcome associations and builds a predictive model to answer questions regarding the influence of predictor variables on outcome variables. Frequently used statistical methods in this category are correlations, regression analysis, and structural equation modeling, including CFA [71, 91, 124]. Researchers who apply this approach often aim to understand the entire dataset in terms of these associations. For example, in the previous

chapter, CFAs were conducted to examine whether and how well a dataset fits a particular theory by examining the relationship between indicator variables and latent variables. In this process of identifying associations among variables, it is assumed that a dataset is composed of relatively homogeneous individuals – homogeneous because they are a sample from a well-defined population – who can all be represented by the same set of parameters explaining associations [71, 91].

A person-centered approach could give further insights on the composition of learners' survey responses and behavioral trace data. A person-centered approach is often applied to differentiate individuals. Clustering and mixture modeling are examples of person-centered approaches. These approaches assume that individuals are heterogeneous enough to be differentiated into groups based on shared attributes [71, 91, 124]. In this particular study's context, the person-centered approach can complement the variable-centered approach by more precisely representing multiple-goal pursuits and suggest how heterogeneous learners can be grouped by similar patterns of pursuing achievement goals. In this chapter, I will employ latent variable mixture modeling, a person-centered approach, to enrich understanding of different goals stated in surveys and goal-relevant behaviors captured through trace data.

5.1.2 Latent variable mixture modeling

Among person-centered approaches, latent variable mixture modeling has consistently garnered the interest of researchers across fields. Latent Variable Mixture Modeling is a statistical technique that reveals clusters of sample populations using indicators [111, 120, 138]. Latent variable mixture modeling is subdivided into Latent Profile Analysis (LPA) and Latent Class Analysis (LCA) depending on which data type indicators form latent variables. LPA uses continuous data while LCA uses categorical data. Yet, because latent variable mixture modeling often includes both continuous and categorical data as indicators, Pastor et al. [124] pointed out that such distinction might be unnecessary. Unlike clustering, latent variable mixture modeling does not require transforming indicators measured in different scales or exhibiting different degrees of variances to match scales for analysis [124]. This is an especially clear advantage for the present study which includes a combination of continuous variables on a large scale and categorical data on a smaller scale. Furthermore, latent variable mixture modeling has a set of more rigorous criteria, such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), than clustering. These criteria are used to evaluate solution fits and decide which solution best represents features in the given dataset best [124, 138].

Some previous studies on achievement goals applied latent variable mixture modeling on survey datasets in order to understand relationships between clusters and various learning outcomes such as task performance and motivations. Pastor et al. [124] aimed to compare LCA solu-

tions on different subset of college students' responses to the Achievement Goal Questionnaire (AGQ) [50]. The authors examined LCA solutions on (1) a dataset only with mastery-approach and performance-approach question responses, (2) a dataset with mastery-approach, performance-approach, and performance-avoidance questions responses, and (3) the full dataset with mastery-approach, mastery-avoidance, performance-approach, and performance-avoidance question responses. Then using cluster outputs from the final solutions, they examined correlations between each cluster and students' motivation and performance. Zhang et al. [184] investigated correlations between German elementary students' multiple goal pursuit and learning outcomes such as motivation, test anxiety, and academic performances by running LCA on survey data. They identified three patterns of multiple goal pursuit: (1) high mastery, (2) low mastery, and (3) high in both mastery and performance-approach. Schwinger and Wild [140] conducted LCA on a five years long longitudinal data. They reported changes in multiple-goal pursuits of students and analyzed relationships between these goal pursuits and learning outcomes.

The focus of the current study is to compare self-reported survey datasets and trace datasets. It was motivated by Chapter 4, where CFA results questioned the correspondence of survey data to trace data in measuring achievement goals of learners taking an online degree program course. Thus, to examine patterns of cluster outputs of each dataset, I shaped one of the first studies conducting latent variable mixture modeling on trace datasets as well as survey datasets. That is, the present study has contributed in three ways: (1) applying latent variable mixture modeling based on goal complex theory (2) on a unique sample population (3) to directly estimate the difference in the information collected by different measures.

5.2 Research Questions

The following research questions were posed in order to further examine the collected data described in Chapter 4.

- RQ1. What goal clusters can be identified from survey data?
- RQ2. What goal clusters can be identified from trace data?
- RQ3. How are goal survey-based goal clusters and trace-based goal clusters related?

RQ1 and RQ2 were posed to compare and contrast the difference of achievement-goal-relevant information captured by survey and trace indicators. Answering RQ1 and RQ2 could show patterns of learners' survey responses and behavioral trace data and enrich understanding of how these responses and behavioral data are different. To answer RQ1 and RQ2, latent variable mixture

modeling was applied to each survey and trace dataset and then parsimonious clusters of learners were identified.

RQ3 is an exploratory research question to deepen understanding of relationship between survey responses and trace indicators through further probes of findings from RQ1 and RQ2. Once differences or similarities between survey and trace datasets are captured, it is natural to question why such differences or similarities were observed. For instance, a learner might have been identified as a mastery learner according to a survey dataset but not as one based on a trace dataset. Such misalignment could explain why the differences or similarities between survey and trace datasets happened.

5.3 Methods

5.3.1 Study context

Survey and trace data were collected during two iterations of an introductory data science course offered in September, 2021 (151 enrollments) and January, 2022 (98 enrollments). This course was the first technical course for an online applied data science Master's degree program at the University of Michigan, School of Information. The course had multiple distinctive characteristics from traditional residential college courses and Massive Open Online Courses (MOOCs). This online course was different from MOOCs in that it was a full tuition credit-bearing course which led to a degree pathway, and enrollment was limited. Furthermore, student-instructor ratio is in line with degree granting program than MOOCs with a plenty of individual synchronous office hours. Compared to traditional residential college courses, learners of this course were more diverse in ages, background knowledge, and level of education, and as many of these learners were employed and had parental responsibilities and had come back to the university degree program after a while to be part-time students.

In this course, learners had to submit four weekly mandatory assignments and each of these assignments was worth 25% of the full credit. Furthermore, learners could earn additional credit by submitting bonus assignments. The top letter grade, $A+$, was only awarded to students who submitted one or more bonus assignments and earned 100% in every mandatory assignments. Extra assignments did not give learners any additional points but provided them with opportunities to learn skills or concepts beyond course. Learners could submit any of these assignments as many times as they wanted until the deadline, and their submissions were graded through automatic code grading system called 'autograder' which would run unit tests on learners' submissions. Figure 5.1 shows the overall timeline of the course.

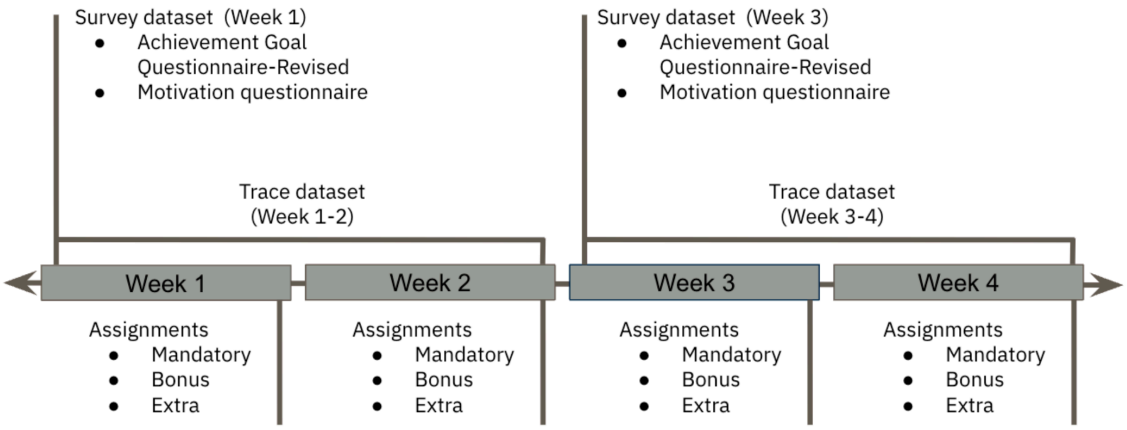

Figure 5.1: A course timeline showing when surveys and assignments were provided to learners. Trace data were collected throughout the course and were broken into two datasets based on when the data were collected.

5.3.2 Data

5.3.2.1 Measurements and indicators

Two questionnaires were employed to collect self-reported achievement goals: Achievement Goal Questionnaire (AGQ-R) [51] and motivation survey data [162]. Multiple previous research combined these two questionnaires to measure self-reported achievement goals from the perspective of goal complex theory [15, 16, 127, 144, 161]. Both questionnaires were shown to learners at the beginning of weeks 1 and 3, as learners' goals could change after the first half of the course.

To design a set of trace indicators, the domain modeling step of the Evidence-Centered Design (ECD) framework was implemented. It is a useful way to identify potential under- and over-representation of target constructs [14, 105, 106]. The results of the framework implementation are reported through 'the design pattern' tables [3, 58].

5.3.2.2 Self-reported: Achievement Goal Questionnaire-Revised (AGQ-R)

Achievement Goal Questionnaire-Revised (AGQ-R) was designed and examined by Elliot and Murayama [51] using American undergraduate students' self-reported achievement goals on an exam for their college course. This questionnaire is composed of twelve 5-point Likert-scale items: four mastery-approach items, four mastery-avoidance items, four performance-approach items, and four performance-avoidance items. In this study, this questionnaire was shown to learners to measure their mastery and performance goals. The AGQ-R used for the study is attached in appendix B.

5.3.2.3 Self-reported: Motivation questionnaire

Motivation questionnaire designed by Vansteenkiste et al. [162] based on Sheldon and Kasser [147]. This questionnaire is composed of four 5-point Likert-scale items where each of these items represents the external, introjected, identified, and intrinsic motivations for performance goal pursuit. In this study, the external and introjected items were used to as two indicators for performance-controlling goals and the identified and intrinsic items were used to measure performance-autonomous goals.

5.3.2.4 Trace: Sharing preference on bonus assignments

Optional weekly bonus assignments provided learners opportunities to earn additional credit which were designed to attract performance learners to whom normative standards such as the final letter grade matters. For week 1 bonus assignment, tasks were given on a specific tool for practicing regular expression [181]. The rest of weekly bonus assignments were Python programming assignments on the Jupyter Notebook.

Learners were also asked to clarify if and how they would like to share their answers with faculty members and learners. Regarding the sharing preference, learners had three options as follows:

- I do not want my answers shared with others after the assignment deadline.
- You may share my answers with others but only anonymously after the assignment deadline.
- You may share my answers with others but only with credit after the assignment deadline.

Considering that performance-controlling learners were likely to seek external approval or recognition from peers and faculty of the degree program, the third option was used as an indicator for performance-controlling goal. On the other hand, the first two options represented performance-autonomous goal. More specific rationale and limitations of the trace indicators are presented on Table 4.2.

5.3.2.5 Trace: Extra assignment submission

Optional weekly extra assignments were designed to attract mastery learners. These assignments did not give any additional credit but provided tasks from which learners can learn new skills beyond other course materials. As the bonus assignments, the first week extra assignments were presented on the tool for practicing regular expression [181] while the rest of the extra assignments were provided on the Jupyter Notebook. Table 4.3 shows more detailed rationale and limitations of the trace indicator design.

5.3.2.6 Trace: Tip-of-the-week email and Jupyter Notebook

In weeks 2 and 4, the course instructor released a tip-of-the-week Jupyter Notebook and sent a notification email about the release. Through these tip-of-the-week Jupyter Notebooks, instructors explained how to write more efficient and readable Python code. The main difference between it and the optional assignments explained above was an absence of particular tasks or deadline for this learning material. Table 4.1 presented more detailed rationale and limitations of the trace indicator design.

5.3.2.7 Trace: Interactions with bonus and extra assignments

From the previous use of the data for other study, it was observed that some learners interacted with bonus and extra assignments yet did not submit their assignments. Some of these incomplete assignment submission might be indicators of achievement goals. For example, learners might have not completed and submitted assignments for time constraint or high perceived difficulty of assignments despite a motivations to earn additional credits or learn advanced concepts. Yet, not all of incomplete submissions would be related to achievement goals. Other learners could also have simply browsed the assignments out of curiosity and did not have serious intention to complete these optional assignments. To consider these different motivations behind incomplete assignment submissions, the number of event log data were counted. This indicator design was based on the rationale that the stronger learners' motivation was to engage with assignments, the more active engagement with assignments would have occurred. Each of these interactions was logged at the telemetry system embedded in the the course. More details of the indicator design process is presented as the design template on Table 5.1.

5.3.2.8 Trace: Additional submissions of mandatory assignments

It was also observed that some learners submitted their mandatory or optional assignments again even after they got 100% on their assignment submission. The additional submission did not add any extra points to their final grade and it seemed that they experimented with alternative ways of coding to get the correct answer. The behavioral pattern was used as another indicator of mastery goal in this study. The indicator was designed to be a count variable which represent the number of assignments learners made additional submissions to even after reaching to the 100%. For example, if a learner made additional submissions for the week 1 mandatory assignment, the week 4 mandatory assignment, the week 1 bonus assignment, and the week 2 extra assignment, the indicator value for the learner was 4. More details of the indicator design process is given on Table 5.2.

Attribute	Definition	Argument Component
Summary	Through actively interacting with an optional weekly assignment, learners expressed their mastery or performance goals.	-
Focal Knowledge, skills, and abilities	Interacting with optional course materials which matches learners' achievement goals.	Claim
Rationale	Through including how actively learners engaged with an optional assignment, It was possible to capture behaviors relevant to achievement goals of learners who were interested in but could not complete assignments for time constraint or other challenges. Through counting the number of log data representing their active engagement, this approach was expected to distinguish these learners from other learners who simply took a look at contents out of curiosity without serious intention to complete assignments.	Warrant
Additional Knowledge, skills, and abilities	Some learners might have generated fewer log data with outstanding skills and knowledge instead of lacking serious intentions. Yet, in this case, most of these learners might have easily submitted their assignments, and there is an indicator capturing if they submitted or not.	Alternative explanation
Potential observation	Learners' interactions such as • opening a notification email from their email inbox, • adding a cell on the Jupyter Notebook, • executing a cell successfully, • executing a cell and receiving an error message. • removing a cell, • changing contents in a cell.	Data
Potential work product	Log data was generated every time learners actively engaged with optional assignments. The log data were used to form continuous variables.	Data

Table 5.1: A design pattern [3, 58] of interactions with optional assignments

Attribute	Definition	Argument Component
Summary	Log data of learners making additional attempts on assignments after receiving the highest possible score were used to shape a count variable form of indicator representing their mastery goal pursuit.	-
Focal Knowledge, skills, and abilities	Submitting assignments even after received 100% may be evidence that they pursued mastery goals.	Claim
Rationale	Additional submission after reaching 100% on one's assignment did not give any extra points. Yet, some learners still experimented with their answers to seek alternative or even more efficient way for answering given task. Log data showing such additional submissions could be used to form an indicator showing learners' mastery goal pursuit. In particular, through counting which assignment they made additional submissions to, it was expected to form an indicator showing how strongly these learners pursued mastery achievement goal.	Warrant
Additional Knowledge, skills, and abilities	A learner could have made such additional submissions simply to play with code which did not involve serious intention to develop their knowledge or understanding. Unfortunately, the present indicator could not identify their precise intention behind additional submissions.	Alternative explanation
Potential observation	If each learner made one or more attempts to submit an assignment even after reaching 100% which was the highest score.	Data
Potential work product	Log data was generated every time learners made an additional assignment submission. The log data were used to form count variables.	Data

Table 5.2: A design pattern [3, 58] of additional submissions of assignments after a learner received 100%

5.3.3 Data analysis

5.3.3.1 Indicators

Previous studies showed differences in how to use the AGQ-R to measure achievement goals based on the goal complex theory. As the goal complex theory does not differentiate the approach-

avoidance dimension, some of the studies combined approach and avoidance items to measure each goal [161]. For example, a combination of mastery-approach and mastery-avoidance items becomes the indicator of a mastery goal. On the other hand, other studies [15, 16, 127, 144] only used approach items as indicators for each goals. That is, a mastery goal and a performance goal were respectively measured through mastery-approach items and performance-approach items.

Thus, there were two different types of survey datasets included in the study. The first survey dataset was composed of sixteen indicator variables: all twelve questions of the AGQ-R [51] as well as four questions from the questionnaire measuring motivations behind performance-oriented goals [147, 162]. The second survey dataset excluded mastery-avoidance questions and performance-avoidance questions which made the number of entire indicator variables ten. Each of these datasets was split by occasion depending on whether the survey data were collected in week 1 or in week 3 of the course. Thus, there were four survey datasets: week 1 survey dataset, week 3 survey dataset, week 1 no-avoidance survey dataset, and week 3 no-avoidance survey dataset. Similarly, trace datasets were used as described in Chapter 4. To this, five indicators were added as per the previous chapter. The trace dataset was also split by occasion based on which weeks the data were collected: week 1-2 trace dataset and week 3-4 trace dataset. A full list of indicators used in this study is presented on Table 5.3.

Source	Description	Theoretical Relationship	Specific Indicator Variable
Tip-of-the-week email	(continuous) How many times learners opened a tip-of-the-week notification email in the Nth week	Mastery	email2_count email4_count
Tip-of-the-week Jupyter Notebook	(continuous) How many log data learners generated while interacting with a tip-of-the-week Jupyter Notebook in the Nth week	Mastery	notebook2_count notebook4_count
Bonus assignment	(binary) If learners submitted the Nth-week bonus assignment and wanted to anonymously share their answers	Performance-autonomous	bonus1_anonymous bonus2_anonymous bonus3_anonymous bonus4_anonymous
Bonus assignment	(binary) If learners submitted the Nth-week bonus assignment and wanted to share their answer with credit	Performance-controlling	bonus1_credit bonus2_credit bonus3_credit bonus4_credit
Extra assignment	(binary) If learners made at least one submission for the extra weekly assignment of the Nth week, which did not give bonus points	Mastery	extra1_no_submit extra3_no_submit
Bonus assignment	(continuous) How many log data learners generated while interacting with bonus assignments on Jupyter Notebook in the Nth week regardless of if the assignment was submitted.	Performance	bonus1_count bonus2_count bonus3_count bonus4_count
Extra assignment	(continuous) How many log data learners generated while interacting with extra assignments on Jupyter Notebook in the Nth week regardless of if the assignment was submitted.	Mastery	extra1_count extra2_count extra3_count extra4_count
Assignment	(count) How many assignments showed a record of additional submissions after a learner receiving 100% credit either in the first two weeks (i.e., additional12_count) or in the last two weeks (i.e., additional34_count).	Mastery	additional12_count additional34_count

Table 5.3: The trace indicators included in the final datasets of study 3

5.3.3.2 Latent variable mixture modeling

After initial data cleaning to remove learners who did not submit all surveys, paralleling the data cleaning process in the previous chapter, latent variable mixture modeling was applied on each of six cleaned datasets. Mplus 8.6 [113] and MplusAutomation R package [66] were used with a Maximum Likelihood with Robust standard errors (MLR) estimator. It was required to choose a statistical software that can run both LCA and LPA in the same model, as trace datasets had both binary indicators and continuous indicators. Mplus was specifically chosen for this purpose [113] and MplusAutomoation was additionally used to write more concise Mplus code. Mplus and MplusAutomation code used in this study is attached in Appendix A. The numbers of clusters across these solution k ranged from one to six (k_1, ..., k_6). The range was decided based on the previous studies which conducted latent variable mixture modeling on survey responses in reference to achievement goals and found no more than six clusters [124, 140, 184].

To answer RQ1 and RQ2, each cluster solution was examined for statistical robustness and theoretical interpretability. Statistical criteria to evaluate the final cluster solution of each datasets were Log Likelihood (LL), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), adjusted Bayesian Information Criterion (aBIC), Bootstrap Likelihood Ratio Test (BLRT) p-value, and Vuo-Lo-Mendell-Rubin likelihood ratio test (VLMR) p-value. The lower the LL, AIC, BIC, and aBIC values were, the more preferred was that solution compared to other solutions. Among these, BIC has been often considered as the most reliable fit statistic criteria. BIC rewards more parsimonious model with fewer parameters [118, 165, 169]. While recent simulation study questioned the reliability of AIC, it was still included as there is no general agreement about if AIC should be completely excluded from reports of cluster solutions [118, 138]. In addition to these, a conventional BLRT p-value or VLMR p-value at 0.05 level are commonly used to decide if a k_n is statistically detectably better than the k_{n-1}. These criteria have been commonly used in latent variable mixture modeling studies [119, 124, 138, 169].

Instead of adopting strict alpha values of p-value at 0.05 as the cutoff ruling decision-making of final solution, p-values in a continuous form were reported and were considered as one evidence along with other statistical criteria and theoretical interpretability[1]. It was also confirmed that the final solution of each dataset was generated through normally terminated solution estimation and that the best log likelihood was replicated, which are necessary conditions for each solution to be trustworthy [113]. Furthermore, to determine a final solution per dataset, cluster sizes and parameter sizes were also checked. There is no existing agreement on how large a cluster should be, but small clusters often do not conceptually add values [169]. This study followed a suggestion from previous work [145, 169]; unless a cluster has distinctive characteristics from other clusters,

[1] See Wasserstein and Lazar [167] and Wasserstein et al. [168] to learn more about why p-values at 0.05 is limited and should not be considered as the only evidence to make scientific decisions

a solution with one or more clusters with less than 10 learners, which is approximately 5% of the sample size per cluster, was not preferred. Furthermore, a parameter value should be smaller than entire learner population size of each dataset (i.e., 193).

It is generally a common practice to consider theoretical interpretability in conjunction with the statistical criteria when latent variable mixture modeling is used. Even if a solution reported superior statistical criteria, it is not recommended to select it as the final solution without theoretically meaningful interpretability [110, 118, 169]. Yet, considering the aim of this work was investigating information as captured by surveys and trace data, theoretical interpretability was not used to reject a certain solution. Instead, theoretical interpretability was discussed to coherently explain and label clusters.

Theoretical interpretability was examined through parameter examination and visual inspection. Parameters included threshold of categorical variables and means of continuous and count variables. Thresholds represent what percentage of the population in a cluster chose a particular option for a type of categorical data. Mplus provides thresholds in the logit scale, so a threshold of 3 shows that approximately 95% of the population in cluster A submitted the assignment. An equation to transform a threshold N on the logit scale to a percentage is presented below.

$$1/(1 + exp(-N)) \tag{5.1}$$

5.3.3.3 Cross tabulation

To answer RQ3 on the relationship between survey dataset and trace dataset, cross tabulation was used. The first table was generated using cluster outputs of latent variable mixture modeling on the week 1 survey dataset and then week 1-2 trace dataset. With clusters of the survey dataset on columns and clusters of the trace dataset on indices, the pivot table could present how learners' survey responses were related with their behavioral trace data. For example, if there is a pattern presenting that learners who responded to mastery achievement goal items mostly positively but did not engage much with learning materials for mastery learners, it shows that learners' expectation or plan on their self-reported achievement goals were not aligned with learners' behavioral indicators of achievement goals during learning.

5.4 Results

5.4.1 Data overview

The total sample of learners was 193 after removing learners who did not complete any of AGQ-R and the motivation survey in weeks 1 and 3. Survey datasets analyzed in this study were same

as the datasets used in Chapter 4. A descriptive statistic summary of these additional trace indicators is presented on Table 5.4. (See Table 4.5 and Table 4.6 in Chapter 4 for summaries of descriptive statistics for the rest of the data). As the datasets used in study 2, bonusN_no_sharing indicators were all collapsed to bonusN_anonymous indicators since bonusN_no_sharing indicators only showed a couple of observations.

Continuous and count variables were also log-transformed and z-scored. These variables were log transformed to contain outliers and then were z-scored to improve the solution convergence process for estimating the parameters of each latent variable mixture modeling solution [70]. Trace datasets had multiple continuous indicator variables: emailN_count, notebookN_count, bonusN_count, extraN_count, and additionalN_count where the N represents from which week the data were collected.

		Performance		Mastery		
		bonus1_count	bonus2_count	extra1_count	extra2_count	additional12_count
Week 1	$\bar{x}$	1107.005	97.010	230.601	19.968	0.098
	σ	1833.645	201.379	594.079	63.506	0.331
		bonus3_count	bonus4_count	extra3_count	extra4_count	additional34_count
Week 3	$\bar{x}$	64.186	52.927	7.637	1.621	0.139
	σ	136.686	160.688	35.147	10.734	0.403

Note. $\bar{x}$ = mean, σ = standard deviation.

Table 5.4: A descriptive statistic summary of newly added trace data which are the number of log events. Variable names are equal to the variable names in the row index of Table 5.3.

5.4.2 Findings

RQ1 and RQ2 respectively asked what clusters could be identified from survey datasets and from trace datasets. To answer these RQs, latent variable mixture modeling was conducted on four survey datasets and two trace datasets.

5.4.2.1 Model estimation: statistical criteria

Multiple solutions with different variance and covariance settings were examined to increase the stability of the final solution. Settings explored were chosen from Pastor et al. [124] and tweaked accordingly. For instance, on Mplus, the MLR estimator setting did not allow categorical variables to be specified and therefore variance settings were not changed for survey datasets which were composed of only categorical variables. For survey datasets, (1) solutions with covariances fixed at 0 and (2) solutions with freely estimated covariances were examined. For trace datasets, there

were six settings examined: (1) variances equal within and across clusters and covariances fixed at 0, (2) variances different within but equal across cluster and covariances fixed at 0, (3) variances different within a cluster but equal across cluster and covariances equal across clusters, and (4) variances different within a cluster but equal across cluster and freely estimated covariances.

Among these, only solutions with covariances fixed at 0 were successfully identified for the survey datasets. For trace datasets, there were two types of solutions identified: (1) solutions with variances equal within and across clusters and covariances fixed at 0 and (2) solutions with variances varying within but equal across clusters and covariances fixed at 0. Thus, there were two types of six solutions (k_1,...,k_6) identified per each trace dataset.

To identify the final solution per each dataset, statistical criteria and cluster sizes of each solution were evaluated. For example, Table 5.6 presents statistical values of solutions for the week 1 survey dataset. BLRT p-values and VLMR p-values agreed that there was no meaningful improvement in solutions beyond the 3-cluster solution (k_3). k_3 also reported the smallest BIC value. It showed large enough cluster size and the number of parameter was smaller than the number of learners (i.e., 193). Therefore, k_3 was further investigated of theoretical interpretability. For the trace datasets which had two sets of six solutions, there were two iterations of the process to select the final solution of each set. Then, these final solutions were compared against each other. For both week 1-2 trace dataset and week 3-4 trace dataset, the solution with variances varying within but equal across a clusters and covariances fixed at 0 was selected for the further investigation.

After applying the same set of statistical criteria, the following were chosen as the final solution for each dataset (Table 5.5): k_2 for the week 1 survey dataset (Table 5.6), k_3 for the week 1 no-avoidance survey dataset (Table 5.7), k_4 for the week 3 no-avoidance survey dataset (Table 5.9), k_3 for the week 1-2 trace dataset (Table 5.10) and k_3 for the week 3-4 trace dataset (Table 5.11). For the week 3 survey dataset, there was no solution in which were meaningfully improved compared to the solution with one less cluster according to BLRT and VLMR p-values (Table 5.8).

Dataset	Final solution
Week 1 survey	k_2
Week 1 no-avoidance survey	k_3
Week 3 survey	–
Week 3 no-avoidance survey	k_4
Week 1-2 trace	k_3
Week 3-4 trace	k_3

Note. k_n = A solution with cluster size n.

Table 5.5: Final solutions per dataset. Week 3 survey solution was not reported since it was not identified.

Model	Parameters	LL	AIC	BIC	aBIC	BLRT p-value	VLMR p-value	Cluster Size
k_1	60	-3703.877	7527.754	7723.515	7533.451	NA	NA	193
k_2	**121**	**-3240.683**	**6723.367**	**7118.152**	**6734.855**	**0.000**	**0.000**	**60, 133**
k_3	182	-3003.358	6370.716	6964.526	6387.996	0.000	0.528	66, 75, 52
k_4	243	-2873.244	6232.488	7025.322	6255.560	0.000	0.765	51, 53, 25, 64
k_5	304	-2790.955	6189.911	7181.769	6218.774	0.000	0.759	40, 55, 41, 25, 32
k_6	365	-2727.422	6184.844	7375.726	6219.499	0.012	0.782	29, 46, 28, 42, 24, 24

Table 5.6: The statistical criteria of each solution for week 1 survey dataset

Model	Parameters	LL	AIC	BIC	aBIC	BLRT p-value	VLMR p-value	Cluster Size
k_1	36	-2199.975	4471.951	4589.407	4475.369	NA	NA	193
k_2	73	-1945.051	4036.101	4274.277	4043.032	0.000	0.000	54, 139
k_3	**110**	**-1827.221**	**3874.442**	**4233.338**	**3884.886**	**0.000**	**0.000**	**51, 78, 64**
k_4	147	-1779.034	3852.069	4331.684	3866.026	0.000	1.000	73, 50, 52, 18
k_5	184	-1737.902	3843.805	4444.140	3861.275	0.000	0.760	49, 63, 42, 17, 22
k_6	221	-1707.752	3857.503	4578.558	3878.486	0.000	0.263	23, 50, 29, 61, 18, 12

Table 5.7: The statistical criteria of each solution for week 1 no-avoidance survey dataset

Model	Parameters	LL	AIC	BIC	aBIC	BLRT p-value	VLMR p-value	Cluster Size
k_1	63	-4076.660	8279.320	8484.870	8285.302	NA	NA	193
k_2	127	-3576.170	7406.339	7820.701	7418.397	0.000	0.3382	53, 140
k_3	191	-3280.754	6943.509	7566.683	6961.643	0.000	0.762	43, 73, 77
k_4	255	-3111.637	6733.274	7565.260	6757.485	0.000	0.2797	62, 49, 44, 38
k_5	319	-2987.364	6612.728	7653.527	6643.016	0.000	0.5973	42, 48, 31, 58, 14
k_6	383	-2907.822	6581.644	7831.254	6618.008	0.000	0.7602	42, 43, 48, 10, 34, 27

Table 5.8: The statistical criteria of each solution for week 3 survey dataset

Model	Parameters	LL	AIC	BIC	aBIC	BLRT p-value	VLMR p-value	Cluster Size
k_1	39	-2492.567	5063.135	5190.380	5066.838	NA	NA	193
k_2	79	-2224.911	4607.822	4865.575	4615.32	0.000	0.000	61, 132
k_3	119	-2059.354	4356.707	4744.967	4368.006	0.000	0.170	45, 62, 86
k_4	**159**	**-1962.351**	**4242.702**	**4761.469**	**4257.798**	**0.000**	**0.016**	**49, 17, 69, 58**
k_5	199	-1913.276	4224.552	4873.827	4243.446	0.000	0.685	45, 26, 49, 14, 59
k_6	239	-1897.810	4273.619	5053.402	4296.311	1.000	0.836	17, 25, 32, 51, 19, 49

Table 5.9: The statistical criteria of each solution for week 3 no-avoidance survey dataset

Model	Parameters	LL	AIC	BIC	aBIC	BLRT p-value	VLMR p-value	Cluster Size
k_1	19	-2205.706	4449.412	4511.403	4451.216	NA	NA	193
k_2	32	-1853.697	3771.395	3875.801	3774.433	0.000	0.000	97, 96
k_3	**45**	**-1654.356**	**3398.712**	**3545.533**	**3402.984**	**0.000**	**0.007**	**44, 97, 52**
k_4	58	-1559.436	3234.871	3424.107	3240.378	0.000	0.055	93, 13, 4, 83
k_5	71	-1433.373	3008.746	3240.397	3015.487	0.000	0.678	93, 38, 10, 45, 7
k_6	84	-1340.834	2849.667	3123.733	2857.643	0.000	0.289	93, 4, 45, 38, 6, 7

Table 5.10: The statistical criteria of each solution for week 1-2 trace dataset

Model	Parameters	LL	AIC	BIC	aBIC	BLRT p-value	VLMR p-value	Cluster Size
k_1	19	-2201.086	4440.172	4502.163	4441.976	NA	NA	193
k_2	32	-1830.972	3725.944	3830.350	3728.983	0.000	0.535	186, 7
k_3	**45**	**-1580.288**	**3250.577**	**3397.398**	**3254.849**	**0.000**	**0.002**	**137, 49, 7**
k_4	58	-1444.111	3004.221	3193.457	3009.728	0.000	0.602	7, 43, 137, 6
k_5	71	-1317.119	2776.238	3007.889	2782.979	0.000	0.900	7, 137, 42, 3, 4
k_6	84	-1279.792	2727.585	3001.651	2735.560	0.000	1.000	1, 32, 6, 7, 137, 10

Table 5.11: The statistical criteria of each solution for week 3-4 trace dataset

5.4.2.2 Model estimation: theoretical interpretability

To continue to answer RQ1 and RQ2, the final solutions' thresholds of categorical variables and means of continuous and count variables were examined for theoretical interpretability.

Overall, all clusters generated by survey-based solutions (RQ1) showed that learners responded highly positively to most of achievement goal questions except performance-controlling items. More specifically, regarding k_2 for the week 1 survey dataset (Figure 5.2), learners in the cluster 1 ($n = 60$) showed positive responses to all question items except both performance-controlling items. Thus, cluster 1 was labelled as 'strongly positive' group. Learners in cluster 2 ($n = 133$) still showed highly positive responses toward mastery approach and mastery avoidance items. Yet, their responses toward performance-autonomous items were more neutral. Thus, the cluster 2 was labelled as 'strongly mastery and lean performance-autonomous' group.

In terms of k_3 for the week 1 no-avoidance survey dataset (Figure 5.3), cluster 1 ($n = 51$) was labelled as 'strongly positive' since it showed similar response proportions to the cluster 1 of the week 1 survey dataset. Response proportion of cluster 2 ($n = 78$) was similar to the 'strongly positive' cluster of the same dataset with a lower proportion of response 'strongly agree' and an increase of response 'agree.' Therefore, cluster 2 was named as 'generally positive.' Responses of learners in cluster 3 ($n = 64$) focused toward mastery goal items, and there were more negative and neutral responses to performance, performance-autonomous, and performance-controlling motivation items. Thus, the cluster 3 was labelled as 'strongly mastery.'

Regarding k_4 for the week 3 no-avoidance survey dataset (Figure 5.4), the cluster 1 ($n = 49$) and the cluster 4 ($n = 58$) were respectively labelled as 'strongly positive' and as 'generally positive' considering dominantly positive responses toward the mastery, performance, and performance-autonomous items. The cluster 3 ($n = 69$) was named as 'strongly mastery' for response proportions which were similar to the cluster 3 for the week 1 no-avoidance survey dataset. The cluster 2 ($n = 17$) were labelled as 'lean mastery' for their less positive answers to mastery items and dominantly neutral or negative responses to performance items.

In contrast to the survey-based clusters showing highly positive toward most of achievement goals, trace-based solutions clustered learners into three distinctive groups including less engaged group which were the biggest in each dataset (RQ2). Regarding the k_3 for the week 1-2 trace dataset (Figure 5.5, Table 5.12), cluster 3 ($n = 52$) shows learners who were comparably more interested in engaging with bonus and extra materials. Learners in the cluster 3 generated the highest number of log data from engagement with bonus assignments, extra assignments, and tip-of-the-week Jupyter Notebook (Table 5.12). Furthermore, they also showed more interest in sharing their answers to the week 1 bonus assignment with credit than sharing them anonymously. This cluster was named as 'mastery and performance-controlling' group. On the other hand, learners in cluster 1 ($n = 44$) showed more targeted interest on bonus materials. While they opened tip-of-the-week

Figure 5.2: Response proportion (i.e., thresholds of categorical indicators) of the 2-cluster solution on the week 1 survey dataset.

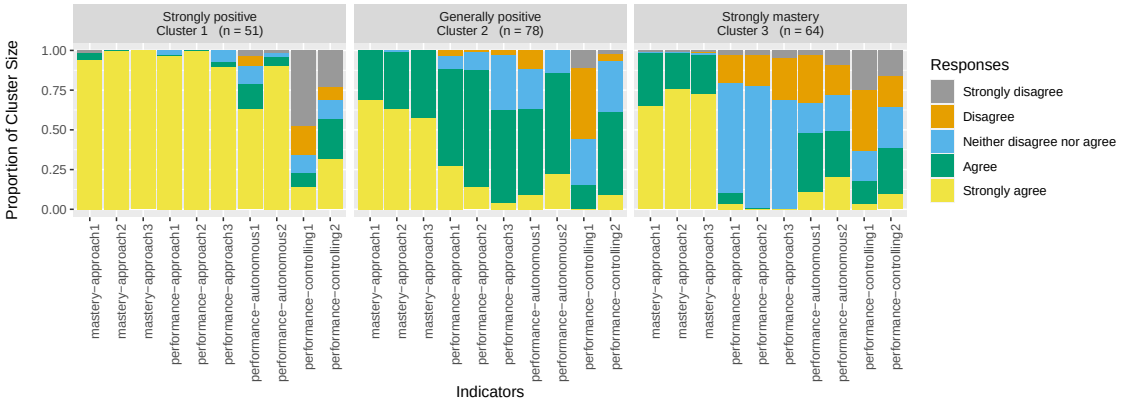

Figure 5.3: Response proportion (i.e., thresholds of categorical indicators) of the 3-cluster solution on the week 1 no-avoidance survey dataset.

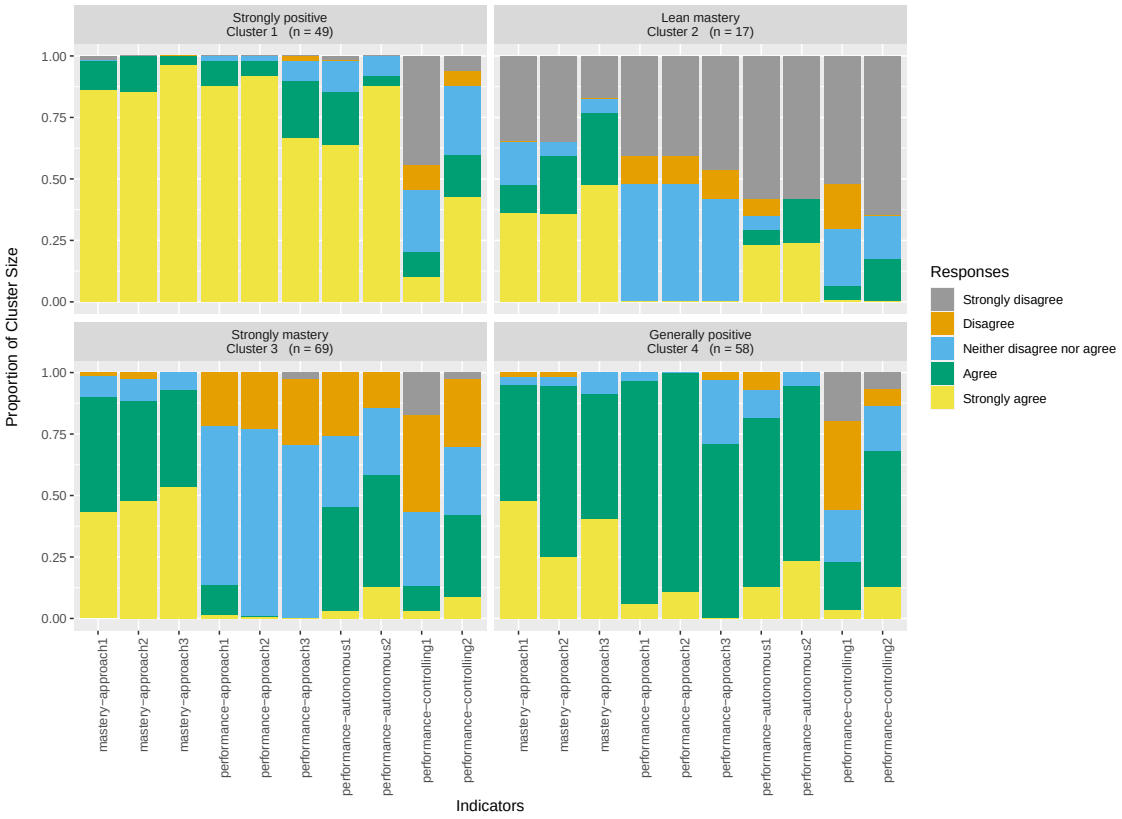

Figure 5.4: Response proportion (i.e., thresholds of categorical indicators) of the 4-cluster solution on the week 3 no-avoidance survey dataset.

notification emails more than learners in cluster 3, notebook2_count variables showed that their engagement with tip-of-the-week Jupyter Notebook was not as high as that of learners in cluster 3 (Table 5.12). Means of indicators extra1_count and extra2_count showing their engagement with extra assignments were also lower than means of bonus1_count and bonus2_count representing learners' engagement with bonus assignments. Learners in cluster 1 also preferred anonymous sharing over sharing with credit for the week 1 bonus assignment (Figure 5.5). This cluster was labelled as 'performance-autonomous' group. Learners in cluster 2 ($n = 97$) overall showed low engagement with materials. This cluster were labelled as 'less engaged' group.

Finally, for the k_3 for the week 3-4 trace dataset (Figure 5.6, Table 5.13), cluster 3 ($n = 7$) reported the smaller number of learners than 5% of the entire sample size but k_3 still remained as the final solution for the dataset. It was not only clearly distinguished from other clusters but also added meaningful interpretation of the data. Cluster 3 showed the highest engagement with overall bonus and extra materials in all three clusters. In particular, the indicators extra3_count, extra4_count (Table 5.13) and extra3_submit (Figure 5.6) showed that learners in cluster 3 engaged with both weeks 3 and 4 extra assignments much more than learners in the other clusters. Furthermore, they also made the most additional submission of assignments even after they received 100%. In terms of performance achievement goal indicators, learners in cluster 3 showed comparably higher engagement with performance materials with mixed preference between anonymous sharing and sharing with credits over weeks 3 and 4. Thus, cluster 3 was labelled as 'mastery and performance-autonomous' group. On the other hand, learners in cluster 2 ($n = 49$) consistently preferred sharing with credit over sharing anonymously in weeks 3 and 4. They also showed more weighed interest in bonus assignments than extra assignments. Values of bonus3_count and bonus4_count were much higher than values of extra3_count and extra4_count. Yet, their notebook4_count value was the highest in all clusters. This cluster was labelled as 'mastery and performance-controlling' group. Lastly, cluster 1 with the largest number of learners ($n = 136$) generally showed less engagement with overall materials and therefore named as 'less engaged' group.

5.4.2.3 Cross tabulation on relationships between survey- and trace-based clusters

RQ3 asked the relationship between the survey response data and trace data. To answer the exploratory RQ, three cross tabulations were generated. If survey-based clusters and trace-based clusters were aligned, similarly labelled clusters should show high overlap of learners. Yet, t hese tabulations showed that these clusters were not closely aligned.

Table 5.14 presents the relationship between cluster outputs of the week 1 survey dataset and cluster outputs of week 1-2 trace dataset. While learners' survey responses did not clearly distinguish mastery, performance-autonomous, and performance-controlling, the behavioral trace indicators identified different achievement goals. Specifically, learners made predominantly positive

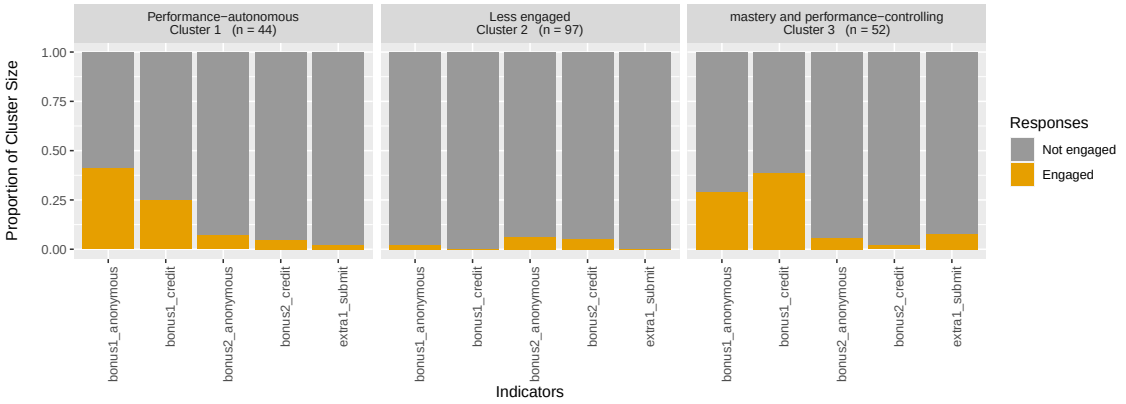

Figure 5.5: Proportions of engagement with learning materials (i.e., thresholds of binary indicators) for the 3-cluster solution on the week 1-2 trace dataset.

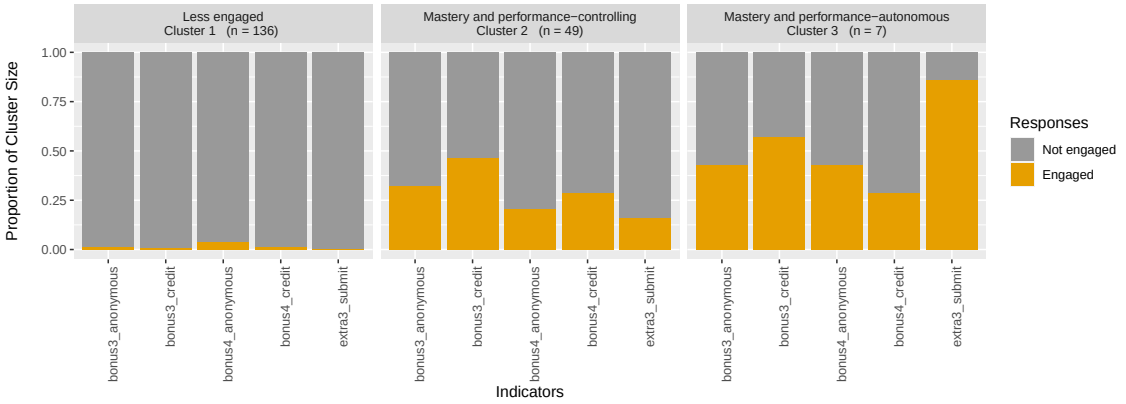

Figure 5.6: Proportions of engagement with learning materials (i.e., thresholds of binary indicators) of the 3-cluster solution on the week 3-4 trace dataset.

	Performance -autonomous Cluster 1 (n = 44)	Less engaged Cluster 2 (n = 97)	Mastery and performance -controlling Cluster 3 (n = 52)
email2_count	0.186	-0.054	-0.013
notebook2_count	0.145	-0.289	0.587
bonus1_count	0.937	-0.987	1.022
bonus2_count	0.343	-0.444	0.749
extra1_count	-0.393	-0.642	1.692
extra2_count	-0.087	-0.298	0.791
additional12_count	0.145	-0.188	0.137

Note 1. Means were computed with z-scored continuous variables.
Note 2. Standard deviations were not reported since each indicator had the same variances across clusters due to the setting of the solution.

Table 5.12: Means of continuous variables in week 1-2 trace dataset.

	Less engaged Cluster 1 (n = 136)	Mastery and performance -controlling Cluster 2 (n = 49)	Mastery and performance -autonomous Cluster 3 (n = 7)
email4_count	0.013	-0.033	0.623
notebook4_count	-0.151	0.408	0.193
bonus3_count	-0.602	1.480	1.546
bonus4_count	-0.376	0.866	1.274
extra3_count	-0.278	0.252	3.550
extra4_count	-0.184	-0.166	4.284
additional34_count	-0.200	0.310	0.985

Note 1. Means were computed in the logit scale on Mplus.
Note 2. Standard deviations were not reported since each indicator had the same variances across clusters due to the setting of the solution.

Table 5.13: Means of continuous variables in week 3-4 trace dataset.

responses to most of survey items except performance-controlling items. However, the trace-based solution identified 97 learners, which is approximately 50% of learners, as 'less engaged.' Furthermore, 44 of learners into 'mastery and performance-controlling' although most learners responded negatively to the performance-controlling items.

Table 5.15 displays the relationship between cluster outputs of the week 1 no-avoidance survey dataset and cluster outputs of week 1-2 trace dataset. This table also showed similar relationship to what the previous table presented. For example, while most learners made positive responses toward survey items except performance-controlling items, approximately a half of them were

clustered into ‘less engaged’ group. Furthermore, about 20 to 30% of learners were clustered as ‘mastery and performance-controlling.’

Finally, Table 5.16 shows the relationship between cluster outputs of the week 3 survey dataset and cluster outputs of week 3-4 trace dataset. According to the trace-based solution, there was a sharp increase of learners who were clustered into ‘less engaged’ group: from 97 learners to 137 learners. These ‘less engaged’ learners were from every survey-based cluster which all represented positive responses toward mastery or overall question items except performance-controlling ones.

Trace \ Survey	Strongly positive	Strongly mastery and lean performance -autonomous	Total
Performance-autonomous	8	36	44
Less engaged	29	68	97
Mastery and performance-controlling	16	36	52
Total	53	140	193

Table 5.14: Relationship between the week 1 survey data and the week 1-2 trace data

Trace \ Survey	Strongly positive	Generally positive	Strongly mastery	Total
Performance-autonomous	12	17	15	44
Less engaged	25	36	36	97
Mastery and performance-controlling	14	25	13	52
Total	51	78	64	193

Table 5.15: Relationship between the week 1 no-avoidance survey data and the week 1-2 trace data

5.5 Discussion and future work

This study posed three RQs as follows:

- RQ1. What goal clusters can be identified from survey data?
- RQ2. What goal clusters can be identified from trace data?
- RQ3. How are goal survey-based goal clusters and trace-based goal clusters related?

Trace \ Survey	Strongly positive	Lean mastery	Strongly mastery	Generally positive	Total
Less engaged	34	13	53	37	137
Mastery and performance-controlling	13	4	13	19	49
Mastery and performance-autonomous	2	0	3	2	7
Total	49	17	69	58	193

Table 5.16: Relationship between the week 3 survey data and the week 3-4 trace data

Findings showed that the survey-based clusters represented a predominantly positive reaction to achievement goals except the performance-controlling goal. On the other hand, trace-based clusters showed that more than half of learners in each dataset were clustered into the 'less engaged' group. The proportion of these 'less engaged' learners sharply increased from weeks 1 and 2 to weeks 3 and 4. Cross tabulations also showed that 'less engaged' learners were present in every survey-based cluster.

It is clear that achievement goals as measured by the surveys did not translate into behaviors with course materials. One possible explanation for this discrepancy specifically in weeks 1 and 2 would be the difference between expectation and the reality. The survey responses were collected before learning when learners have vague assumption of how the course would be. Furthermore, these responses could have reflected learners' hopes about how they would like to learn and how they would like to present themselves to teaching staff, peers, and themselves. It is less likely that learners have considered precise challenges which could change their goals or make them difficult pursue. Thus, once the course started, these vaguely defined goals could have been challenged. For example, learners might have thought that mastery goals were more socially desirable since some instructors encourage students to be internally motivated. Yet, they may forget to consider that pursuing mastery goals often requires more time commitment.

Even after 2 weeks of the course, which should allow enough time for learners to familiarize themselves with the course environment, the discrepancy between survey-based clusters and trace-base clusters did not change. On week 3 surveys, many learners still declared that their goals were either a combination of mastery and performance-autonomous goals or solely mastery goals. On the other hand, trace data indicated that more than half of learners did not engage with learning materials designed for each achievement goal. One explanation is that learners recognized their failure to meet the initial goals but decided to try again, which was not successful. Another explanation is that learners did not seriously commit to the goal. Either way, considering the increase of 'less engaged' learners in weeks 3 and 4, keeping a goal in their head and following a goal with

actions seem to be two different things.

The finding that achievement goals as measured by surveys do not translate into learning behaviors questions the common interpretation of existing goal theory and associated instruments. One example is why performance goals have been reported less than mastery goals. Multiple studies showed that students described their goals as performance goals much less frequently than mastery goals during interviews or surveys [26, 69, 157]. One of these studies even concluded that performance goals are not pursued enough for researchers to study them [26].

Many previous studies also assumed that learners maintained achievement goals measured before learning. Based on the assumption, these studies often aim to find causal relationships between achievement goals and learning outcomes [75, 98, 143, 180]. There is no consensus on a clear correlation between achievement goals and academic achievement for mastery-focused learners [143].

A potential explanation for these observations might be that many learners stated mastery goals before learning but did not actually follow the goal during learning. That is, it might not be appropriate to label mastery learners based on surveys before learning. It is crucial to conduct more studies with trace data in various contexts in order to investigate if there are other discrepancies which could be better explained.

One fundamental implication for future researchers is that they have to clearly define in their study what 'an achievement goal' is: (1) learners' expectations on how to learn before learning, (2) their actual process of goal pursuit during learning, or (3) their recall of how they learned after learning. The first and the third measurements could be valuable for researchers to understand learners' perception on their achievement goals and these could be collected through pre- and post-surveys. Yet, if researchers would like to identify the goal-relevant behaviors during learning, a survey would not be the best measure to adopt.

The implication for instructors is that such discrepancies between trace and survey data could be a useful technique to identify learners that need help with self-regulation. For example, if such a gap is detected in a course, instructors could intervene and see why that happened – was there any challenge? Or was it a healthy goal adjustment as learners get used to the course and recognize various contexts such as their skill level and task difficulty? After identifying reasons for the gap, instructors could decide if learners need support such as improving time management skills.

It is a limitation that the present study did not consider behavioral indicators outside of the course. It is possible that learners in the 'less engaged' cluster pursued goals in a way that could not be detected through trace data. For example, instead of engaging with additional materials and experimenting with alternative answers even after receiving 100% on their assignments, they might have read blog articles or followed tutorials outside of the course. A future study with broader scope could expand the understanding of how online learners accomplish their achievement goals outside of their courses.

CHAPTER 6

Discussion and conclusion

The field of learning analytics reflects the rise and development of data-intensive approaches to education [166]. For instance, designing and interpreting trace data has become a common practice for learning analytics researchers to study SRL. Adoption of trace data was also partially due to concerns over limitations of self-report measures in capturing SRL constructs. Survey questions, in particular, generally ask respondents to aggregate their experiences and therefore easily lose contextual information of SRL [60, 170, 188]. It has been also questioned respondents' ability to interpret or recall their own experiences to answer survey questions [79, 164, 170]. These limitations could cause low construct validity and detachment from dynamic SRL theories.

Although trace data do not suffer from the same limitations surveys have, trace data are not completely free from validity issues. Validity issues of trace data could appear when researchers' perspective on theories or interpretation of raw data into education construct are not coherent or well-thought-out. Kovanovic et al. [86], for example, showed that the validity of a time-on-task indicator depends on research contexts. That is, both trace data and survey data have their own limitations which could undermine research contribution.

Understanding how to appropriately employ survey and trace data is a timely task for learning analytics community. Learning analytics integrates knowledge from traditional instruments such as surveys and new findings from digital data such as traces [57]. In this integration process, it is an important step to investigate when and where to use these measures which have different characteristics. One such type of investigation would be identifying validity issues and limitations of measures in specific contexts [6, 78, 79, 86, 164, 170]. To build on these previous studies, in this dissertation, I answered the following overarching research question:

- How can the different nature of **self-reported surveys** and **trace data** be understood in order to adopt **more appropriate indicators** to capture specific **SRL constructs**?

To address this question, I identified two approaches from previous SRL works and implemented them: the complement approach and the comparison approach [19, 80, 122, 185]. Complement approach could address researchers' concern on how to answer SRL-related questions

with valid adoption of measurements. In Chapter 3, I used complementary surveys and trace indicators in order to understand the long-term effects of reflection prompts on learning. In particular, I assigned each method to constructs based on understanding of nature of each methods. I used surveys to measure overall satisfaction and perceived learning after learning with reflection prompts. These constructs were measured through surveys since they are aggregations of inner state of mind throughout learning. On the other hand, trace indicators were employed to measure learners' behaviors during learning, which could be better captured with more fine-grained and real-time data. The implication is that complementing both measures could help researchers understanding the overall pictures of SRL.

Although the complement approach is helpful in capturing different constructs together to build a holistic picture of SRL, this approach is not suitable for researchers who aim to focus on improving understanding of measurements by comparing the validity of measurements on the same constructs. This is where comparison approach is necessary. In Chapters 4 and 5, I compared the difference of information that is captured through surveys and trace indicators to understand what is more appropriate in measuring achievement goals. In particular, in Chapter 4, I revealed that surveys measure goal stated before learning which is different from learners' goal pursued during learning. This is striking since traditionally surveys have been commonly used to capture goals with assumptions that the goals would be maintained or rarely change throughout learning. One implication is that using trace data could be more valid in capturing conceptual constructs such as achievement goals when constructs are highly context-specific and could potentially change throughout learning.

Studies in this dissertation not only contribute to the better understanding of how to adopt measures for specific SRL constructs, but also shed lights on under-explored theoretical aspects of these constructs. After revealing the misalignment between survey data and the goal complex theory in Chapter 4, I investigated causes of misalignment by data-driven approach in Chapter 5. The study confirmed that goals stated before learning does not translate to the goal-relevant behaviors during learning. This finding questions not only the validity of surveys in measuring achievement goals during learning, but also generates new theoretical understanding of achievement goal constructs. For example, many previous studies have failed to find a clear relationship between mastery-focused goals and various learning outcomes, especially academic achievement. The current study suggests a potential explanation for these previous findings. That is, researchers might have labeled mastery learners based on survey responses before learning which does not equate to learners' actual goal pursuit behaviors during learning. In summary, the studies in this dissertation contribute to both methodological and theoretical understanding of SRL and suggest guidance for future learning analytics studies.

This dissertation did not aim to suggest a generalizable solution applicable across contexts

and constructs. As Messick [102] said, the validity should be understood in consideration with contexts for which measures or indicators are adopted. That is, the study findings might not be generalizable across future research in different contexts. For example, Chapters 4 and 5 were conducted in an online degree program where learners' motivation is different from that of college students taking residential undergraduate courses. This difference could cause findings diverging from the previous study findings.

Furthermore, trace indicators used in the studies have room for further improvement and generalization. Although trace indicators were designed through design patterns to understand alternative claims, the design process did not remove the inherent equivocality of trace data and its context-specific nature. Thus, there should be future study investigating and improving trace indicator designs for achievement goals tailored to each study context. When there are enough studies to begin showing patterns of appropriate trace indicator designs, learning analytics researchers can also propose more generalizable principles.

Beyond the issue of context, there is a need for reflection on how measures have been created and used in learning analytics and SRL studies. A large portion of theories have been built on self-reported data (e.g., surveys) and not all of these studies showed high validity in employing self-reported measures. Through better understanding of how self-reported surveys and trace data relate to one another, it is possible that previous findings may be questioned and challenged. Reflecting on previous methodological practices in light of modern technical opportunities may direct researchers toward novel findings which could explain previously under-investigated or misunderstood concepts.

APPENDIX A

Mplus and R Code

A.1 Code for CFA

Mplus code for CFA used in Chapter 4 is given below.

A.1.1 CFA of three-factor model on survey data collected in week1

```
cfa_survey12_three_factor  <- mplusObject(

  TITLE = "cfa_survey12_three_factor",

  VARIABLE =
   "
   usevariables = at1-pv3;
   categorical = at1-pv3;
   missing = ALL (999);

  ",


  MODEL =
    "
    Mastery BY mp1* mp2 mp3 mv1 mv2 mv3;
    Per_at BY pp1* pp2 pp3 pv1 pv2 pv3 at1 at2;
    Per_ct BY pp1* pp2 pp3 pv1 pv2 pv3 ct1 ct2;

    Mastery@1;
    Per_at@1;
```

```
    Per_ct@1;
    ",

  ANALYSIS = "
    estimator = WLSMV;
    starts = 100;
    stscale=1;
    PROCESSORS = 6;
  ",

  OUTPUT = "sampstat standardized residual
            mod(0) tech1 tech2 tech4;",

  PLOT =
    "type = plot3;",

   SAVEDATA = "FILE = save_cfa_survey12_three_factor.txt;
   SAVE = fscores;",

  rdata = df12)

cfa_survey_fit_w12_three_factor
            <- mplusModeler(cfa_survey12_three_factor,
                dataout="cfa_survey12_three_factor.dat",
                modelout="cfa_survey12_three_factor.inp" ,
                check=TRUE, run = TRUE, hashfilename = FALSE)
```

A.1.2 CFA of three-factor model on survey data collected in week3

```
cfa_survey34_three_factor  <- mplusObject(

  TITLE = "cfa_survey34_three_factor",

  VARIABLE =
   "
   usevariables = at1-pv3;
```

```
   categorical = at1-pv3;
   missing = ALL (999);
  ",


  MODEL =
    "
    Mastery BY mp1* mp2 mp3 mv1 mv2 mv3;
    Per_at BY pp1* pp2 pp3 pv1 pv2 pv3 at1 at2;
    Per_ct BY pp1* pp2 pp3 pv1 pv2 pv3 ct1 ct2;
    Mastery@1;
    Per_at@1;
    Per_ct@1;
    ",

  ANALYSIS = "
    estimator = WLSMV;
    starts = 100;
    stscale=1;
    PROCESSORS = 6;
  ",

  OUTPUT = "sampstat standardized
            mod(0) tech1 tech2 tech4;",

  PLOT =
    "type = plot3;",

  SAVEDATA = "FILE = save_cfa_survey34_three_factor.txt;
  SAVE = fscores;",

  rdata = df34)

cfa_survey_fit_w34_three_factor
            <- mplusModeler(cfa_survey34_three_factor,
                dataout="cfa_survey34_three_factor.dat",
```

```
                modelout="cfa_survey34_three_factor.inp" ,
                check=TRUE, run = TRUE, hashfilename = FALSE)
```

A.1.3 CFA of three-factor no-avoidance model on survey data collected in week1

```
cfa_survey12_three_factor_nov  <- mplusObject(

  TITLE = "cfa_survey12_three_factor_nov",

  VARIABLE =
   "
   usevariables = mp1 mp2 mp3 pp1 pp2 pp3 at1 at2 ct1 ct2;
   categorical = mp1 mp2 mp3 pp1 pp2 pp3 at1 at2 ct1 ct2;
   missing = ALL (999);

  ",


  MODEL =
    "
    Mastery BY mp1* mp2 mp3;
    Per_at BY pp1* pp2 pp3 at1 at2;
    Per_ct BY pp1* pp2 pp3 ct1 ct2;

    Mastery@1;
    Per_at@1;
    Per_ct@1;
    ",


  ANALYSIS = "
    estimator = WLSMV;
    starts = 100;
    stscale=1;
    PROCESSORS = 6;
```

```
  ",

  OUTPUT = "sampstat standardized residual
            mod(0) tech1 tech2 tech4;",

  PLOT =
    "type = plot3;",

  SAVEDATA = "FILE = save_cfa_survey12_three_factor_nov.txt;
SAVE = fscores;",

  rdata = df12)
fit_cfa_survey12_three_factor_nov
                <- mplusModeler(cfa_survey12_three_factor_nov,
                    dataout="cfa_survey12_three_factor_nov.dat",
                    modelout="cfa_survey12_three_factor_nov.inp" ,
                    check=TRUE, run = TRUE, hashfilename = FALSE)
```

A.1.4 CFA of three-factor no-avoidance model on survey data collected in week3

```
cfa_survey34_three_factor_nov  <- mplusObject(

  TITLE = "cfa_survey34_three_factor_nov",

  VARIABLE =
   "
   usevariables= mp1 mp2 mp3 pp1 pp2 pp3 at1 at2 ct1 ct2;
   categorical = mp1 mp2 mp3 pp1 pp2 pp3 at1 at2 ct1 ct2;
   missing = ALL (999);
  ",


  MODEL =
    "
    Mastery BY mp1* mp2 mp3 ;
```

```
    Per_at BY pp1* pp2 pp3 at1 at2;
    Per_ct BY pp1* pp2 pp3 ct1 ct2;

    Mastery@1;
    Per_at@1;
    Per_ct@1;
    ",


  ANALYSIS = "
    estimator = WLSMV;
    starts = 100;
    stscale=1;
    PROCESSORS = 6;
  ",

  OUTPUT = "sampstat standardized residual
            mod(0) tech1 tech2 tech4;",

  PLOT =
    "type = plot3;",

  SAVEDATA = "FILE = save_cfa_survey34_three_factor_nov.txt;
  SAVE = fscores;",

  rdata = df34)

fit_cfa_survey34_three_factor_nov
                <- mplusModeler(cfa_survey34_three_factor_nov,
                    dataout="cfa_survey34_three_factor_nov.dat",
                    modelout="cfa_survey34_three_factor_nov.inp" ,
                    check=TRUE, run = TRUE, hashfilename = FALSE)
```

A.1.5 CFA of three-factor model on trace data collected in week1-2

```
cfa_trace_w12_three_factor  <- mplusObject(
```

```
 TITLE = "cfa_trace12_three_factor",

VARIABLE =
  "
    usevariables = es1 TIPI_C1 TIPE_C1
    bs1_2 bs1_3 bs2_2 bs2_3 BS2_1;
    categorical = es1
    bs1_2 bs1_3 bs2_2 bs2_3 BS2_1;
    missing = ALL (999);
  ",



 MODEL =
   "
   Goal1 BY ES1* TIPI_C1 TIPE_C1;
   Goal2 BY BS1_3*  BS2_3;
   goal3 BY BS1_2*  BS2_2  BS2_1;

   goal1@1;
   goal2@1;
   goal3@1;
    ",

 ANALYSIS = "
   estimator = WLSMV;
   stscale=1;
   starts = 300;
   iterations = 10000;
   PROCESSORS = 6;
 ",

 OUTPUT = "standardized residual
           mod(0) tech1 tech2 tech4;",
```

```
  PLOT =
    "type = plot3;",

  SAVEDATA = "FILE = save_cfa_trace_week12_three_factor.txt;
  SAVE = fscores;",

  rdata = df12)

cfa_trace_fit_w12_three_factor
            <- mplusModeler(cfa_trace_w12_three_factor,
                dataout="cfa_trace_week12_three_factor.dat",
                modelout="cfa_trace_week12_three_factor.inp",
                check=TRUE, run = TRUE, hashfilename = FALSE)
```

A.1.6 CFA of three-factor model on trace data collected in week3-4

```
cfa_trace_w34_three_factor  <- mplusObject(

  TITLE = "cfa_trace_w34_three_factor_",


 VARIABLE =
   "
     usevariables =  TIPE_C4 TIPI_C4 ES3
     BS3_2 BS3_3 BS4_2 BS4_3;
     categorical = ES3
     BS3_2 BS3_3 BS4_2 BS4_3;
     missing = ALL (999);
    ",



  MODEL =
    "
    goal1 BY ES3* TIPI_C4 TIPE_C4;
```

```
    goal2 BY BS3_3* BS4_3;
    goal3 BY BS3_2* BS4_2;

    goal1@1;
    goal2@1;
    goal3@1;

     ",

  ANALYSIS = "
    estimator = WLSMV;
    starts = 300;
    stscale=1;
    iterations = 30000;
    PROCESSORS = 6;
  ",

  OUTPUT = "sampstat standardized residual
            mod(0) tech1 tech2 tech4;",

  PLOT =
    "type = plot3;",

  SAVEDATA = "FILE = save_cfa_trace_week34_three_factor.txt;
  SAVE = fscores;",

  rdata = df34)

cfa_trace_fit_w34_three_factor
         <- mplusModeler(cfa_trace_w34_three_factor,
            dataout="cfa_trace_week34_three_factor.dat",
            modelout="cfa_trace_week34_three_factor.inp",
            check=TRUE, run = TRUE, hashfilename = FALSE)
```

A.2 Code for latent variable mixture modeling (LVMM)

Mplusautomation code for LVMM used in the third study (Chapter 5) is given below.

A.2.1 LVMM on the survey data collected in week 1

```
lca_summary_w12_6   <- lapply(1:6, function(k) {
  lca_survey_w12  <- mplusObject(

    TITLE = glue("Survey model (week 1-2) class {k}"),

    VARIABLE = glue(
     "
     classes = c({k});
     usevariables = at1-pv3;
     categorical = at1-pv3;
     missing = ALL (999);
     "),


    MODEL = "

    %OVERALL%
    at1 WITH at2@0 ct1@0 ct2@0 mp1@0 mp2@0 mp3@0
    mv1@0 mv2@0 mv3@0 pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
    at2 WITH ct1@0 ct2@0 mp1@0 mp2@0 mp3@0 mv1@0
    mv2@0 mv3@0 pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
    ct1 WITH ct2@0 mp1@0 mp2@0 mp3@0 mv1@0 mv2@0
    mv3@0 pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
    ct2 WITH mp1@0 mp2@0 mp3@0 mv1@0 mv2@0 mv3@0
    pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
    mp1 WITH mp2@0 mp3@0 mv1@0 mv2@0 mv3@0 pp1@0
    pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
    mp2 WITH mp3@0 mv1@0 mv2@0 mv3@0 pp1@0 pp2@0
    pp3@0 pv1@0 pv2@0 pv3@0;
    mp3 WITH mv1@0 mv2@0 mv3@0 pp1@0 pp2@0 pp3@0
    pv1@0 pv2@0 pv3@0;
```

```
    mv1 WITH mv2@0 mv3@0 pp1@0 pp2@0 pp3@0 pv1@0
    pv2@0 pv3@0;
    mv2 WITH mv3@0 pp1@0 pp2@0 pp3@0 pv1@0 pv2@0
    pv3@0;
    mv3 WITH pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
    pp1 WITH pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
    pp2 WITH pp3@0 pv1@0 pv2@0 pv3@0;
    pp3 WITH pv1@0 pv2@0 pv3@0;
    pv1 WITH pv2@0 pv3@0;
    pv2 WITH pv3@0;
    ",


  ANALYSIS =
    "estimator = mlr;
    ALGORITHM=INTEGRATION;
    PARAMETERIZATION=RESCOV;
    integration = montecarlo;
    type = mixture;
    starts = 400 40;
    processors = 8;",

  OUTPUT = "residual tech11 tech14;",

  PLOT =
    "type = plot3;",

  SAVEDATA = glue("FILE = save_c{k}_lca_survey_week12.txt;
            SAVE = cprob;"),

  rdata = df12)

lca_survey_fit_w12 <- mplusModeler(lca_survey_w12,
                      dataout=glue("c_lca_survey_week12.dat"),
                      modelout=glue("c{k}_lca_survey_week12.inp"),
                      check=TRUE, run = TRUE, hashfilename = FALSE)
```

```
})
```

A.2.2 LVMM on the no-avoidance survey data collected in week 1

```
lca_summary_w12_6   <- lapply(1:6, function(k) {
  lca_survey_w12  <- mplusObject(

    TITLE = glue("(Nov) Survey model (week 1-2) class {k}"),

    VARIABLE = glue(
     "
     classes = c({k});
     usevariables = at1 at2 ct1 ct2 mp1 mp2 mp3
     pp1 pp2 pp3;
     categorical = at1 at2 ct1 ct2 mp1 mp2 mp3
     pp1 pp2 pp3;
     missing = ALL (999);
     "),

    MODEL = "
    %OVERALL%
    at1 WITH at2@0 ct1@0 ct2@0 mp1@0
    mp2@0 mp3@0 pp1@0 pp2@0 pp3@0;
    at2 WITH ct1@0 ct2@0 mp1@0 mp2@0
    mp3@0 pp1@0 pp2@0 pp3@0;
    ct1 WITH ct2@0 mp1@0 mp2@0 mp3@0
    pp1@0 pp2@0 pp3@0;
    ct2 WITH mp1@0 mp2@0 mp3@0 pp1@0
    pp2@0 pp3@0;
    mp1 WITH mp2@0 mp3@0 pp1@0 pp2@0
    pp3@0;
    mp2 WITH mp3@0 pp1@0 pp2@0 pp3@0;
    mp3 WITH pp1@0 pp2@0 pp3@0;
    pp1 WITH pp2@0 pp3@0;
    pp2 WITH pp3@0;
    ",
```

```
   ANALYSIS =
   "estimator = mlr;
    type = mixture;
    ALGORITHM=INTEGRATION;
    integration = montecarlo;
    PARAMETERIZATION=RESCOV;
    processors = 8;",

   OUTPUT = "residual tech11 tech14;",

   PLOT =
    "type = plot3;",

   SAVEDATA = glue("FILE = save_c{k}_lca_survey_week12_nov.txt;
            SAVE = cprob;"),

   rdata = df12)

lca_survey_fit_w12 <- mplusModeler(lca_survey_w12,
                    dataout=glue("c_lca_survey_week12_nov.dat"),
                    modelout=glue("c{k}_lca_survey_week12_nov.inp"),
                    check=TRUE, run = TRUE, hashfilename = FALSE)
})
```

A.2.3 LVMM on the survey data collected in week 3

```
lca_summary_w34_6   <- lapply(1:6, function(k) {
  lca_survey_w34  <- mplusObject(

    TITLE = glue("Survey model week 3-4 Class {k}"),

    VARIABLE = glue(
     "
```

```
 classes = c({k});
 usevariables = at1-pv3;
 categorical = at1-pv3;
 missing = ALL (999);
 "),


MODEL = "
%OVERALL%
at1 WITH at2@0 ct1@0 ct2@0 mp1@0 mp2@0 mp3@0
mv1@0 mv2@0 mv3@0 pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
at2 WITH ct1@0 ct2@0 mp1@0 mp2@0 mp3@0 mv1@0
mv2@0 mv3@0 pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
ct1 WITH ct2@0 mp1@0 mp2@0 mp3@0 mv1@0 mv2@0
mv3@0 pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
ct2 WITH mp1@0 mp2@0 mp3@0 mv1@0 mv2@0 mv3@0
pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
mp1 WITH mp2@0 mp3@0 mv1@0 mv2@0 mv3@0 pp1@0
pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
mp2 WITH mp3@0 mv1@0 mv2@0 mv3@0 pp1@0 pp2@0
pp3@0 pv1@0 pv2@0 pv3@0;
mp3 WITH mv1@0 mv2@0 mv3@0 pp1@0 pp2@0 pp3@0
pv1@0 pv2@0 pv3@0;
mv1 WITH mv2@0 mv3@0 pp1@0 pp2@0 pp3@0 pv1@0
pv2@0 pv3@0;
mv2 WITH mv3@0 pp1@0 pp2@0 pp3@0 pv1@0 pv2@0
pv3@0;
mv3 WITH pp1@0 pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
pp1 WITH pp2@0 pp3@0 pv1@0 pv2@0 pv3@0;
pp2 WITH pp3@0 pv1@0 pv2@0 pv3@0;
pp3 WITH pv1@0 pv2@0 pv3@0;
pv1 WITH pv2@0 pv3@0;
pv2 WITH pv3@0;
",
```

```
   ANALYSIS =
   "estimator = mlr;
    ALGORITHM=INTEGRATION;
    PARAMETERIZATION=RESCOV;
    integration = montecarlo;
    type = mixture;
    starts = 200 20;
    processors = 8;",

   OUTPUT = "sampstat residual tech4  tech11 tech12 tech14;",

   PLOT =
    "type = plot3;
    ",

   SAVEDATA = glue("FILE = save_c{k}_lca_survey_week34.txt;
             SAVE = cprob;"),

   rdata = df34)

lca_survey_fit_w34 <- mplusModeler(lca_survey_w34,
                      dataout=glue("c_lca_survey_week34.dat"),
                      modelout=glue("c{k}_lca_survey_week34.inp"),
                      check=TRUE, run = TRUE, hashfilename = FALSE)
})
```

A.2.4 LVMM on the no-avoidance survey data collected in week 3

```
lca_summary_w34_nov_6   <- lapply(1:6, function(k) {
  lca_survey_w34_nov  <- mplusObject(

    TITLE = glue("(Nov) Survey model week 3-4 Class {k}"),

    VARIABLE = glue(
     "
```

```
   classes = c({k});
   usevariables = mp1 mp2 mp3 pp1 pp2 pp3 at1 at2 ct1 ct2;
   categorical = mp1 mp2 mp3 pp1 pp2 pp3 at1 at2 ct1 ct2;
   missing = ALL (999);
   "),

  MODEL = "
    %OVERALL%
    at1 WITH at2@0 ct1@0 ct2@0 mp1@0
    mp2@0 mp3@0 pp1@0 pp2@0 pp3@0;
    at2 WITH ct1@0 ct2@0 mp1@0 mp2@0
    mp3@0 pp1@0 pp2@0 pp3@0;
    ct1 WITH ct2@0 mp1@0 mp2@0 mp3@0
    pp1@0 pp2@0 pp3@0;
    ct2 WITH mp1@0 mp2@0 mp3@0 pp1@0
    pp2@0 pp3@0;
    mp1 WITH mp2@0 mp3@0 pp1@0 pp2@0
    pp3@0;
    mp2 WITH mp3@0 pp1@0 pp2@0 pp3@0;
    mp3 WITH pp1@0 pp2@0 pp3@0;
    pp1 WITH pp2@0 pp3@0;
    pp2 WITH pp3@0;
",



ANALYSIS =
 "estimator = mlr;
  type = mixture;
  ALGORITHM=INTEGRATION;
  integration = montecarlo;
  PARAMETERIZATION=RESCOV;
  processors = 8;",

OUTPUT = "sampstat residual tech4  tech11 tech12 tech14;",
```

```
  PLOT =
    "type = plot3;
    ",

  SAVEDATA = glue("FILE = save_c{k}_lca_survey_week34_nov.txt;

  rdata = df34)

lca_survey_fit_w34_nov <- mplusModeler(lca_survey_w34_nov,
                dataout=glue("c_lca_survey_week34_nov.dat"),
                modelout=glue("c{k}_lca_survey_week34_nov.inp") ,
                check=TRUE, run = TRUE, hashfilename = FALSE)
})
```

A.2.5 LVMM on the trace data collected week 1-2

```
lca_k1_6  <- lapply(1:6, function(k) {
  lca_enum  <- mplusObject(

    TITLE = glue("Model Trace 1-2 Class {k}"),
   VARIABLE = glue(
   "
     classes = c({k});
     usevariables = tipi_c1 tipe_c1 es1
     bs1_2 bs1_3 bs2_2 bs2_3
     es_c1 bs_c1 es_c2 bs_c2
     as_c12;
     categorical = es1
     bs1_2 bs1_3 bs2_2 bs2_3;
     missing = ALL (999);
     "),

   MODEL = "
     %OVERALL%
    tipi_c1 tipe_c1 es_c1 bs_c1 es_c2 bs_c2 as_c12;
```

```
tipi_c1 WITH tipe_c1@0;
tipi_c1 WITH es_c1@0;
tipi_c1 WITH bs_c1@0;
tipi_c1 WITH es_c2@0;
tipi_c1 WITH bs_c2@0;
tipi_c1 WITH as_c12@0;
tipe_c1 WITH es_c1@0;
tipe_c1 WITH bs_c1@0;
tipe_c1 WITH es_c2@0;
tipe_c1 WITH bs_c2@0;
tipe_c1 WITH as_c12@0;
es_c1 WITH bs_c1@0;
es_c1 WITH es_c2@0;
es_c1 WITH bs_c2@0;
es_c1 WITH as_c12@0;
bs_c1 WITH es_c2@0;
bs_c1 WITH bs_c2@0;
bs_c1 WITH as_c12@0;
es_c2 WITH bs_c2@0;
es_c2 WITH as_c12@0;
bs_c2 WITH as_c12@0;
es1 WITH bs1_2@0;
es1 WITH bs1_3@0;
es1 WITH bs2_2@0;
es1 WITH bs2_3@0;
bs1_2 WITH bs1_3@0;
bs1_2 WITH bs2_2@0;
bs1_2 WITH bs2_3@0;
bs1_3 WITH bs2_2@0;
bs1_3 WITH bs2_3@0;
bs2_2 WITH bs2_3@0;
  ",


ANALYSIS =
 "estimator = mlr;
```

```
     ALGORITHM=INTEGRATION;
     PARAMETERIZATION=RESCOV;
     integration = montecarlo (100);
     starts = 400 40;
     type = mixture;
     processors = 8;",

    OUTPUT = "sampstat residual tech4
    tech7 tech11 tech12 tech13 tech14;",

    PLOT =
      "type = plot2 plot3;",

    SAVEDATA = glue("FILE = save_c{k}_lca_trace_week12.txt;
                SAVE = cprob;"),

    rdata = df12)



lca_enum_fit <- mplusModeler(lca_enum,
                dataout=glue("c_lca_trace_week12.dat"),
                modelout=glue("c{k}_lca_trace_week12.inp"),
                check=TRUE, run = TRUE, hashfilename = FALSE)
})
```

A.2.6 LVMM on the trace data collected week 3-4

```
```{r}
lca_k1_6  <- lapply(1:6, function(k) {
  lca_enum  <- mplusObject(

   TITLE = glue("Model trace 3-4 Class {k}"),
   VARIABLE = glue(
   "
```
```

```
 classes = c({k});
 usevariables = tipi_c4 tipe_c4 es3
 bs3_2 bs3_3 bs4_2 bs4_3
 es_c3 bs_c3 es_c4 bs_c4
 as_c34;
 categorical = es3
 bs3_2 bs3_3 bs4_2 bs4_3;
 missing = ALL (999);
 "),

MODEL = "
%OVERALL%
tipi_c4 tipe_c4 es_c3 bs_c3 es_c4 bs_c4 as_c34;
tipi_c4 WITH tipe_c4@0;
tipi_c4 WITH es_c3@0;
tipi_c4 WITH bs_c3@0;
tipi_c4 WITH es_c4@0;
tipi_c4 WITH bs_c4@0;
tipi_c4 WITH as_c34@0;
tipe_c4 WITH es_c3@0;
tipe_c4 WITH bs_c3@0;
tipe_c4 WITH es_c4@0;
tipe_c4 WITH bs_c4@0;
tipe_c4 WITH as_c34@0;
es_c3 WITH bs_c3@0;
es_c3 WITH es_c4@0;
es_c3 WITH bs_c4@0;
es_c3 WITH as_c34@0;
bs_c3 WITH es_c4@0;
bs_c3 WITH bs_c4@0;
bs_c3 WITH as_c34@0;
es_c4 WITH bs_c4@0;
es_c4 WITH as_c34@0;
bs_c4 WITH as_c34@0;
es3 WITH bs3_2@0;
es3 WITH bs3_3@0;
```

```
    es3 WITH bs4_2@0;
    es3 WITH bs4_3@0;
    bs3_2 WITH bs3_3@0;
    bs3_2 WITH bs4_2@0;
    bs3_2 WITH bs4_3@0;
    bs3_3 WITH bs4_2@0;
    bs3_3 WITH bs4_3@0;
    bs4_2 WITH bs4_3@0;

    ANALYSIS =
     "estimator = mlr;
     ALGORITHM=INTEGRATION;
     PARAMETERIZATION=RESCOV;
     integration = montecarlo;
     starts = 300 30;
     type = mixture;
     processors = 8;",

    OUTPUT = "sampstat residual tech4 tech7
    tech11 tech12 tech13 tech14;",

    PLOT =
      "type = plot2 plot3;",

    SAVEDATA = glue("FILE = save_c{k}_lca_trace_week34.txt;
                    SAVE = cprob;"),
    rdata = df34)



lca_enum_fit <- mplusModeler(lca_enum,
                dataout=glue("c_lca_trace_week34.dat"),
                modelout=glue("c{k}_lca_trace_week34.inp"),
                check=TRUE, run = TRUE, hashfilename = FALSE)
})
```

APPENDIX B

Survey Instruments

B.1 Survey description

The following questionnaire is a combination of the AGQ-R [51] and motivation survey [162] which were modified according to the context of the second and the third study. the questionnaire was presented to students at the beginning of week 1 and week 3 of a 4-week long online course. All the questions provided 5-point Likert response scale as the original design as follows:

1. Strongly disagree
2. Disagree
3. Neither disagree nor agree
4. Agree
5. Strongly agree

B.1.1 Survey contents

This quiz isn't really a quiz at all! It's a chance for me to gain some insight into how you are approaching this course, how you work, and how I might best support you. It is not worth any grades and is completely optional.

Here are a few statements that might describe how you approach learning. Please indicate how much you agree with each statement.

1. My aim is to completely master the material presented in this class.
2. My aim is to avoid learning less than I possibly could.
3. My aim is to perform well relative to other students.

4. My aim is to avoid doing worse than other students.
5. I am striving to understand the content of this course as thoroughly as possible.
6. I am striving to avoid an incomplete understanding of the course material.
7. I am striving to do well compared to other students.
8. I am striving to avoid performing worse than others.
9. My goal is to learn as much as possible.
10. My goal is to avoid learning less than it is possible to learn.
11. My goal is to perform better than the other students.
12. My goal is to avoid performing poorly compared to others.
13. "I am striving to do well compared to others in this class" Assume you agreed, even if only a little bit. What reason(s) motivate you to pursue this goal in your class?
    (a) Because I find this a highly stimulating and challenging goal.
    (b) Because I find this a personally valuable goal.
    (c) Because I have to comply with the demands of others such as parents, friends, and teachers.
    (d) Because I would feel bad, guilty, or anxious if I didn't do so.

APPENDIX C

Modification Index Tables

The tables presented here report modification indices of each survey dataset and trace dataset from Chapter 4. A modification index was used as a statistical criteria to evaluate CFA solutions. Commonly, 3.84 has been used as a cutoff point that corresponds to 1 degree of freedom at $p < 0.05$. That is, a modification index equal to or greater than the cutoff for a particular connection shows that the fit of a particular solution could be improved. This improvement could be achieved when a new connection between indicators or between an indicator and a factor is established.

In this study, to avoid misusing the $p < 0.05$ as a dichotomous cutoff of statistical significance, the entire modification indices were reported as continuous values.

Indicators	Mastery	Performance-autonomous	Performance-controlling
mastery-approach1		1.003	0.127
mastery-approach2		**9.477**	**14.321**
mastery-approach3		**10.181**	**15.990**
mastery-avoidance1		0.241	0.859
mastery-avoidance2		0.658	0.753
mastery-avoidance3		**10.666**	**17.509**
performance-approach1	0.127		
performance-approach2	0.102		
performance-approach3	0.987		
performance-avoidance1	2.256		
performance-avoidance2	0.050		
performance-avoidance3	3.233		
performance-autonomous1	2.780		0.018
performance-autonomous2	0.554		0.018
performance-controlling1	**11.157**	**4.432**	
performance-controlling2	3.141	**4.432**	

Note. Modification indices larger than 3.84 appear bold for readability.

Table C.1: Modification indices of the solution for the week 1 survey dataset

Indicators	Mastery	Performance-autonomous	Performance-controlling
mastery-approach1		**11.872**	**7.787**
mastery-approach2		0.057	0.021
mastery-approach3		0.982	0.435
mastery-avoidance1		**12.675**	**11.966**
mastery-avoidance2		2.412	1.151
mastery-avoidance3		0.210	1.227
performance-approach1	0.740		
performance-approach2	1.891		
performance-approach3	0.007		
performance-avoidance1	0.898		
performance-avoidance2	1.262		
performance-avoidance3	1.554		
performance-autonomous1	**4.172**		1.127
performance-autonomous2	0.232		1.127
performance-controlling1	**30.133**	2.889	
performance-controlling2	**11.356**	2.889	

Note. Modification indices larger than 3.84 appear bold for readability.

Table C.2: Modification indices of the solution for the week 3 survey dataset

Indicators	Mastery	Performance-autonomous	Performance-controlling
mastery-approach1		**9.684**	0.243
mastery-approach2		**5.317**	0.567
mastery-approach3		1.014	0.058
performance-approach1	0.232		
performance-approach2	3.206		
performance-approach3	0.006		
performance-autonomous1	3.246		0.289
performance-autonomous2	1.063		0.289
performance-controlling1	**12.616**	**11.231**	
performance-controlling2	**6.286**	**11.231**	

Note. Modification indices larger than 3.84 appear bold for readability.

Table C.3: Modification indices of the solution for the week 3 no-avoidance survey dataset

Indicators	Mastery	Performance -autonomous	Performance -controlling
email2_count		0.222	0.451
notebook2_count		0.301	2.660
bonus1_anonymous	0.000	0.002	
bonus1_credit	0.374		0.374
bonus2_no_sharing	0.037	0.194	
bonus2_anonymous	0.039	0.221	
bonus2_credit	0.375		0.375
extra1_submit		1.511	**4.827**

Note. Modification indices larger than 3.84 appear bold for readability.

Table C.4: Modification indices of the solution for the week 1-2 trace dataset (between factors and indicators)

	email2 _count	notebook2 _count	bonus1 _anonymous	bonus1 _credit	bonus2_no _sharing	bonus2 _anonymous	bonus2 _credit	extra1_no _submit
email2 _count	-							
notebook2 _count	**5.815**	-						
bonus1 _anonymous	0.031	0.077	-					
bonus1 _credit	0.027	0.208	2.745	-				
bonus2 _no_sharing	1.131	0.429	0.211	0.257	-			
bonus2 _anonymous	0.438	0.067	0.622	0.584	0.001	-		
bonus2 _credit	0.131	0.525	0.377	NA	0.630	0.826	-	
extra1 _submit	**4.480**	0.406	0.310	0.231	**3.995**	0.001	0.254	-

Note. Modification indices larger than 3.84 appear bold for readability.

Table C.5: Modification indices of the solution for the week 1-2 trace dataset (between indicators)

Indicators	Mastery	Performance -autonomous	Performance -controlling
email4_count		2.109	1.621
notebook4_count		0.792	1.012
bonus3 _submit	0.174		0.150
bonus3 _anonymous	0.150	0.174	
bonus3 _credit	0.174		0.150
bonus4 _anonymous	0.150	0.174	
extra3 _submit	0.061	0.002	0.011

Note. Modification indices larger than 3.84 appear bold for readability.

Table C.6: Modification indices of the solution for the week 3-4 trace dataset (between factors and indicators)

Indicators	email4 _count	notebook3 _count	bonus3 _anony -mous	bonus3 _credit	bonus4 _anony -mous	bonus4 _credit	extra3 _no _submit
email4 _count	-						
notebook4 _count	2.220	-					
bonus3 _anonymous	1.243	0.924	-				
bonus3 _credit	1.829	1.335	0.006	-			
bonus4 _anonymous	1.221	0.110	NA	0.191	-		
bonus4 _credit	2.641	0.007	0.159	NA	0.003	-	
extra3 _submit	0.579	0.070	0.006	2.086	0.018	1.129	-

Note. Modification indices larger than 3.84 appear bold for readability.

Table C.7: Modification indices of the solution for the week 3-4 trace dataset (between indicators)

BIBLIOGRAPHY

[1] Vincent Aleven and Kenneth R Koedinger. Limitations of student control: Do students know when they need help? In *Intelligent Tutoring Systems*, pages 292–303. Springer Berlin Heidelberg, 2000.

[2] Vincent Aleven, Bruce Mclaren, Ido Roll, and Kenneth Koedinger. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, 16(2):101–128, 2006.

[3] Christopher Alexander, Sara Ishikawa, and Murray Silverstein. *A Pattern Language: Towns, buildings, construction*. Oxford University Press, New York, 1977.

[4] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for educational and psychological testing*. American Educational Research Association, 2014.

[5] Carole Ames. Classrooms: Goals, structures, and student motivation. *J. Educ. Psychol.*, 84(3):261, 1992.

[6] Roger Azevedo. Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educ. Psychol.*, 50(1):84–94, January 2015.

[7] Maria Bannert. Effects of reflection prompts when learning with hypermedia. *Journal of Educational Computing Research*, 35(4):359–375, December 2006.

[8] Maria Bannert and Peter Reimann. Supporting self-regulated hypermedia learning through prompts. *Instructional Science*, 40(1):193–211, January 2012.

[9] Maria Bannert, Christoph Sonnenberg, Christoph Mengelkamp, and Elisabeth Pieger. Short- and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. *Comput. Human Behav.*, 52:293–306, 2015.

[10] Susan M Barnett and Stephen J Ceci. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, 128(4):612, 2002.

[11] Sarit Barzilai and Ina Blau. Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. *Comput. Educ.*, 70:65–79, January 2014.

[12] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

[13] James W Beck and Aaron M Schmidt. State-level goal orientations as mediators of the relationship between time pressure and performance: a longitudinal study. *J. Appl. Psychol.*, 98(2):354–363, March 2013.

[14] John T Behrens, Philip Piety, Kristen E DiCerbo, and Robert J Mislevy. Inferential foundations for learning analytics in the digital ocean. *Learning analytics in education*, pages 1–48, 2018.

[15] Moti Benita and Lennia Matos. Internalization of mastery goals: The differential effect of teachers' autonomy support and control. *Front. Psychol.*, 11:599303, 2020.

[16] Moti Benita, Lennia Matos, and Yasmin Cerna. The effect of mastery goal-complexes on mathematics grades and engagement: The case of Low-SES peruvian students. *Learning and Instruction*, page 101558, October 2021.

[17] Bernadette Berardi-Coletta, Linda S Buyer, Roger L Dominowski, and Elizabeth R Rellinger. Metacognition and problem solving: A process-oriented approach. *J. Exp. Psychol. Learn. Mem. Cogn.*, 21(1):205–223, January 1995.

[18] Matthew L Bernacki. Examining the cyclical, loosely sequenced, and contingent features of Self-Regulated learning. In *Handbook of Self-Regulation of Learning and Performance*, chapter 24, pages 370–387. Routledge, 2018.

[19] Matthew L Bernacki, Timothy J Nokes-Malach, and Vincent Aleven. Examining self-efficacy during learning: variability and relations to behavior, performance, and learning. *Metacognition and Learning*, 10(1):99–117, April 2015.

[20] Kirsten Berthold, Matthias Nückles, and Alexander Renkl. Do learning protocols support learning strategies and outcomes? the role of cognitive and metacognitive prompts. *Learning and Instruction*, 17(5):564–577, October 2007.

[21] Monique Boekaerts. Self-regulated learning: Where we are today. *International journal of educational research*, 31(6):445–457, 1999.

[22] Monique Boekaerts and Markku Niemivirta. Chapter 13 - Self-Regulated learning: Finding a balance between learning goals and Ego-Protective goals. In *Handbook of Self-Regulation*, pages 417–450. Academic Press, San Diego, 2000.

[23] George Botsas and Susana Padeliadu. Goal orientation and reading comprehension strategy use among students with and without reading difficulties. *Int. J. Educ. Res.*, 39(4):477–495, January 2003.

[24] David Boud, Rosemary Keogh, and David Walker. What is reflection in learning. *Reflection: Turning experience into learning*, pages 7–17, 1985.

[25] John D Bransford, Ann L Brown, and Rodney R Cocking. *How people learn: Brain, mind, experience, and school.* National Academy Press, 1999.

[26] Jere Brophy. Goal theorists should move on from performance goals. *Educ. Psychol.*, 40(3):167–176, September 2005.

[27] Timothy A Brown. *Confirmatory Factor Analysis for Applied Research, Second Edition*. Guilford Publications, January 2015.

[28] Timothy A Brown and Michael T Moore. Confirmatory factor analysis. *Handbook of structural equation modeling*, pages 361–379, 2012.

[29] Violet A Brown. An introduction to linear mixed-effects modeling in r. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920960351, 2021.

[30] Feinian Chen, Patrick J Curran, Kenneth A Bollen, James Kirby, and Pamela Paxton. An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociol. Methods Res.*, 36(4):462–494, January 2008.

[31] Zhi-Hong Chen, Calvin C Y Liao, Hercy N H Cheng, Charles Y C Yeh, and Tak-Wai Chan. Influence of game quests on pupils' enjoyment and goal-pursuing in math learning. *Journal of Educational Technology & Society*, 15(2):317–327, 2012.

[32] Michelene T H Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cogn. Sci.*, 13(2):145–182, April 1989.

[33] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[34] David A Cook, Richmond M Castillo, Becca Gas, and Anthony R Artino, Jr. Measuring achievement goal motivation, mindsets and cognitive load: validation of three instruments' scores. *Med. Educ.*, 51(10):1061–1074, October 2017.

[35] Melanie M Cooper, Santiago Sandi-Urena, and Ron Stevens. Reliable multi method assessment of metacognition use in chemistry problem solving. *Chemistry Education Research and Practice*, 9(1):18–24, 2008.

[36] William A Cunningham, Kristopher J Preacher, and Mahzarin R Banaji. Implicit attitude measures: consistency, stability, and convergent validity. *Psychol. Sci.*, 12(2):163–170, March 2001.

[37] Elizabeth A Davis. Prompting middle school science students for productive reflection: Generic and directed prompts. *Journal of the Learning Sciences*, 12(1):91–142, January 2003.

[38] Samuel B Day and Robert L Goldstone. The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, 47(3):153–176, 2012.

[39] Edward L Deci and Richard M Ryan. The“ what” and“ why” of goal pursuits: Human needs and the self-determination of behavior. *Psychol. Inq.*, 11(4):227–268, 2000.

[40] Anneline Devolder, Johan van Braak, and Jo Tondeur. Supporting self-regulated learning in computer-based learning environments: systematic review of effects of scaffolding in the domain of science education. *Journal of Computer Assisted Learning*, 28(6):557–573, 2012.

[41] John Dewey. *How We Think*. Courier Corporation, January 1997.

[42] Oliver Dickhäuser, Stefan Janke, Martin Daumiller, and Markus Dresel. Motivational school climate and teachers' achievement goal orientations: A hierarchical approach. *Br. J. Educ. Psychol.*, 91(1):391–408, March 2021.

[43] N Ding, R J Bosker, and E G Harskamp. Exploring gender and gender pairing in the knowledge elaboration processes of students using computer-supported collaborative learning. *Comput. Educ.*, 56(2):325–336, February 2011.

[44] Ellen A Drost. Validity and reliability in social science research. *Education Research and perspectives*, 38(1):105, 2011.

[45] Carol S Dweck. Motivational processes affecting learning. *Am. Psychol.*, 41(10):1040, 1986.

[46] Anastasia Efklides. Metacognition. *Eur. Psychol.*, 13(4):277–287, January 2008.

[47] Anastasia Efklides. Interactions of metacognition with motivation and affect in Self-Regulated learning: The MASRL model. *Educ. Psychol.*, 46(1):6–25, January 2011.

[48] Andrew J Elliot. Approach and avoidance motivation and achievement goals. *Educ. Psychol.*, 34(3):169–189, June 1999.

[49] Andrew J Elliot. A conceptual history of the achievement goal construct. In Andrew J Elliot and Carol S Dweck, editors, *Handbook of competence and motivation*, volume 16, pages 52–72. New York: The Guilford Press, 2005.

[50] Andrew J Elliot and Holly A McGregor. A 2$\times$ 2 achievement goal framework. *J. Pers. Soc. Psychol.*, 80(3):501, 2001.

[51] Andrew J Elliot and Kou Murayama. On the measurement of achievement goals: Critique, illustration, and application. *J. Educ. Psychol.*, 100(3):613, 2008.

[52] Andrew J Elliot, Kou Murayama, and Reinhard Pekrun. A 3 $\times$ 2 achievement goal model. *J. Educ. Psychol.*, 103(3):632–648, 2011.

[53] Andrew J Elliot and Todd M Thrash. Achievement goals and the hierarchical model of achievement motivation. *Educ. Psychol. Rev.*, 13(2):139–156, 2001.

[54] Katharina Engelmann, Maria Bannert, and Nadine Melzner. Do self-created metacognitive prompts promote short- and long-term effects in computer-based learning environments? *Research and Practice in Technology Enhanced Learning*, 16(1):1–21, February 2021.

[55] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

[56] Sara J Finney, Suzanne L Pieper, and Kenneth E Barron. Examining the psychometric properties of the achievement goal questionnaire in a general academic context. *Educ. Psychol. Meas.*, 64(2):365–382, April 2004.

[57] Society for Learning Analytics Research (SoLAR). What is learning analytics?, Mar 2021.

[58] Erich Gamma, Richard Helm, Ralph Johnson, Ralph E . Johnson, and John Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Pearson Deutschland GmbH, 1995.

[59] Joanna Giota and Daniel Bergh. Adolescent academic, social and future achievement goal orientations: Implications for achievement by gender and parental education. *Scandinavian Journal of Educational Research*, 65(5):831–850, July 2021.

[60] Jeffrey A Greene and Roger Azevedo. The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educ. Psychol.*, 45(4):203–209, 2010.

[61] Jeffrey Alan Greene and Roger Azevedo. A theoretical review of winne and hadwin's model of self-regulated learning: New perspectives and directions. *Rev. Educ. Res.*, 77(3):334–372, 2007.

[62] Jeffrey Alan Greene and Dale H Schunk. Historical, contemporary, and future perspectives on Self-Regulated learning and performance. In *Handbook of Self-Regulation of Learning and Performance*, pages 17–32. Routledge, 2018.

[63] Leslie Morrison Gutman. How student and parent goal orientations and classroom goal structures influence the math achievement of african americans during the high school transition. *Contemp. Educ. Psychol.*, 31(1):44–63, January 2006.

[64] Allyson F Hadwin, John C Nesbit, Dianne Jamieson-Noel, Jillianne Code, and Philip H Winne. Examining trace data to explore self-regulated learning. *Metacognition Learning*, 2(2-3):107–124, December 2007.

[65] Joseph F Hair, William C Black, Babin Barry J., and Anderson Rolph E. *Multivariate Data Analysis, Eighth Edition*. Cengage Learning, EMEA, 2018.

[66] Michael N Hallquist and Joshua F Wiley. MplusAutomation: An R package for facilitating Large-Scale latent variable analyses in mplus. *Struct. Equ. Modeling*, 25(4):621–638, January 2018.

[67] Spencer E Harpe. How to analyze likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, 7(6):836–850, November 2015.

[68] Caroline O Hart, Christian E Mueller, Royal, Kenneth D., and Martin H Jones. Achievement goal validation among african american high school students: CFA and rasch results. *J. Psychoeduc. Assess.*, 31(3):284–299, June 2013.

[69] Daphne Hijzen, Monique Boekaerts, and Paul Vedder. Exploring the links between students' engagement in cooperative learning, their goal preferences and appraisals of instructional conditions in the classroom. *Learning and Instruction*, 17(6):673–687, December 2007.

[70] T George Hornby, Christopher E Henderson, Carey L Holleran, Linda Lovell, Elliot J Roth, and Jeong Hoon Jang. Stepwise regression and latent profile analyses of locomotor outcomes poststroke. *Stroke*, 51(10):3074–3082, October 2020.

[71] Matt C Howard and Michael E Hoffman. Variable-Centered, Person-Centered, and Person-Specific approaches: Where theory meets the method. *Organizational Research Methods*, 21(4):846–876, October 2018.

[72] Li-tze Hu and Peter M Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Modeling*, 6(1):1–55, January 1999.

[73] Chris S Hulleman, Sheree M Schrager, Shawn M Bodmann, and Judith M Harackiewicz. A meta-analytic review of achievement goal measures: different labels for the same constructs or different constructs with similar labels? *Psychol. Bull.*, 136(3):422–449, May 2010.

[74] Dirk Ifenthaler. Determining the effectiveness of prompts for self-regulated learning in problem-solving scenarios. *Journal of Educational Technology & Society*, 15(1):38–52, 2012.

[75] Leong Yeok Jang and Woon Chia Liu. 2×2 achievement goals and achievement emotions: a cluster analysis of students' motivation. *European Journal of Psychology of Education*, 27(1):59–76, March 2012.

[76] Stefan Janke and Oliver Dickhäuser. A neglected tenet of achievement goal theory: Associations between life aspirations and achievement goal orientations. *Pers. Individ. Dif.*, 142:90–99, May 2019.

[77] Hogyeong Jeong, Amit Gupta, Rod Roscoe, John Wagster, Gautam Biswas, and Daniel Schwartz. Using hidden markov models to characterize student behaviors in Learning-by-Teaching environments. In *Intelligent Tutoring Systems*, pages 614–625. Springer Berlin Heidelberg, 2008.

[78] Il-Hyun Jo, Dongho Kim, and Meehyun Yoon. Analyzing the log patterns of adult learners in LMS using learning analytics. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, LAK '14, pages 183–187, New York, NY, USA, 2014. ACM.

[79] Stuart A Karabenick, Michael E Woolley, Jeanne M Friedel, Bridget V Ammon, Julianne Blazevski, Christina Rhee Bonney, Elizabeth D E Groot, Melissa C Gilbert, Lauren Musu,

Toni M Kempler, and Kristin L Kelly. Cognitive processing of Self-Report items in educational research: Do they think what we mean? *Educ. Psychol.*, 42(3):139–151, July 2007.

[80] Stuart A Karabenick and Akane Zusho. Examining approaches to research on self-regulated learning: conceptual and methodological considerations. *Metacognition and Learning*, 10(1):151–163, April 2015.

[81] David A Kenny, Burcu Kaniskan, and D Betsy McCoach. The performance of RMSEA in models with small degrees of freedom. *Sociol. Methods Res.*, 44(3):486–507, August 2015.

[82] Jasmine Kim and Panayiota Kendeou. Knowledge transfer in the context of refutation texts. *Contemporary Educational Psychology*, page 102002, 2021.

[83] Alison King. Facilitating elaborative learning through guided Student-Generated questioning. *Educ. Psychol.*, 27(1):111–126, January 1992.

[84] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, and Others. Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, pages 87–90. books.google.com, 2016.

[85] Richard Koestner, Nancy Otis, Theodore A Powers, Luc Pelletier, and Hugo Gagnon. Autonomous motivation, controlled motivation, and goal progress. *J. Pers.*, 76(5):1201–1230, October 2008.

[86] Vitomir Kovanovic, Dragan Gašević, Shane Dawson, Srećko Joksimovic, Ryan S Baker, and Marek Hatala. Does time-on-task estimation matter? implications on validity of learning analytics findings. *1*, 2(3):81–110, 2015.

[87] Bracha Kramarski and Zehavit Kohen. Promoting preservice teachers' dual self-regulation roles as learners and as teachers: Effects of generic vs. specific prompts. *Metacognition and Learning*, 12(2):157–191, 2017.

[88] Ulrike-Marie Krause and Robin Stark. Reflection in example- and problem-based learning: effects of reflection prompts, feedback and cooperative learning. *Educ. Res. Eval.*, 23(4):255–272, November 2010.

[89] Andrew E Krumm, Andrew Coulson, and Julie Neisler. Defining productive struggle in ST math: Implications for developing indicators of learning behaviors and strategies in digital learning. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery, 2022.

[90] Emily R Lai. Metacognition: A literature review. *Always learning: Pearson research report*, 24, 2011.

[91] Brett Laursen and Erika Hoff. Person-Centered and Variable-Centered approaches to longitudinal data. *Merrill. Palmer. Q.*, 52(3):377–389, 2006.

[92] Chun-Yi Lee and Ming-Puu Chen. A computer game as a context for non-routine mathematical problem solving: The effects of type of question prompt and level of prior knowledge. *Comput. Educ.*, 52(3):530–542, April 2009.

[93] Magdeleine D N Lew and Henk G Schmidt. Self-reflection and academic performance: is there a relationship? *Adv. Health Sci. Educ. Theory Pract.*, 16(4):529–545, October 2011.

[94] Xiaodong Lin. Designing metacognitive activities. *Educ. Technol. Res. Dev.*, 49(2):23–40, 2001.

[95] Xiaodong Lin and James D Lehman. Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 36(7):837–858, 1999.

[96] Elizabeth A Linnenbrink and Paul R Pintrich. Achievement goal theory and affect: An asymmetrical bidirectional model. *Educ. Psychol.*, 37(2):69–78, June 2002.

[97] Jacob A. Long. *jtools: Analysis and Presentation of Social Scientific Data*, 2020. R package version 2.1.0.

[98] Wenshu Luo, Scott G Paris, David Hogan, and Zhiqiang Luo. Do performance goals promote learning? a pattern analysis of singapore students' achievement goals. *Contemp. Educ. Psychol.*, 36(2):165–176, April 2011.

[99] Martin L Maehr and John G Nicholls. Culture and achievement motivation: A second look. *Studies in cross-cultural psychology*, 2:221–267, 1980.

[100] Herbert W Marsh, Kit-Tai Hau, and Zhonglin Wen. In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and bentler's (1999) findings. *Struct. Equ. Modeling*, 11(3):320–341, July 2004.

[101] Samiha Marwan, Joseph Jay Williams, and Thomas Price. An evaluation of the impact of automated programming hints on performance and learning. In *Proceedings of the 2019 ACM Conference on International Computing Education Research*, ICER '19, pages 61–70, New York, NY, USA, July 2019. Association for Computing Machinery.

[102] Samuel Messick. Validity. *ETS Research Report Series*, 1987(2):i–208, December 1987.

[103] Carol Midgley, Avi Kaplan, Michael Middleton, Martin L Maehr, Tim Urdan, Lynley H Anderman, Eric Anderman, and Robert Roeser. The development and validation of scales assessing students' achievement goal orientations. *Contemp. Educ. Psychol.*, 23(2):113–131, April 1998.

[104] Carol Midgley, Martin L Maehr, Ludmila Z Hruda, Eric Anderman, Lynley Anderman, Kimberley E Freeman, T Urdan, and Others. Manual for the patterns of adaptive learning scales. *Ann Arbor: University of Michigan*, 2000.

[105] R J Mislevy, L S Steinberg, and others. Focus article: On the structure of educational assessments. *research and perspectives*, 2003.

[106] Robert J Mislevy and Geneva D Haertel. Implications of evidence-centered design for educational testing. *Educ. Meas. Issu. Pr.*, 25(4):6–20, April 2007.

[107] Sandra Monteiro, Gail M Sullivan, and Teresa M Chan. Generalizability theory made simple(r): An introductory primer to g-studies. *J. Grad. Med. Educ.*, 11(4):365–370, August 2019.

[108] Jennifer A Moon. *A handbook of reflective and experiential learning: Theory and practice*. Routledge, London, England, April 2013.

[109] Pedro J Muñoz-Merino, José A Ruipérez Valiente, and Carlos Delgado Kloos. Inferring higher level learning information from low level data for the khan academy platform. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, LAK '13, pages 112–116, New York, NY, USA, April 2013. Association for Computing Machinery.

[110] Bengt Muthen and Linda K Muthen. Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcohol. Clin. Exp. Res.*, 24(6):882–891, June 2000.

[111] Bengt O Muthén. Latent variable mixture modeling. In *New developments and techniques in structural equation modeling*, pages 21–54. Psychology Press, 2001.

[112] Bengt O Muthén, Stephen H C du Toit, and Damir Spisic. Robust interference using weighted least squares and quadratic estimating equations in the latent variable modeling with categorical and continuous outcomes. *Unpublished manuscript, University of California, Los Angeles, USA*, 1997.

[113] Bengt O Muthén and Linda Muthén. *Mplus*. Chapman and Hall/CRC, 2017.

[114] Lap Trung Nguyen and Mitsuru Ikeda. The effects of eportfolio-based learning model on student self-regulated learning. *Active Learning in Higher Education*, 16(3):197–209, November 2015.

[115] John G Nicholls. Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychol. Rev.*, 91(3):328–346, July 1984.

[116] Timothy J Nokes, Robert G M Hausmann, Kurt VanLehn, and Sophia Gershman. Testing the instructional fit hypothesis: the case of self-explanation prompts. *Instr. Sci.*, 39(5):645–666, September 2011.

[117] Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Adv. Health Sci. Educ. Theory Pract.*, 15(5):625–632, December 2010.

[118] Karen L Nylund, Tihomir Asparouhov, and Bengt O Muthén. Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Struct. Equ. Modeling*, 14(4):535–569, October 2007.

[119] Karen Nylund-Gibson and Andrew Young Choi. Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, 4(4):440–461, December 2018.

[120] Daniel Oberski. Mixture models: Latent profile and latent class analysis. In Judy Robertson and Maurits Kaptein, editors, *Modern Statistical Methods for HCI*, pages 275–287. Springer International Publishing, Cham, 2016.

[121] Rebecca O'Rourke. The learning journal: from chaos to coherence. *Assessment & Evaluation in Higher Education*, 23(4):403–413, December 1998.

[122] Cindy Paans, Inge Molenaar, Eliane Segers, and Ludo Verhoeven. Temporal variation in children's self-regulated hypermedia learning. *Comput. Human Behav.*, 96:246–258, July 2019.

[123] Ernesto Panadero. A review of self-regulated learning: Six models and four directions for research. *Front. Psychol.*, 8:422, April 2017.

[124] Dena A Pastor, Kenneth E Barron, B J Miller, and Susan L Davis. A latent profile analysis of college students' achievement goal orientation. *Contemp. Educ. Psychol.*, 32(1):8–47, January 2007.

[125] Paul R Pintrich. A manual for the use of the motivated strategies for learning questionnaire (MSLQ). 1991.

[126] Paul R Pintrich. Chapter 14 – the role of goal orientation in Self-Regulated learning. In Monique Boekaerts, Paul R Pintrich, and Moshe Zeidner, editors, *Handbook of Self-Regulation*, pages 451–502. Academic Press, San Diego, 2000.

[127] Caroline Julia Pulfrey, Maarten Vansteenkiste, and Aikaterina Michou. Under pressure to achieve? the impact of type and style of task instructions on student cheating. *Front. Psychol.*, 10:1624, August 2019.

[128] Minna Puustinen and Lea Pulkkinen. Models of self-regulated learning: A review. *Scandinavian Journal of Educational Research*, 45(3):269–286, September 2001.

[129] Daniela Raccanello and Margherita Brondino. Assessing primary and secondary students' achievement goals for italian and mathematics domains: The italian version of the achievement goal Questionnaire-Revised (AGQ-R). *BPA-Applied Psychology Bulletin (Bollettino di Psicologia Applicata)*, 64(277), 2016.

[130] Catherine F Ratelle, Frédéric Guay, Robert J Vallerand, Simon Larose, and Caroline Senécal. Autonomous, controlled, and amotivated types of academic motivation: A person-oriented analysis. *J. Educ. Psychol.*, 99(4):734–746, November 2007.

[131] Richard Remedios and John T E Richardson. Achievement goals in adult learners: evidence from distance education. *Br. J. Educ. Psychol.*, 83(Pt 4):664–685, December 2013.

[132] Ido Roll, Vincent Aleven, Bruce M McLaren, and Kenneth R Koedinger. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and instruction*, 21(2):267–280, 2011.

[133] Jeremy Roschelle, Ricardo Rosas, and Miguel Nussbaum. Towards a design framework for mobile computer-supported collaborative learning. In *Computer Supported Collaborative Learning 2005: The Next 10 Years!*, pages 520–524. Routledge, 2017.

[134] Jeremy Roschelle and Stephanie D Teasley. The construction of shared knowledge in collaborative problem solving. In *Computer Supported Collaborative Learning*, pages 69–97. Springer Berlin Heidelberg, 1995.

[135] Richard M Ryan and Edward L Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.*, 55(1):68–78, January 2000.

[136] Javier Sánchez Rosas. validation of the achievement goal questionnaire – revised in argentinean university students (A-AGQ-R). *Int. J. Psychol. Res. (Medellin)*, 8(1):10–23, January 2015.

[137] Thomas A Schmitt. Current methodological considerations in exploratory and confirmatory factor analysis. *J. Psychoeduc. Assess.*, 29(4):304–321, August 2011.

[138] James B Schreiber. Latent class analysis: An example for reporting results. *Res. Social Adm. Pharm.*, 13(6):1196–1201, November 2017.

[139] Dale H Schunk and Barry J Zimmerman. Self-regulated learning and performance: an introduction and an overview. In *Handbook of Self-Regulation of Learning and Performance*, pages 15–26. Routledge, 2011.

[140] Malte Schwinger and Elke Wild. Prevalence, stability, and functionality of achievement goal profiles in mathematics from third to seventh grade. *Contemp. Educ. Psychol.*, 2012.

[141] Silke Schworm and Alexander Renkl. Learning by solved example problems: Instructional explanations reduce self-explanation activity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 24. escholarship.org, 2002.

[142] Gerard H Seijts, Gary P Latham, Kevin Tasa, and Brandon W Latham. Goal setting and goal orientation: An integration of two different yet related literatures. *AMJ*, 47(2):227–239, April 2004.

[143] Corwin Senko, Chris S Hulleman, and Judith M Harackiewicz. Achievement goal theory at the crossroads: Old controversies, current challenges, and new directions. *Educ. Psychol.*, 46(1):26–47, January 2011.

[144] Corwin Senko and Katie L Tropiano. Comparing three models of achievement goals: Goal orientations, goal standards, and goal complexes. *J. Educ. Psychol.*, 108(8):1178–1192, November 2016.

[145] Lilly Shanahan, William E Copeland, Carol M Worthman, Alaattin Erkanli, Adrian Angold, and E Jane Costello. Sex-differentiated changes in c-reactive protein from ages 9 to 21: the contributions of bmi and physical/sexual maturation. *Psychoneuroendocrinology*, 38(10):2209–2217, 2013.

[146] Kennon M Sheldon and Andrew J Elliot. Goal striving, need satisfaction, and longitudinal well-being: the self-concordance model. *J. Pers. Soc. Psychol.*, 76(3):482–497, March 1999.

[147] Kennon M Sheldon and Tim Kasser. Coherence and congruence: two aspects of personality integration. *J. Pers. Soc. Psychol.*, 68(3):531–543, March 1995.

[148] Melody Siadaty, Dragan Gašević, Jelena Jovanović, Kai Pata, Nikola Milikić, Teresa Holocher-Ertl, Zoran Jeremić, Liaqat Ali, Aleksandar Giljanović, and Marek Hatala. Self-regulated workplace learning: A pedagogical framework and semantic web-based environment. *Journal of Educational Technology & Society*, 15(4):75–88, 2012.

[149] George Siemens and Phil Long. Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, 46(5):30, 2011.

[150] James H Steiger. Understanding the limitations of global fit assessment in structural equation modeling. *Pers. Individ. Dif.*, 42(5):893–898, May 2007.

[151] Robert J. Sternberg. Metacognition, abilities, and developing expertise: What makes an expert student? 26(1):127–140, 1998.

[152] Kamden K Strunk. A factor analytic examination of the achievement goal Questionnaire–Revised supports a Three-Factor model. *Psychol. Rep.*, 115(2):400–414, October 2014.

[153] Ana N A Susac, Andreja Bubic, Jurica Kaponja, Maja Planinic, and Marijan Palmovic. Eye movements reveal students' strategies in simple equation solving. *International Journal of Science and Mathematics Education*, 12(3):555–577, 2014.

[154] Gita Taasoobshirazi and Shanshan Wang. The performance of the SRMR, RMSEA, CFI, and TLI: An examination of sample size, path size, and degrees of freedom. *Journal of Applied Quantitative Methods*, 2016.

[155] Pierre Tchounikine, Nikol Rummel, and Bruce M McLaren. Computer supported collaborative learning and intelligent tutoring systems. In Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi, editors, *Advances in Intelligent Tutoring Systems*, pages 447–463. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[156] Sigmund Tobias and Howard T Everson. Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring. research report no. 2002-3. *College Entrance Examination Board*, 2002.

[157] Tim Urdan. Using multiple methods to assess students' perceptions of classroom goal structures. *Eur. Psychol.*, 9(4):222–231, January 2004.

[158] Tim Urdan and Miranda Mestas. The goals behind performance goals. *J. Educ. Psychol.*, 98(2):354–365, May 2006.

[159] Ellen L Usher and Dale H Schunk. Social cognitive theoretical perspective of Self-Regulation. In *Handbook of Self-Regulation of Learning and Performance*, chapter 2. Routledge, 2018.

[160] Gerard van den Boom, Fred Paas, Jeroen J G van Merriënboer, and Tamara van Gog. Reflection prompts and tutor feedback in a web-based learning environment: effects on students' self-regulated learning competence. *Comput. Human Behav.*, 20(4):551–567, July 2004.

[161] Maarten Vansteenkiste, Willy Lens, Andrew J Elliot, Bart Soenens, and Athanasios Mouratidis. Moving the achievement goal approach one step forward: Toward a systematic examination of the autonomous and controlled reasons underlying achievement goals. *Educ. Psychol.*, 49(3):153–174, July 2014.

[162] Maarten Vansteenkiste, Stijn Smeets, Bart Soenens, Willy Lens, Lennia Matos, and Edward L Deci. Autonomous and controlled regulation of performance-approach goals: Their relations to perfectionism and educational outcomes. *Motiv. Emot.*, 34(4):333–353, December 2010.

[163] Marcel V J Veenman. Learning to self-monitor and self-regulate. In Richard E Mayer and Patricia Alexander, editors, *Handbook of research on learning and instruction*, pages 197–218. New York: Routledge, 2011.

[164] Marcel V J Veenman and Dorit van Cleef. Measuring metacognitive skills for mathematics: students' self-reports versus on-line assessment methods. *ZDM*, pages 1–11, 2018.

[165] Jeroen K Vermunt. Latent class analysis of complex sample survey data: Application to dietary data. *Journal of the American Statistical Association*, 97(459):736–737, 2002.

[166] Olga Viberg, Mathias Hatakka, Olof Bälter, and Anna Mavroudi. The current landscape of learning analytics in higher education. *Comput. Human Behav.*, 89:98–110, December 2018.

[167] Ronald L Wasserstein and Nicole A Lazar. The ASA statement on p-values: Context, process, and purpose. *Am. Stat.*, 70(2):129–133, April 2016.

[168] Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. Moving to a world beyond "p < 0.05". *Am. Stat.*, 73(sup1):1–19, March 2019.

[169] Bridget E Weller, Natasha K Bowen, and Sarah J Faubert. Latent class analysis: A guide to best practice. *J. Black Psychol.*, 46(4):287–311, May 2020.

[170] Philip H Winne. Improving measurements of Self-Regulated learning. *Educ. Psychol.*, 45(4):267–276, October 2010.

[171] Philip H Winne. Issues in researching self-regulated learning as patterns of events. *Metacognition and Learning*, 9(2):229–237, August 2014.

[172] Philip H Winne. Chapter 21: Learning analytics for self-regulated learning. In Charles Lang, George Siemens, Alyssa Wise, and Dragan Gašević, editors, *Handbook of Learning Analytics*. solaresearch.org, 2017.

[173] Philip H Winne. Construct and consequential validity for learning analytics based on trace data. *Comput. Human Behav.*, 112:106457, November 2020.

[174] Philip H Winne and Allyson F Hadwin. Studying as self-regulated learning. *Metacognition in educational theory and practice*, 93:27–30, 1998.

[175] Philip H Winne and Dianne Jamieson-Noel. Exploring students' calibration of self reports about study tactics and achievement. *Contemp. Educ. Psychol.*, 27(4):551–572, 2002.

[176] Phillip H Winne. Self-regulated learning viewed from models of information processing. *Self-regulated learning and academic achievement: Theoretical perspectives*, 2:153–189, 2001.

[177] Christopher A Wolters. Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *J. Educ. Psychol.*, 96(2):236–250, June 2004.

[178] Christopher A Wolters and Paul R Pintrich. Contextual differences in student motivation and self-regulated learning in mathematics, english, and social studies classrooms. *Instructional science*, 26(1):27–47, 1998.

[179] Christopher A Wolters and Sungjun Won. Validity and the use of Self-Report questionnaires to assess Self-Regulated learning. In *Handbook of Self-Regulation of Learning and Performance*, chapter 20. Routledge, 2018.

[180] Stephanie Virgine Wormington and Lisa Linnenbrink-Garcia. A new look at multiple goal pursuit: the promise of a Person-Centered approach. *Educ. Psychol. Rev.*, 29(3):407–445, February 2016.

[181] Zihan Wu, Barbara Ericson, and Christopher Brooks. Regex parsons: Using horizontal parsons problems to scaffold learning regex. In *21st Koli Calling International Conference on Computing Education Research*, pages 1–3, 2021.

[182] Yan Xia and Yanyun Yang. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behav. Res. Methods*, 51(1):409–428, February 2019.

[183] Cristina D Zepeda, J Elizabeth Richey, Paul Ronevich, and Timothy J Nokes-Malach. Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation: An in vivo study. *Journal of Educational Psychology*, 107(4):954, 2015.

[184] Ying Zhang, Rainer Watermann, and Annabell Daniel. Are multiple goals in elementary students beneficial for their school achievement? a latent class analysis. *Learn. Individ. Differ.*, 51:100–110, October 2016.

[185] Mingming Zhou and Philip H Winne. Modeling academic achievement by self-reported versus traced goal orientation. *Learning and Instruction*, 22(6):413–419, December 2012.

[186] Yiqiu Zhou, Juan Miguel Andres-Bray, Stephen Hutt, Korinn Ostrow, and Ryan S Baker. A comparison of hints vs. scaffolding in a MOOC with adult learners. In *Artificial Intelligence in Education*, pages 427–432. Springer International Publishing, 2021.

[187] Barry J Zimmerman. Self-Efficacy: An essential motive to learn. *Contemp. Educ. Psychol.*, 25(1):82–91, January 2000.

[188] Barry J Zimmerman. Investigating Self-Regulation and motivation: Historical background, methodological developments, and future prospects. *Am. Educ. Res. J.*, 45(1):166–183, March 2008.