Pieces in the Puzzle of Language Learning

On the Roles of Morphological Knowledge, App-Based Implicit Learning and Child-App Interactions

Jarl Kleppe Kristensen

Thesis submitted for the degree of Ph.D.

Centre for Educational Measurement (CEMO)
Faculty of Educational Sciences
University of Oslo

2023

Series of dissertations submitted to the
Faculty of Educational Sciences, University of Oslo
No. 377

ISSN 1501-8962

Cover: UiO.
Print production: Graphic center, University of Oslo.

Acknowledgements

First I would like to thank my brilliant supervisors, Björn Andersson and Janne von Koss Torkildsen. You are both fantastic in your own right, but the sum was even greater than its parts. Your combined knowledge and experience gave me the opportunity to develop broadly within diverse areas of statistics, linguistics, critical thinking, writing, and much, much more. Thank you so much for helping and guiding me, for making me push myself to reach goals I did not even know I had, for being flexible and understanding, and for making me believe in my ability to do important and meaningful research. I would also like to thank Siri Steffensen Bratlie for the collaboration and support.

Next, I am very grateful for the opportunity to do my doctoral research at the Centre for Educational Measurement. Starting out, statistical papers with all their alphas, betas, gammas and worse were very much Greek to me. It has been a challenging, but highly rewarding journey to where I stand today; a journey I do not think I could have made so successfully anywhere else. I owe all the staff at CEMO many thanks for the support, the interest, and for making every day at the office a joy. A special thanks to Ronny Scherer for all the valuable things you have taught me about structural equation modeling, and particularly how to teach it. It has been a pleasure to be part of the measurement models team! Also a big thanks to my fellow Ph.D. candidates for helping make these years memorable, for sharing the experience, and for the good times both at and outside of work. To Henrik, thanks for making every day a bit less appropriate and a lot more fun! And to Yuriko, thank you for always being there, and for sharing all the ups and downs, cakes and chelas. And to a very special former CEMOnian, Alexandra, you may have left the country, but not my heart.

Another big thank you goes to the Vocabulary Learning Challenge project team. You all played big parts on my road to CEMO and a Ph.D. Being a research assistant in the project gave me a lot of valuable experience, and countless opportunities to learn and develop, and I appreciate your support and encouragement to continue on an academic path. I also want to thank Luc Paquette for invaluable advice and great discussions at my midway assessment and final reading.

Further, I owe a lot of gratitude to all of my family and friends. To my parents, Kjersti and Rolf Morten, for their love and constant support. Thank you for all your help and for allowing me to follow all the strange paths that lead me here. Thank you for bestowing on me the love of books, of fairytales and science, of computers and nature. On the note of computers, I would also like to thank my brother, Tord, for teaching me the value of

“googling it” - the solution is out there, and learning is best achieved by doing. To Tom Kenneth, thank you for supporting and encouraging me, and for everything you have done for me. To Jean, thank you for reminding me of the path I wanted to take, and for the peaches on the balcony. And, to Steffen, thanks for all the show! There are many more I should have thanked, but the list is far too long.

Last but not least, to my grandparents, Unn and Reidulf, and Oddveig and Odd, for all the love and happiness. You have all played, and still play, immense parts in my life. Finally, a special thanks to my grandfather Odd for instilling in me the great value of education.

Jarl Kleppe Kristensen
Oslo, March 2024

Abstract

Vocabulary, knowledge of word meanings, is essential for comprehension and production of oral and written language. It is, however, an unconstrained skill that covers a vast content area. This makes explicit teaching of every single word in a language an impossible task. Additionally, teaching the meanings of specific words is unlikely to provide transfer of knowledge to other words. We need to target knowledge and skills that are generalizable. Furthermore, educational games and apps have come to play large parts in education. While educational software can provide efficient and effective parts of education, learning gains depend on how children interact with the software. The overarching aim of this thesis is to unravel some of the pieces in the puzzle of app-based language learning, focusing specifically on morphological knowledge, implicit learning, and child-app interaction.

In Article 1, we found evidence that implicit app-based morphological training provides generalizable and durable effects on children's word knowledge. Article 2 focused on morphological knowledge, and provides support of a multidimensional view, where morphological knowledge consists of morphological awareness, morphological decoding and morphological analysis. Article 3 examined repeated mistakes, a child-app interaction pattern, and found that children with less prior knowledge are more likely to repeat mistakes. Furthermore, children with a high propensity to repeat mistakes have lower expected learning gains than low-propensity children with the same pre-test score. Finally, Article 4 investigated task and child covariates of repeated mistakes. We found the number of repeated mistakes made relates to task type and task position in a session, as well as children's receptive knowledge of morphologically complex words and non-verbal ability.

In the extended abstract, I discuss these four articles, highlighting key findings and contributions, and providing recommendations for both research and practice. To sum up briefly, an implicit app-based morphological intervention can contribute to language learning, but we need to pay close attention to how children interact with the app, as some interaction patterns may hinder children's learning.

Table of Contents

List of Articles

1) Torkildsen, J. V. K., Bratlie, S. S., Kristensen, J. K., Gustafsson, J.-E., Lyster, S.-A. H., Snow, C., Hulme, C., Mononen, R.-M., Næss, K.-A. B., López-Pedersen, A., Wie, O. B., & Hagtvet, B. (2022). App-based morphological training produces lasting effects on word knowledge in primary school children: A randomized controlled trial. *Journal of Educational Psychology*, *114*(4), 833–854. https://doi.org/10.1037/edu00 00688

2) Kristensen, J. K., Andersson, B., Bratlie, S. S., & Torkildsen, J. V. K. (2023). Dimensionality of Morphological Knowledge—Evidence from Norwegian Third Graders. *Reading Research Quarterly*, *58*(3), 406-424. https://doi.org/10.1002/rrq.497

3) Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2024). Repeated Mistakes in App-Based Language Learning: Persistence and Relation to Learning Gains. *Computers & Education*. https://doi.org/10.1016/j.compedu.2023.104966

4) Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2023). *Who repeats mistakes and when? Task and child covariates of repeated mistakes in app-based learning* [Manuscript in preparation]. Centre for Educational Measurement, University of Oslo.

Part I: Extended Abstract

Chapter 1: Introduction

Ash nazg durbatulûk, Ash nazg gimbatul,
Ash nazg thrakatulûk, Agh burzum-ishi krimpatul
(J.R.R. Tolkien, *The Lord of the Rings*)

The quote may seem a strange place to begin a doctoral thesis on app-based language learning. Yet, in many ways this was how my journey began. My parents reading The Lord of the Rings to me was a major catalyst for my later interest in language. Through Tolkien's languages, including the black speech in which the quote is written, I got interested in how languages were developed and built up. Looking more closely at the quote, we can see that there are common elements recurring throughout the verse. Clearly recurring is the phrase *ash nazg*, or "one ring". A bit more hidden are *tul* (them) and *ûk* (all). These are suffixes added on to the bases of the verbs *durba*, *gimba*, *thraka*, and *krimpa* (rule, find, bring, and bind) to signal that the actions are performed on "them" and "them all".[1]

These small meaningful parts of words are morphemes. They are important building blocks of languages, and knowing their meanings, functions and the processes by which they are combined can make language comprehension a much easier task. Teaching morphemes is not necessarily such an easy task, though, and this thesis aims to unravel some of the constituents of successful morphological instruction.

1.1 Background and Relevance of the Thesis

Education is becoming ever more digitalized and technology-dependent, for better or worse. Recent years have seen a vast number of educational apps and games becoming integral parts of students' learning resources (e.g., Montazami et al., 2022). While educational software can provide excellent resources for learning, it is crucial that we understand what works and what does not, as well as how and why different approaches may support or hinder learning. Broadly speaking, my goal with this thesis is to add to the knowledge of what works when providing digital educational interventions, as well as factors associated with different learning outcomes. My focus is on app-based language learning. As stated in the title of the thesis, I aim to reveal some of the puzzle pieces related to language learning. Specifically, my research revolves around a language learning app based on principles of morphology and

[1] With reservations about possible missteps in word segmentation, as black speech is sparsely described.

implicit learning. Furthermore, I examine how children's app interaction patterns relate to learning gains.

Morphology, in linguistic terms, refers to the study of morphemes, the smallest meaning-bearing elements of language. For example, the word *misplace* consists of two morphemes – *mis-* and *place*. Morphology and morphological knowledge is discussed in detail in section 2.2. *Implicit learning* refers to learning that happens without explicit instruction or intent. We can, without conscious effort, pick up statistical patterns in our environment, such as elements that frequently co-occur in a stream of spoken language. With this information, we can categorize and assign meaning to different elements. This is covered in section 2.3. Finally, *child-app interaction* refers to the way children approach and use an app, from engagement and learning strategies to specific actions and the time spent on each action. My research focuses specifically on an interaction pattern where children repeat the same mistake multiple times during a task. Child-app interactions, and repeated mistakes, is the focus of section 2.4.

This thesis, while representing an independent doctoral research project, is closely connected to a larger project, the Vocabulary Learning Challenge[2] (VLC). The overarching goal of the VLC project was to develop and test an app-based intervention to support primary school children's vocabulary development. Vocabulary is an elementary part of language, and vocabulary knowledge is both concurrently and predictively associated with reading comprehension, writing and academic achievement more generally (e.g. Hulme et al., 2020; Ricketts et al., 2020; Vellutino et al., 2007; Wagner & Quinn, 2019). Since vocabulary is a basic and necessary requirement to understand oral and written language, efficient interventions are essential to support vocabulary development.

There are two common challenges related to vocabulary interventions, namely lack of generalization and fade-out effects. Vocabulary is an unconstrained area of language in terms of the vast content space (Paris, 2005; Snow & Matthews, 2016). It is impossible to teach every word of a language explicitly, especially within the confines of a time-limited intervention. Thus, interventions need to provide generalizable knowledge in order to promote understanding that goes beyond the specific intervention content. Focusing on a selection of words frequently lead to children learning those specific words, but the knowledge is rarely generalized to words not included in the intervention. Additionally, while vocabulary interventions focusing on specific words may have immediate effects in terms of learning

[2] https://www.uv.uio.no/isp/english/research/projects/the-vocabulary-learning-challenge-vlc-/index.html

gains, the control group often catches up to the experimental group relatively quickly, leading to fade-out effects (Bailey et al., 2017).

To address these challenges, the VLC project developed "Kaptein Morf", an app targeting morphology. Morphology, as mentioned, is the study of morphemes. Morphemes typically recur in many combinations. Returning to the example above, *mis-* in *misplace* also occurs in words such as *mistake*, *misconduct* and *misanthropy*. Thus, if a child knows the meaning of *mis-* in misplace, it can help them uncover the meaning of unknown words containing *mis-*. This may decrease potential fade-out effects as children continue to generalize their knowledge and apply it in new contexts (Bailey et al., 2017). The viability of using app-based morphological training as a pathway to word knowledge is addressed in Article 1 of this thesis. The article reports on the randomized controlled trial (RCT) conducted in the VLC project, and provides evidence that morphological training, delivered through a gamified app, contributes to generalizable and lasting effects on knowledge of word meanings.

Furthermore, through the remaining articles of the thesis, I expand upon the research conducted in the VLC project by doing in-depth analyses of morphological knowledge and child-app interactions. The findings add to the current knowledge in the fields of morphological knowledge and app-based learning in several important ways. First, there are substantial discrepancies in the empirical evidence provided by studies investigating the dimensionality of morphological knowledge. Knowing whether morphological knowledge is a unidimensional construct or consists of several distinct, but related, subskills is important in order to understand how we can teach morphology and how it relates to other areas of language, such as vocabulary. Knowledge of dimensionality is also crucial when developing interventions and assessments. For example, to get a complete picture of children's ability levels in multidimensional skills, we need to assess each of the dimensions. When designing interventions, we must similarly make decisions on which dimensions to target, and whether learning might transfer across dimensions. These issues are addressed in Article 2.

Articles 3 and 4 examine how children interacted with the app, the association between interaction and learning gains, and how child-app interaction relate to different task and child characteristics. Article 3 focuses on a specific response pattern that is largely unexplored in the extant literature on response behavior and child-app interactions, namely repeated mistakes (described in section 2.4). We examine children's propensity to repeat mistakes and how it relates to learning gains from the morphological intervention. In Article 4, we conduct an in-depth study of repeated mistakes, examining task and child covariates of

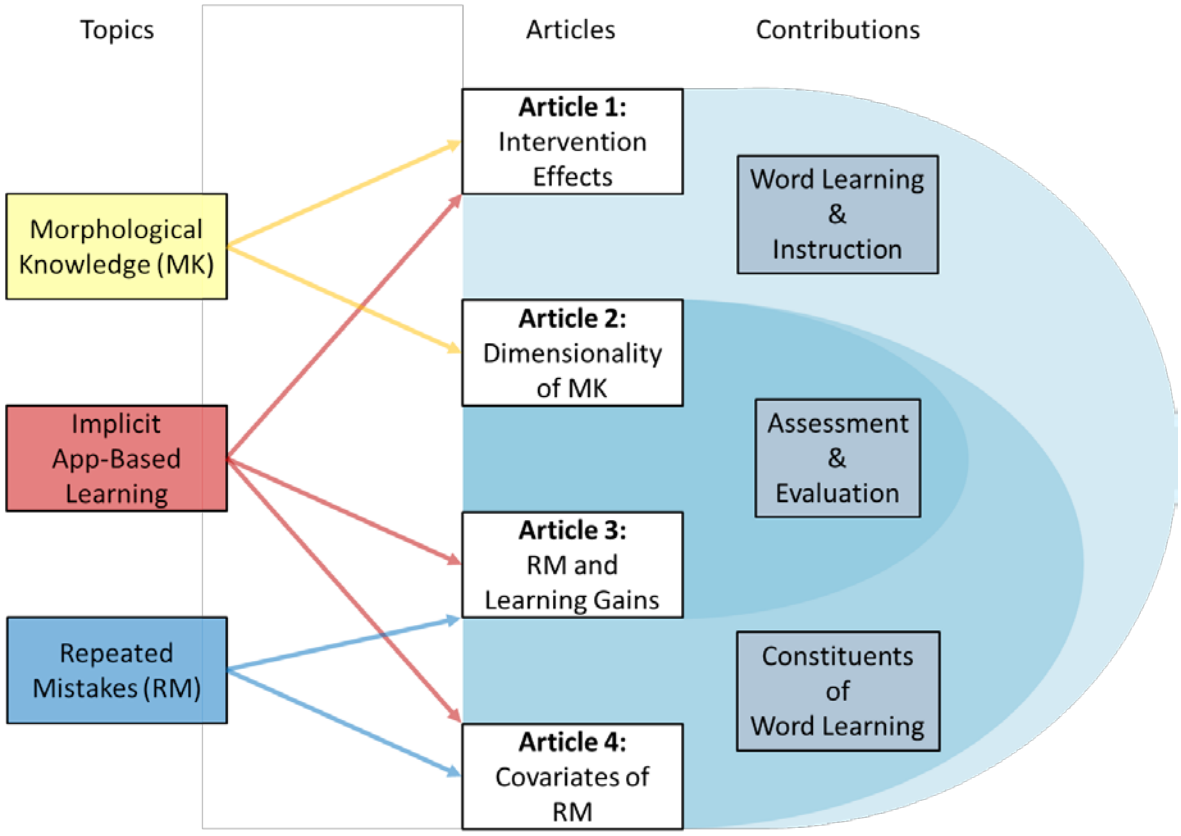
repeated mistakes to uncover mechanisms related to such behavior. Taken together, the four articles have important implications for language instruction and assessment, as well as the design and development of interventions, in particular those based on apps or other educational software.

1.2 Overarching Aim

The overarching aim of the thesis was to examine *whether* and *how* implicit app-based morphological training facilitates word learning in primary school children. The latter part on *how* relates both to the concept of morphological knowledge, its structure, and how it relates to word knowledge more generally, and to child-app interactions and how they relate to task and child characteristics, as well as to learning. Figure 1 provides an overview of the relations between the key topics of the thesis, the research articles, and their contributions to research and practice.

Figure 1

Relations between thesis topics, research articles, and contributions



Article 1 reports on the effects of the Kaptein Morf app, Article 2 examines the dimensionality of morphological knowledge, and Articles 3 and 4 investigate the propensity to repeat mistakes, its impact on learning gains, and task and person covariates of mistake repetition. All four articles contribute to the current knowledge of technology enhanced language learning and instruction, with examples of what works and what does not work. Articles 2, 3 and 4 carry important insights into constituent elements of app-based language learning, namely morphological interventions and child-app interaction, and provide suggestions for future development of language learning apps. Finally, Articles 2 and 3 provide substantial insights into assessment and evaluation, with Article 2 carrying implications for the development of morphological assessments, and Article 3 outlining important considerations for the evaluation of app-based interventions.

1.3 Outline of the Thesis

The main objective of the extended abstract is to present and discuss the research conducted in the four articles as a coherent whole. To this end, I first present the theoretical foundations of the work as a whole, covering morphological knowledge, implicit learning, and child-app interaction, including repeated mistakes, in chapter 2. In chapter 3, I present methods and methodological considerations. Key results from the four articles are presented in chapter 4, before discussing the main findings and implications of the studies in chapter 5.

Chapter 2: Theoretical Framework

This chapter consists of three parts. In section 2.1, I provide an overview of morphological knowledge and the rationale for morphological interventions. This includes the dimensionality of morphological knowledge and its relations to reading and writing generally, and vocabulary specifically. Section 2.2 gives an outline of the principles of implicit learning. Finally, in section 2.3, I review extant literature on disengagement and how it relates to our conceptualization of repeated mistakes.

2.1 Morphological Knowledge

I will begin this section by clarifying some key terms. There is substantial variation in the terminology used to describe skills and knowledge relating to morphology (e.g., Apel, 2014; Berthiaume et al., 2018). *Morphologically complex* words are words consisting of two or more morphemes. This is also referred to as *multimorphemic* words. In this thesis, I define morphological knowledge in line with Article 2, as "…the ability to recognize, understand, manipulate and produce spoken and written morphemes" (p. 407). Furthermore, I follow the definitions of Levesque et al. (2021) for the concepts morphological awareness, morphological decoding and morphological analysis. *Morphological awareness* is explicit knowledge about morphemes and the processes in which they are combined. *Morphological decoding* refers to knowledge pertaining to the written forms of morphemes, whereas *morphological analysis* refers to knowledge of morpheme meanings.

Morphemes can be divided into base words (free morphemes) and affixes (bound morphemes). Base words are morphemes that can stand alone, such as *rail*, hence the term free morphemes. Affixes, on the other hand, cannot stand alone, hence the term bound morphemes. Rather, they change the meaning and/or grammatical properties of base morpheme, like *-s* in *rails* which makes a plural, or *de-* in *derail* which changes both the meaning and the word class.

There are three processes in which morphemes are commonly combined in word formation: inflection, derivation and compounding (Gonnerman, 2018). Inflections modify the grammatical category of a word. They may for example change the tense of verbs, number of nouns, or degree of adjectives. Derivational affixes are affixes added to a base to change the meaning and/or word class of a word, e.g. *un-* in *unhappy* and *-ness* in *happiness*. Derivational affixes recur in combination with many base words, such as *happiness, sadness, greyness and suddenness*, providing possibilities for generalization of knowledge across base

words. Knowing the meaning of an affix, or how it changes the base word, can provide support when trying to work out the meaning of a new word (e.g., Crosson, et al., 2019). Finally, compounds consist of two or more base words. Closed compounds, where the base words are written as one, are highly frequent in the Norwegian language. For example, *doktorgradsavhandlingsunderkappiteloverskrift* (doctoral thesis subchapter heading), while constructed here to underline a point, is entirely viable in Norwegian. Reading lengthy compounds is an enormous task unless we can segment them into their meaning-bearing constituent parts, e.g. doktor|grad|s|av|handl|ing|s|under|kapittel|over|skrift. Similarly, being able to identify individual elements of such long words can support comprehension, as we piece together the meanings of separate parts of the word (e.g., Nagy & Townsend, 2012). While base words carry the same challenges as general vocabulary in terms of content space, there is a limited number of inflectional and derivational affixes, as well as compounding patterns. Thus, focusing on these areas of morphology provides a constrained pool of target content that may generalize to an unconstrained number of contexts.

2.2.1 Dimensionality of Morphological Knowledge

There has been much debate over the dimensionality of morphological knowledge, i.e. is it a single skill relating to morphology in general, or does it consist of several distinct, but related subskills? Over the past two decades, several attempts have been made to unravel the dimensional structure of morphological knowledge (for an overview, see table S1 in the supplementary materials of Article 2). Much of the discrepancy is likely due to how morphological knowledge has been conceptualized and measured, as is discussed in detail in Article 2. Recent studies, however, seem to agree that morphological knowledge is a multidimensional construct, although there is still considerable differences in how these dimensions are conceptualized (e.g., Apel et al., 2023; Goodwin et al., 2022; Han et al., 2022; Levesque & Deacon, 2022; Shen & Crosson, 2023; Varga et al., 2022, Wang & Zhang, 2023)

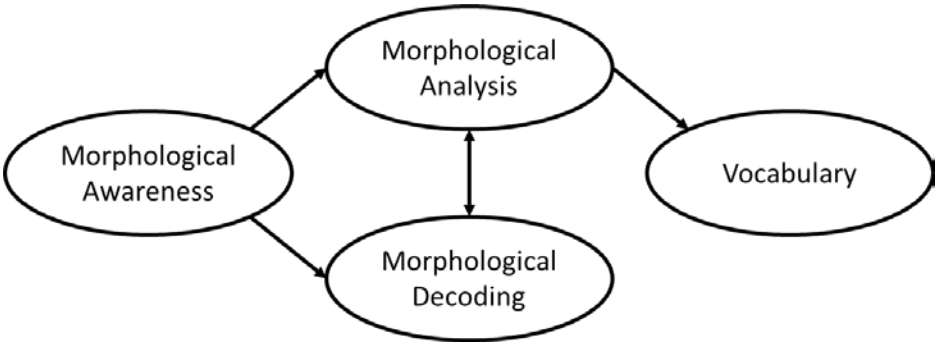
In an endeavor to provide a unified view of morphological knowledge based on current theoretical perspectives and empirical evidence, Levesque et al. (2021) proposed the Morphological Pathways Framework. In this framework, morphological knowledge consists of three separate, related skills: morphological awareness, morphological analysis and morphological decoding. Morphological awareness, or the explicit knowledge of morphemes and morphological processes, relates to literacy in different ways, through the form-based skills of morphological decoding and the meaning-based skills of morphological analysis.

Morphological decoding relates to skills such as word reading and spelling, and morphological analysis relates to vocabulary. All three skills contribute to reading

comprehension and text generation (coherent reading and writing of longer texts). Of particular interest for this thesis is the relation to vocabulary, shown in Figure 2. Levesque et al. (2021) suggests two ways in which morphological knowledge can contribute to general vocabulary. First, there is a direct relation between morphological analysis and vocabulary. Furthermore, the framework suggests that morphological awareness is indirectly linked to vocabulary through the association with morphological analysis. There is no direct association between morphological decoding and vocabulary in the framework.

Figure 2

Associations between Morphological Knowledge and Vocabulary



2.2.3 Rationale for Morphological Interventions

Several studies have found evidence that morphological instruction can enhance both word knowledge and reading development (e.g., Bowers et al., 2010; Carlisle, 2010; Goodwin & Ahn, 2010; Lyster et al., 2016; Reed, 2008). As mentioned, morphological knowledge is a constrained skill that can serve as a gateway to unconstrained skills such as vocabulary and reading comprehension. Thus, it can provide an accessible and efficient target for interventions. Bailey et al. (2017) argue that interventions should target *trifecta* skills: skills that are fundamental, malleable and would not attained if not for the intervention. Skills are *fundamental* if they are important for success in a given area, and they are *malleable* if they can be changed or influenced through intervention. In the present thesis I argue that one such trifecta skill is morphological knowledge.

Morphology is fundamental, as morphemes are the smallest meaning-bearing units of language, and provide building blocks to help us understand complex word and sentences. Morphological knowledge is also a constrained skill, since there is a limited number of affixes and compounding processes involved in word formation in any given language. Thus it

provides a set amount of building blocks that can be put together to form countless words. Since knowledge of morphemes and morphological processes can be improved, morphological knowledge represents a malleable skill. Furthermore, affixes and compounding patterns are applicable in many contexts. For example, knowing that *un*bound means *not* bound can help children understand other words such as unhappy and unburdened. Hence, morphology can provide knowledge that is generalizable to new words. This generalization can also build durable knowledge that evolves and persists after an intervention, as is addressed in Article 1.

Finally, relating to the Norwegian context, morphology is not specifically mentioned in the national curriculum for primary school children (The Norwegian Directorate for Education and Training, 2020). While primary school children do have and develop morphological knowledge to a certain extent, more morphological knowledge, and particularly knowledge pertaining to derivational affixes would be highly beneficial. For example, Nagy and Anderson (1984) found that 60% of words in English school texts are morphologically complex. The proportion is likely to be even higher in the Scandinavian languages, which are more morphologically complex than English (Kettunen, 2014). Since morphology is not specifically mentioned in the Norwegian curriculum, it is also unlikely that is receives much attention in language instruction. Thus, it may not develop to a desirable extent without intervention.

One challenge with morphological instruction is that it requires relatively high levels of metalinguistic knowledge if taught explicitly, at least for morphemes with a primarily grammatical meaning (e.g. morphemes that change a word's part of speech) and morphemes with complex semantics or multiple (often related) meanings (e.g. 'over' which can mean both 'too much' and spatially over). However, given that affixes and compounding patterns occur in many different combinations, morphology is well suited for interventions based on implicit learning.

2.3 Implicit Learning

While incidental implicit learning plays an important part in language development, from birth and throughout life (e.g., Romberg & Saffran, 2010; Erickson & Thiessen, 2015), interventions building on implicit learning depend on activities strategically designed to target specific elements and goals. When employed in interventions, implicit learning should not be incidental, but rather carefully planned and executed. For example, in a meta-analysis of 97 studies, Boeve et al. (2023) found that different statistical learning paradigms relate

differentially to language outcomes. Thus, the intervention design can have a large impact on learning gains. In order to facilitate the pattern learning, the target needs to appear often, and in many different contexts. With high variability in non-target elements, the target becomes the most salient feature of the input, and we are more likely to internalize it. Torkildsen et al. (2013) found that as many as 24 different variations may be necessary to facilitate implicit learning. Furthermore, implicit learning depends on continuous accumulation of information. Hence, disengagement or other interruptions may hinder efficient implicit learning. Another potential threat to implicit learning is that "all input is input" (Plante & Gómez, 2018). This means that we can potentially learn from any element in the input, and if we are exposed to large amounts of incorrect input, this may become the salient feature and thus what we recall at a later stage.

Educational apps are well suited to deliver interventions based on implicit learning, since they can accommodate tasks with high variability in non-target elements, and do not necessarily need large amounts of explicit instructions or explanations in order for tasks to make sense. However, in line with the comments about potential threats to implicit learning, the success of such apps depend on how children interact with them.

2.4 Engagement and Child-App Interaction

The way children interact with educational technology has important implications for the learning process. It is well established that engagement is associated with positive learning outcomes (see Fredricks et al., 2004 for an overview). On the other hand, disengagement is associated with lower learning gains and higher drop-out rates. The spectrum of disengagement ranges from total (off-task, not interacting) to partial (interacting with the task, but not the content, e.g. gaming the system, Baker et al., 2004).

Engagement can be separated into three distinct, but related categories: emotional, behavioral and cognitive engagement. *Emotional engagement* is defined by the affective reactions of students to any given learning situation, including for example interest or boredom (Fredricks et al, 2004). Interest can be personal or situational. Personal interest, while necessarily directed towards a specific situation, refers to an individual's preference to certain topics, and their willingness to take on and persist in difficult tasks relating to areas of interest. Situational interest relates to the activity itself, such as the storyline and design of an educational app.

Behavioral engagement is defined differently according to contexts. Here, I define it as being involved in learning processes and tasks, in line with the second definition provided

by Fredricks and colleagues (2004). In the current context, it relates to children's physical interaction with apps or other educational software.

Finally, *cognitive engagement* can be viewed from two related, yet sometimes conflicting, perspectives. The first views cognitive engagement as psychological investment in learning and mastering the knowledge that a learning context is intended to promote, while the other focuses on self-regulation and the use of metacognitive strategies when undertaking tasks (Fredricks et al., 2004). While students may be both highly strategic and highly invested in their work, they could also employ avoidance strategies as a means to finish tasks without having to engage with the content.

2.4.1 Interaction Patterns

In this section, I will first review *gaming the system* and *wheel-spinning*, two patterns of interaction that have received much attention in recent research, and how these relate to engagement. I then give an overview of repeated mistakes, and how this behavioral pattern may relate to gaming the system, wheel-spinning, and engagement.

Gaming the system refers to behavior where the goal is to complete tasks without having to engage with the task content (Baker et al., 2004). This manifests as hint abuse or systematic guessing (Baker et al., 2009). Hint abuse can be an issue in educational software that provides help functions giving a series of hint, of which the last one provides the solution to the task. Students may then click through all the hints in order to get to the answer with a minimum of effort. Systematic guessing, on the other hand, involves systematically attempting different answers until hitting the correct one. When gaming the system, students may be behaviorally engaged, i.e. they are interacting with the app. However, they are not likely to be emotionally engaged, and cognitive engagement is reduced to the application of avoidance strategies. Gaming the system is related to boredom and confusion (e.g., Baker et al., 2010; Rodrigo et al., 2007) and is associated with poorer short and long term learning gains (Baker et al., 2004; Pardos et al., 2013).

Wheel-spinning, unlike gaming the system, is an interaction pattern where students are fully engaged with tasks and content. Yet, they do not achieve mastery of the content even after attempting many tasks targeting the same skill (Beck & Gong, 2013). This could be due to misconceptions or that children simply do not understand the tasks they are presented with. When students are wheel-spinning, presenting more tasks is unlikely to facilitate learning without some form of intervention, e.g. scaffolding or presenting easier content. In a study of wheel-spinning and productive persistence, Owen et al. (2019) found that wheel-spinning was associated with lower levels of motor skills and prior knowledge. Thus, age or developmental

stage may affect wheel-spinning. Beck and Gong (2013) found a relation between wheel-spinning and gaming the system. Furthermore, this relation may be causal in nature, with wheel-spinning leading to gaming the system (Beck and Rodrigo, 2014).

It is clear that wheel-spinning may lead children to repeat mistakes as they attempt to solve tasks on the basis of misconceptions or without understanding the purpose of what they are doing. Systematic guessing could also lead to an increase in repeated mistakes when response options are shuffled after making mistakes. Thus, both gaming the system and wheel-spinning could be potential explanations of why children repeat mistakes. Here, *repeated mistakes* are defined in line with Articles 3 and 4, as making the same mistake more than once within a task. The *propensity* to repeat mistakes is conceptualized as the likelihood that a child will make the same mistake more than once. This is an unobserved (latent) variable, indicated by the average number of repeated mistakes across sessions. The actual count of repeated mistakes in any given task is an observable expression of propensity combined with context (task and session characteristics). As pointed out in Articles 3 and 4, repeated mistakes could pose a dual threat to learning, depending on its underlying mechanisms. If seen as an indication of disengagement from the content, as in gaming the system, repeated mistakes can hinder the accumulation of information from the input, thus disrupting the process of implicit learning. If, on the other hand, the children are engaged, in line with wheel-spinning, the increased exposure to incorrect answers could make the incorrect information the most salient feature of the input, causing children to learn incorrect patterns. Thus, whether repeated mistakes are related to gaming or spinning has great implications for how we interpret the behavior and its relation to learning, and to child and task covariates, as well as the types of support children may need in order to break the patterns.

2.4.2 Feedback

Feedback varies in form, specificity and complexity (e.g., Nikolayev et al., 2021). Verification and correction are simple forms of feedback, while more complex forms includes elaboration and scaffolding (Nicolayev et al., 2021; Tärning, 2018). Non-specific forms of feedback simply indicates whether an answer is correct or incorrect, whereas specific forms for example supplies the correct answer (simple specific feedback) or an explanation of why an answer is correct or incorrect (complex specific feedback, e.g. Callaghan & Reich, 2018; Nikolayev et al., 2021). Non-specific feedback such as verification is by far the most common form (Nikolayev et al., 2021). Relating to child-app interaction, Tärning (2018) notes that verification feedback can lead to trial-and-error strategies and increased propensity to game

the system. The effect of feedback, however, is dependent on app design. Specifically, trial-and-error in the context of verification feedback can be separated into three categories: low-cost, risky, and time-consuming (Tärning, 2018). Low-cost trial-and-error may promote gaming the system, since it involves little time and effort. Risky and time-consuming trial-and-error, however, could for example lead children to lose points or spend very large amounts of time, making these less rewarding strategies.

Chapter 3: Methods and Methodological Considerations

In this chapter, I first present the app and the child, task and session samples used in the research papers. I then describe measures and validity, before discussing latent variable models, and ethical considerations.

4.1 Design and samples

4.1.1 Kaptein Morf

The app we developed in the VLC project is called “Kaptein Morf og Stjernestøvet” (Captain Morph and the Stardust). The app builds on principles of morphology and implicit learning with high variability in non-target elements to support generalization. The app and the randomized controlled trial are described in great detail in article 1 and article 3, so rather than reiterate all the information here I will focus on some key features that are especially important for the interpretation of the findings in this thesis.

The app contains 40 sessions, arranged to be played at a rate of one session per day, Monday through Friday, for the eight weeks of the intervention. The first four sessions each week introduces new content, whereas every fifth session contains tasks from the previous four sessions to promote consolidation of the content. The first two sessions target inflections, and are meant to introduce the children to the app and the format of the tasks. The content of these two sessions is relatively easy compared to the other sessions. Aside from the consolidation sessions, the remaining ones target 26 different derivational affixes, three compounding processes, as well as one session covering words with multiple derivational affixes. Table S1 in the supplementary material for Article 1 provides an overview of the sessions. We selected derivations based on four factors: 1) frequency of use, 2) number of words containing the derivation, 3) the utility for school children, and 4) fourth graders’ knowledge of the derivations. See supplementary material A of Article 1 for details on the selection procedure.

There are twelve different task types in the app. Each session begins with two “warm up” tasks (type 1), and end with a word generation task (type 12, e.g. “How many words ending in -ist can you think of?”). Except for the word generation task, where children are asked to write their answers, the other task types involves selecting images, dragging-and-dropping images, morphemes or words, or drawing arrows between elements (see supplementary material A from Article 1 for examples and details concerning all task types). Aside from the warm up and word generation tasks, all tasks in a session are presented in

random order. This was meant to minimize “cheating” by looking at someone else’s answer, and also enabled the examination of potential effects of task position. To discourage systematic guessing, the response options are reshuffled after an incorrect answer in most task types.

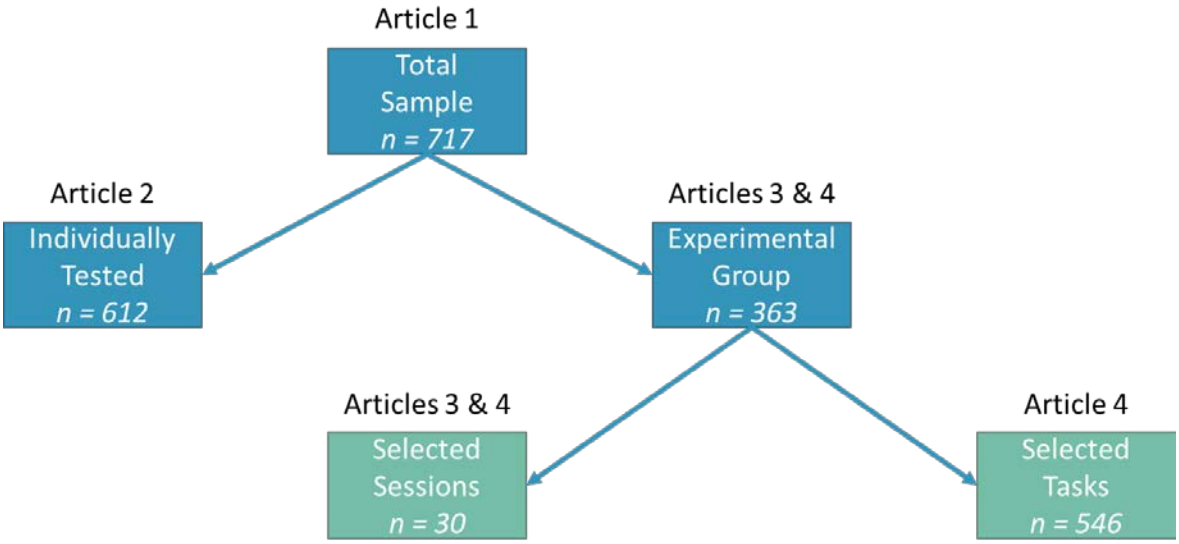
Positive feedback is given in the form stars provided for correct answers, along with the correct answer being shown on the screen. Negative feedback is provided in the form of incorrect answers disappearing and being reshuffled into the remaining response options. Since the app builds on implicit learning with a minimum of explicit instructions, we did not provide any elaborative feedback.

4.1.2 Child, Task and Session Samples

All four articles of the thesis analyzed data from the randomized controlled trial conducted in the Vocabulary Learning Challenge project. Due to the differing aims of the articles, we used different subsets of the data in the papers. Figure 2 shows the distribution of the subsamples in the different articles in relation to the total sample.

Figure 2

Relations between samples in the different articles



Note. Blue boxes describe the samples of children, green boxes describe the samples of app sessions and tasks.

Article 1 is based on the full sample, including all recruited participants from the 12 participating schools. We conducted intention-to-treat analyses, so children were kept in the sample even if they did not complete the intervention. In Article 2, the aim was to do in-depth

analyses of the morphological tests. Due to the late addition of the last school in our sample, we did not have time or resources to administer individual tests to the 105 participating children from this school. Since these children did not answer the majority of the morphological tests, we excluded them from the analyses. Articles 3 and 4 focus on child-app interaction. We did not have access to process data from the math app used by the control group. Therefore, we used the data from the experimental group. Since the first two sessions and the eight consolidation sessions differed in content, these were excluded from our analyses in Articles 3 and 4. Additionally, not all task types allowed for repeated mistakes or provided the necessary details in the process data log files for us to identify repeated mistakes, resulting in a total of 546 tasks available for analyses in Article 4. For more details, see the individual research papers.

4.2 Measures and Validity

Table 1

Overview of Performance Variables

Performance variables	In Article(s)
Morphological Awareness	
Word Analogy	2
Morphological Analysis	
Receptive	1, 2, 3, 4
Productive	1, 2
Morphological Decoding	
Word Reading	1, 2
Spelling	1, 2
WISC-IV Vocabulary	1, 2
Repeated Mistakes	3, 4

Before the intervention, we collected background data using parental questionnaires, as well as cognitive measures administered to the children. These are described in detail in the research papers, particularly in Article 1. The main performance measures in the VLC project were the Vocabulary subtest from WISC-IV (Wechsler, 2009), along with four researcher-developed measures of morphological knowledge: receptive and productive word knowledge (measuring morphological analysis), and word reading efficiency and spelling (measuring morphological decoding). These four tests all focus on morphologically complex words. The

five performance measures were administered at pre-test immediately before the intervention, at post-test directly after the intervention and again at a follow-up test six months after the intervention. We also administered a word analogy test, a measure of morphological awareness, at the follow-up six months after the intervention. In addition to these measures, Article 3 and Article 4 use process data from the app to identify repeated mistakes. Table 1 gives an overview of the performance variables and indicates where more details can be found. In the remainder of this section, I will focus on the validity of the use and interpretation of the scores from the WISC-IV Vocabulary, the morphological tests, and the measures of repeated mistakes from the process data.

The WISC-IV Vocabulary test is a part of the larger WISC-IV battery of tests constructed to measure children's general abilities (Wechsler, 2009). The test measures children's ability to explain the meanings of words. Thus, scores from the test can be seen as indicators of productive word knowledge (vocabulary). To ensure correct administration of the test, all test administrators followed the official manual. A potential validity threat is that this is a single test measuring productive skills, i.e. the ability to explain word meanings. Keeping this in mind, WISC-IV Vocabulary is a well-tested measure of *productive* word knowledge.

The morphological tests were all developed by researchers within the VLC project. While we did pilot the tests, they are not normed, and the RCT was the first large-scale use of these measures. Content-wise, they all focus on morphologically complex words, with a special focus on derivations. Test items were developed with content experts in the project team. Furthermore, they were largely modelled on existing tests in English (see Article 1 for details). The tests, while focusing on morphologically complex words, do not measure morphological knowledge in isolation. For example, the morphological word reading efficiency test likely taps into phonological awareness as well as morphological knowledge. Furthermore, in line with Levesque et al. (2021), we consider morphological knowledge a multidimensional construct. Hence, a single test does not capture the entirety of the construct. By administering tests tapping into different parts of the construct, we get a more complete picture of morphological knowledge. Since the tests capture construct-irrelevant variance (for example due to phonological awareness) in addition to variance relating to morphological knowledge, we need to separate the relevant parts from the irrelevant parts. By using confirmatory factor analysis (CFA) models, we could extract the common variance in items within and between tests. Article 1 used a higher-order model with word reading and spelling as indicators of form-based knowledge (analogous to morphological decoding), and

productive and receptive word knowledge as indicators of meaning-based knowledge, or morphological analysis. Article 2, comparing different models, underlines the importance of model selection for subsequent interpretation. In Articles 3 and 4, we used a single measure of morphological knowledge in the test of receptive word knowledge. In line with using a single measure, we considered the test an indicator of receptive knowledge of morphologically complex words in these papers, rather than as a proxy for general morphological knowledge.

Finally, repeated mistakes is a largely unexamined construct in the field of app-based learning. Thus, extra care must be taken in interpreting the counts of repeated mistakes. First off, there is a difference between the *propensity* to repeat mistakes, as conceptualized in Article 3, and the *count* of repeated mistakes in Article 4. The propensity measure is subject to CFA modeling, extracting the common variance from children's average count of repeated mistakes in each session. We considered this a "baseline" likelihood of repeating mistakes. The observed counts at the task level are expressions of a combination of a child's propensity to repeat mistakes and the context, for example task format and content. While counting mistakes is relatively straightforward, interpretation requires care. Since this is a "new" construct, we do not yet know what underlying mechanisms it represents, and whether there are several different mechanisms potentially leading to similar outcomes. While we discuss some potential mechanisms in Articles 3 and 4, more research is needed to uncover *why* children repeat mistakes.

4.3 Latent Variable Models[3]

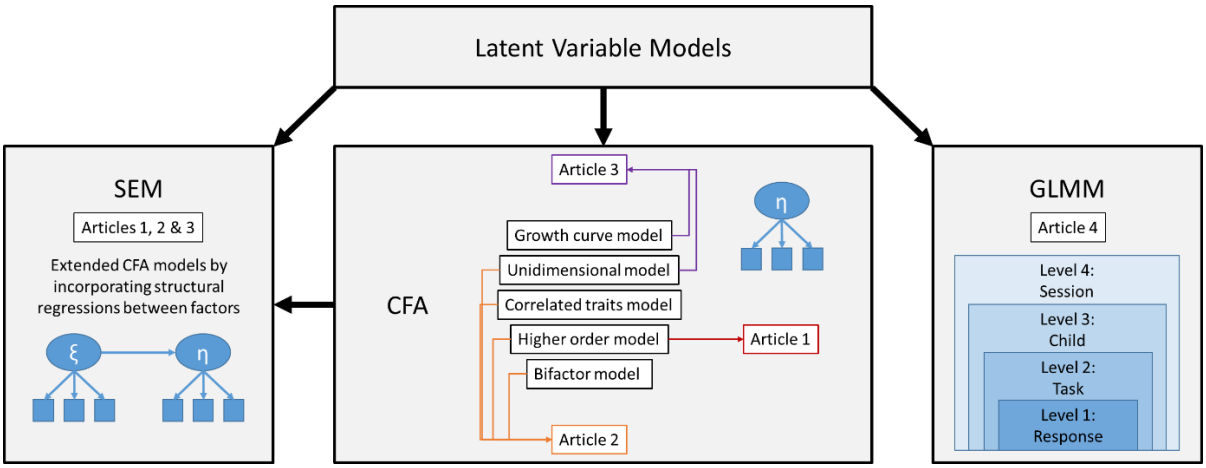
The articles in this thesis all rely on latent variable models. In Articles 1-3, we used confirmatory factor analysis (CFA) and structural equation modeling (SEM) frameworks in the analyses. These are common modeling frameworks in educational and psychological research, where the variables of interest are often impossible to measure directly. Examples are intelligence, personality traits, mathematical abilities and reading comprehension. To examine such constructs, we use observable variables as indicators of the underlying construct we wish to measure. In Article 4, we used a generalized linear mixed model (GLMM). This class of models combines generalized linear models for non-normal data, e.g. binary or count data, with mixed models that incorporate random effects. These random effects are not observed and hence they are in essence latent variables that we derive from the model. The

[3] This section is partially based on my course paper (unpublished) written for UV9002 – Philosophy of Science.

details of each specific model is outlined in the articles, and will not be reiterated here. Rather, in this section, I will focus on the theoretical background and rationale of the models.

Figure 3

Overview of the latent variable modeling framework and the statistical models applied in the research papers



Note. CFA = confirmatory factor analysis, SEM = structural equation modeling, GLMM = generalized linear mixed models.

In addition to providing a framework for modeling unobservable variables, latent variable models can help separate different sources of variance in the observed measures. Both morphological knowledge and repeated mistakes serve as a good examples of this. Regarding repeated mistakes, the observed count in any given task is a result of a specific child solving a specific task in a specific session. Thus, it does not depend only upon the child's propensity to repeat mistakes, but also on characteristics of the task and session. One way to model such complex clustering structures in data is through a mixed model framework. Mixed models, also known as hierarchical models or multilevel models, allow us to separate variance relating to different "levels" of the data, such as child, task and session. We are not likely able to provide all the necessary fixed effects to account for the different sources of variance, so we instead use random effects relating to unexplained variance at different levels.

Returning to the example of morphological knowledge, it is clear that while it is possible to measure someone's knowledge of different morphemes, covering all morphemes and combinations thereof is not possible. Hence, we need to select a subset of morphemes and combinations and ensure that these capture the relevant variance in the construct. Additionally, we need to separate out the variance that is not related to morphology.

Morphological knowledge is measured using tasks that vary in many ways. Task types include decomposition, definition, derivation, lexical decision, morphological relation judgement, naming, plausibility judgment, spelling, prefix/suffix choice and word analogy (Berthiaume et al., 2018). Tasks can employ real words, pseudo-words and non-words, and they also vary in the amount of linguistic context that is provided. Unless we assume that all tasks measure a common underlying ability, it is difficult to defend any comparisons between tasks. The underlying ability is seen as a cause of variation in observed test scores, but even with such an assumption, it is clear that different tasks require different skills in addition to morphological knowledge. This can lead to a form of holist underdetermination (e.g., Stanford, 2017) of the theory of morphological knowledge, as we cannot test hypotheses about morphological knowledge independent of hypotheses and assumptions regarding other language skills and the relationships among these. Our interpretation of unanticipated results, such as the failure to find a predicted relationship, is underdetermined because we cannot know which specific part of our hypotheses and assumptions need revision (Stanford, 2017). How then, can we check whether we are measuring morphological knowledge rather than other skills such as phonological awareness or general vocabulary? First, we need to use several different measures to capture the construct of interest across different contexts. Second, we need to capture the common variance of these measures in order to separate out the construct-irrelevant information related to other skills. Latent variable modeling frameworks such as CFA can help us separate what is relevant from what is not, and to evaluate the ability of different items to describe the phenomenon of interest.

As discussed in Article 2, the interpretation of the resulting construct(s) is substantially affected by the choice of model. Different models that fit the data similarly well may lead to entirely different explanations of the phenomena we wish to study. This can be seen as a form of contrastive underdetermination. Different models outline different theories that are, if not empirically equivalent, at least equally well supported by the current evidence (Stanford, 2017).

How then should we interpret latent variables, and what should guide us in model selection? Borsboom et al. (2003) argue that latent variable modeling necessitates realism. The models used in this thesis are reflective, meaning the latent variables cause variance in the observed variables. In CFA models, this entails that different tests can be seen as measuring the same construct(s), since the underlying ability causes the responses to test items. If we abandon the realist perspective, the latent variable would have to be caused by the observed variable (a formative model). This would entail that every single test measures a

separate construct (Borsboom et al., 2003). The fact that we consider morphological knowledge to be a real ability does not, however, help us determine the structure of that entity. To meet the challenge of model selection, we can take a pragmatic approach, focusing on the purpose and value of a model (Winther, 2021). Returning to the morphological example, we need to consider what different models can tell us about morphological knowledge and its relation to other areas of language and literacy. Furthermore, to fully understand morphological knowledge, we must unravel what is construct-relevant and construct-irrelevant. A model of morphological knowledge is sufficiently detailed if, and only if, it captures the common variance associated with morphological knowledge while explicitly separating out the specific variance of other language skills involved.

The underdetermination of morphological knowledge stems, at least to a large extent, from the way we measure the construct. Tests cannot themselves disentangle morphological knowledge from other areas of language and literacy. Differences in dimensionality across studies could stem from different tests on one hand measuring too little of morphological knowledge and on the other hand measuring too much of other skills. As different studies use different tests, the degree to which they measure a common construct is bound to vary; it depends on which parts of morphological knowledge they measure, and the degree to which the tests measure other skills in addition to morphological knowledge. If the tests used are similar in other regards than morphological knowledge, such as general vocabulary, this might even cloud the common construct(s) that are extracted. While there is still much to be discovered about morphological knowledge, the field seems to be moving towards a multidimensional view. Thus, in order to capture the complete picture, and to investigate the constituent parts measured by tests of morphological knowledge, we need to use several measures, and modeling frameworks that allow for separation of different sources of variance.

To meet this challenge, we used a higher order model of morphological knowledge in Article 1, where morphological analysis is measured by both receptive and productive tests, and morphological decoding is measured with by word reading and spelling. Thus the different indicators include different “support” skills. Article 2 provides further argumentation for such a framework, while in Articles 3 and 4 we acknowledge the issue by referring to receptive knowledge of morphological words rather than general morphological knowledge when using the single test.

4.4 Ethical considerations[4]

Several ethical considerations were made during the planning and execution of this project, relating to the collection, storage and use of sensitive data, as well as specific considerations relating consent and participation when conducting research with children. The proposed collection, storage and use of data were evaluated by the Norwegian Agency for Shared Services in Education and Research (Sikt, previously NSD). The sensitive nature of parts of the data required specific consideration. The data included audio recordings and information such as whether students received special education, which can be readily identifiable. To ensure data security, we used the data storage platform provided by Services for sensitive data (TSD), and personal information was de-identified.

Voluntary informed consent is a fundamental requirement in research on human participants and when involving children, specific ethical considerations are required (Backe-Hansen, 2009; The Norwegian National Committees for Research Ethics, 2016). In addition to consent, important concerns include student well-being on an individual level, and interference with the everyday organization of teaching on a group level.

Beauchamp and Childress (2001) present four clusters of ethical principles: *respect for autonomy, non-maleficence, benevolence,* and *justice.* Although the original context for these is biomedicine, the principles are similarly applicable in other areas when conducting research with human participants. This is perhaps especially true for research involving vulnerable groups such as children. *Respect for autonomy* refers to respecting participants' ability to make their own decisions. According to Norwegian law, parents or legal guardians (hereafter referred to only as parents) are required to consent on behalf of children under the age of 15. However, the researcher is still responsible for ensuring that the child also consents. In the VLC project, we provided parents with written information about the project and the consequences of participating and we required a signed consent form before a child was included in the study. Additionally, the children were given age-appropriate information about the project, and about their right to withdraw if they did not want to participate. For a few children, a different issue arose when they wanted to participate, but their parents did not consent. To counteract feelings of exclusion and unjust treatment, these children were offered to work with the math app, as well as to take part in group assessments if they wanted to, without having any information recorded.

[4] This section is based on my course paper (unpublished) written for UV9010 – Research Ethics.

The principle of *non-maleficence* emphasizes that you should avoid causing harm to participants. When conducting intervention research, one concern is that the intervention does not have the intended effect. For untested interventions, this may be difficult to foresee and counter. To meet this challenge, we piloted a short version of the Kaptein Morf app, to establish the potential of the proposed intervention. Vektor, the app used by the control group had already been subject to previous research by the developers (e.g., Nemmi et al., 2016).

Another concern is whether some students will experience repeated failure. The app is not adaptive, and some sessions are relatively difficult. If a student keeps failing at a task, this may lead to frustration, dejection and loss of motivation. It is important to support students that struggle and help them understand the task, to avoid students getting stuck and giving up. In the long run, this process may be helped by providing automatic notifications to teachers about students who struggle, as well as by automated feedback to the children, and adaptive task selection within the app.

For schools, one major concern with participation in research is increasing teacher workload. At the current state, the app is self-contained and should not put much strain on the teachers. However, it is important that the teachers receive necessary information and support. Another concern is that it takes time away from ordinary teaching activities. Even for an intervention with relatively short daily sessions this could be an issue for teachers.

Beneficence relates to the positive gains for research participants. The aim for any intervention is to change something for the better. It is therefore natural to consider what students and schools stand to gain from participation. The main benefit for children in the experimental group was increased word knowledge, along with strategies to help them learn new words. For the control group, the benefits were related to skills and strategies in mathematics. These are intrinsic gains, and not necessarily ones the students themselves will notice. In addition, the students received extrinsic gains, for example in the form of diplomas for completing different stages of the intervention, which served as an external source of motivation.

For the schools and teachers, participation gave the opportunity to try a research-based educational application in their teaching. Through participation in the project and related seminars, teachers also gained knowledge about morphology and how it can be incorporated in teaching to support vocabulary development.

Finally, the principle of *justice* focuses on fair distribution of risks and benefits connected to research. Certain groups or individuals should not suffer greater risks, nor gain greater benefits, than others. For participating children, the issue of justice concerned equity.

One concern was whether there equal opportunities for students to participate. The project aimed to include all children in the second grade at the schools that were invited, and there were no exclusion criteria for participation, meaning, for example, that the sample included language minority children and children with special educational needs. However, an issue regarding opportunities to participate was that the app is not adaptive. This made it difficult for children with poor language comprehension to participate. This was addressed by teachers providing individual support, however, a few children (approximately 1% of the sample) still had to withdraw from the intervention due to the relative difficulty of the tasks.

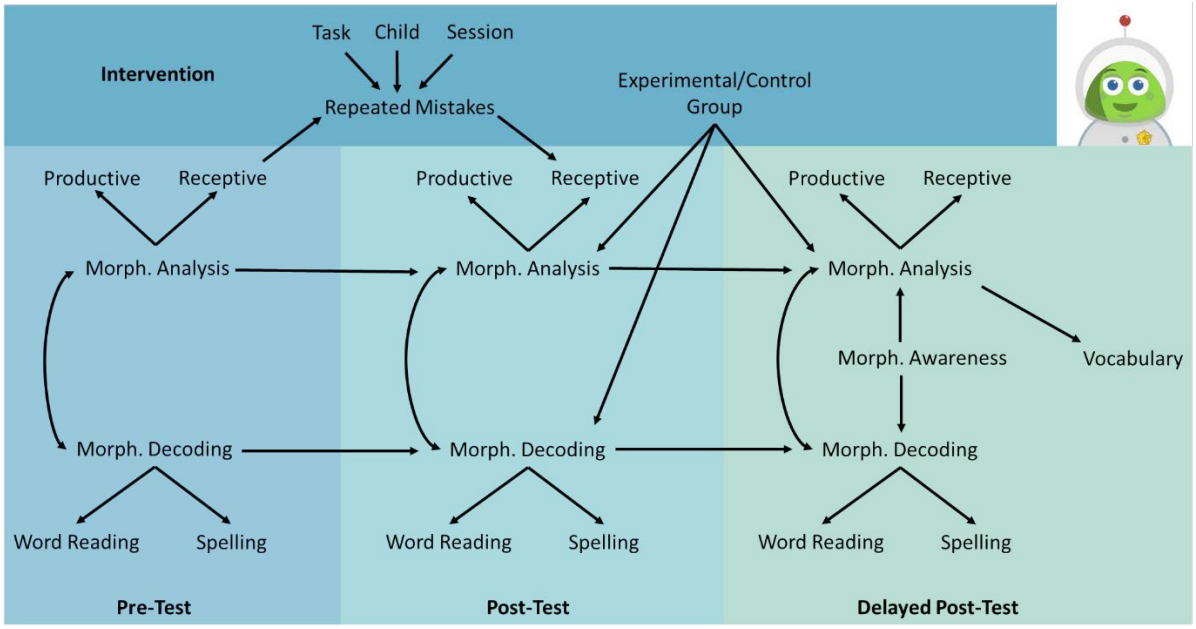
Another consideration to make is whether all students receive equal gains from participating. The extrinsic gains were the same across students (with rewards also given to the class as a whole, so as not to exclude non-participating students). The intrinsic gains may be a somewhat different story. The Kaptein Morf app did result in learning gains in the experimental group seen as a whole, and the groups were offered to switch apps after the six month follow-up to allow for the children to also receive the alternate intervention. However, Article 3 shows that children with lower prior knowledge were more likely to repeat mistakes, and that those who repeated more mistakes had lower expected learning gains. Thus, some children may have benefitted less from the app than others. A final consideration when it comes to justice for the students is whether some (groups of) students are more at risk of harm. For example, children with language disorders are probably more at risk of experiencing failure, frustration and dejection. Future iterations of the app should address these issues.

Chapter 4: Article Summaries

In this chapter, I give a brief summary of each of the articles in this thesis, mainly focusing on the key findings in each study. Figure 4 shows a collective overview of the results across the articles and their interrelations.

Figure 4

Summary of Results across Articles



Note. The lower three panels show the key performance variables at each time point, as well as the relations between these. The upper panel shows intervention variables and how they relate to the performance measures.

5.1 Article 1: Effects of App-Based Morphological Training

Torkildsen, J. V. K., Bratlie, S. S., Kristensen, J. K., Gustafsson, J.-E., Lyster, S.-A. H., Snow, C., Hulme, C., Mononen, R.-M., Næss, K.-A. B., López-Pedersen, A., Wie, O. B., & Hagtvet, B. (2022). App-based morphological training produces lasting effects on word knowledge in primary school children: A randomized controlled trial. *Journal of Educational Psychology, 114*(4), 833–854. https://doi.org/10.1037/edu00 00688

This article reports on the RCT where we investigated the effects of using the Kaptein Morf app in an eight week intervention. The study included 717 Norwegian second graders

who were tested before and immediately after the intervention, and at a six-month follow up to examine long term effects.

The intervention led to improvements in meaning-based knowledge (morphological analysis) and code-based knowledge (morphological decoding) of words exposed during the app sessions. It also led to generalization to unexposed words containing morphemes encountered in the app. This generalization shows that children acquired knowledge about affixes and compounding patterns in addition to knowledge about specific words. This is likely an effect of the high nontarget (base word) variability, increasing the saliency of affixes and compounding patterns (invariant elements), thus promoting generalization. These findings align with results from previous research indicating that high nontarget variability in language input facilitates the learning of underlying grammatical regularities (Gómez, 2002; Torkildsen et al., 2013).

In addition to the generalization of knowledge to unexposed words, the results indicated transfer from implicit learning to explicit skills. The intervention did not involve any oral language production on the part of the children and the app contained no explicit explanation of the meanings of words or affixes. However, after the eight-week intervention, the children were better at giving explicit explanations of the words' meanings. In fact, raw scores for the different outcome measures suggested that training effects were largest for productive word knowledge (word definitions).

The effects for meaning-based knowledge of exposed and unexposed words and code-based knowledge for exposed words were sustained at the 6-month follow-up in third grade. Additionally, we found a significant indirect effect of meaning-based knowledge of exposed words at the posttest on unexposed words at the follow-up. This suggested that sustained generalization effects depended on learning gains in exposed words. While there were sustained effects on code-based knowledge of exposed words, we did not find sustained effects on code-based knowledge of unexposed words.

We found no significant far-transfer effects to a measure of general productive vocabulary that did not focus specifically on morphologically complex words. The lack of effects on the measure of general vocabulary were likely due to the fact that the program did not train knowledge of base words but focused on affixes and compounding patterns.

The findings in Article 1 indicate that the Kaptein Morf app is similarly beneficial to children who differ widely in language skills and general ability, with children with lower than average general abilities showing slightly larger gains for meaning-based skills for unexposed words at the posttest than children with higher initial general abilities did.

However, the findings in Articles 3 and 4 nuance this view. Finally, there was evidence that shorter app completion time was associated with better outcomes for code-based skills (word reading fluency and spelling), even when the corresponding pretest scores and cognitive and linguistic background variables were taken into consideration.

5.2 Article 2: Dimensionality of Morphological Knowledge

Kristensen, J. K., Andersson, B., Bratlie, S. S., & Torkildsen, J. V. K. (2023). Dimensionality of Morphological Knowledge—Evidence from Norwegian Third Graders. *Reading Research Quarterly*, *58*(3), 406-424. https://doi.org/10.1002/rrq.497

In this article, we compared several CFA models to examine the dimensionality of morphological knowledge, using data from the 612 children in the VLC project who participated in individual testing. Furthermore, we regressed productive vocabulary on the morphological factors in three competing models, a five-factor (correlated traits) model, a higher order model and a bifactor model, to further examine how informative each model was.

Our findings indicate that morphological knowledge is a multidimensional construct. Furthermore, the five-factor model was less informative than the other two models. The results provide evidence in support of the Morphological Pathways Framework (Levesque et al., 2021), as the higher order model with factors representing morphological awareness, morphological analysis and morphological decoding fit the data well. Furthermore, the higher order SEM supported the relations to vocabulary suggested by the Morphological Pathways Framework. We found a significant direct relation between vocabulary and morphological analysis, and an indirect relation with morphological awareness, through morphological analysis. We did not find a significant association between vocabulary and morphological decoding.

The bifactor model also fit the data well, and can provide valuable insights about children's overall (general) morphological knowledge, as well as the additional skills the tests tap into, such as phonological awareness. This is perhaps especially valuable in test development, to understand the different skills a test measures. However, since this model assumes that the factors are uncorrelated, it is less useful if we wish to examine the relations between different morphological skills, and their relations to other language skills. Thus, for research purposes, we recommend a higher order model in line with the Morphological Pathways Framework.

5.3 Article 3: Repeated Mistakes in App-Based Language Learning

Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2024). Repeated Mistakes in App-Based Language Learning: Persistence and Relation to Learning Gains. *Computers & Education*. https://doi.org/10.1016/j.compedu.2023.104966

In Article 3, we investigated children's propensity to repeat mistakes while working with Kaptein Morf. Using test and process data from the experimental group, we examined whether the propensity changed across sessions or remained stable, as well as the relations between prior knowledge, repeated mistakes and learning gains. To examine potential changes over time in the propensity to repeat mistakes, we estimated a growth curve model with one factor representing baseline propensity (intercept) and another representing changes (slope). As we did not have a prior hypothesis about the shape of the slope, we allowed the factor loadings on the slope factor to vary freely. Inspecting the results, however, we found no significant loadings on the slope factor, indicating that the baseline propensity explained all the common variance across the sessions. Along with results indicating that a unidimensional model fit well, this indicated that the propensity to repeat mistakes is best represented as a trait that remains stable over time.

We proceeded to estimate a structural model including pre-test and post-test scores on the test of receptive knowledge of morphologically complex words. Post-test scores were regressed on pre-test scores and on the propensity to repeat mistakes. The propensity factor was also regressed on the pre-test scores. We found a negative association between pre-test scores and repeated mistakes, indicating that children with less prior knowledge tend to repeat more mistakes. We also found a negative association between repeated mistakes and post-test scores, indicating that children who repeat more mistakes are likely to have lower learning gains. Importantly, our results indicated that repeated mistakes mediate the relation between prior knowledge and learning gains. Children with a high propensity to repeat mistakes have lower expected gains than low propensity children with the same pre-test score.

5.4 Article 4: Task and child covariates of repeated mistakes

Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2023). *Who repeats mistakes and when? Task and child covariates of repeated mistakes in app-based learning* [Manuscript in preparation]. Centre for Educational Measurement, University of Oslo.

In this article, we examined task and child covariates of repeated mistakes with the aim to unravel some of the underlying mechanisms related to mistake repetition. Using data from the experimental group, we estimated a four-level GLMM where the count of repeated mistakes was modelled with a negative binomial distribution. Level 1 represented the individual responses. Level 2 represented individual tasks, with fixed effects of task type and task position. At level 3, we modelled the child characteristics with gender, language background, tablet use, receptive morphological knowledge and non-verbal ability as fixed effects. Finally, at level 4, we included session number as a random effect, without any fixed effects. The intercept at level 1 was allowed to vary as a function of random effects at the task, child and session levels.

The random effects indicated that the number of repeated mistakes vary across tasks and sessions, as well as between the children. Furthermore, they suggest that there is significant unexplained variance at each level, not captured by the fixed effects included in our model. Thus, there are unobserved mechanisms influencing the number of repeated mistakes a child makes.

At the task level, we found significant relations to both task type and position. The results indicated that tasks where response options are reshuffled following an incorrect answer, as well as tasks with more complex combinations of response options and targets, are associated with higher numbers of repeated mistakes. For task position, we found a negative association to repeated mistakes, indicating that later tasks elicit fewer repeated mistakes.

At the child level, we found that repeated mistakes were negatively related to receptive morphological word knowledge and non-verbal ability. We did not find significant associations to gender, language background or tablet use.

The findings regarding task types are in line with research on gaming the system, and the relations to prior knowledge and non-verbal ability are in line with both gaming the system and wheel-spinning. However, the negative association to task position is not in line with gaming the system, while it does make sense in terms of wheel-spinning. Thus, we believe that interpreting repeated mistakes in line with wheel-spinning might be more viable than an interpretation in line with gaming the system. We cannot, however, exclude the possibility that repeated mistakes could relate to both, and perhaps to other underlying mechanisms, and that there are differences among children repeating mistakes.

Chapter 5: Discussion and contributions

The work presented in this thesis carries important implications for the development of interventions in support of language learning, and particularly for app-based vocabulary interventions for primary school children. Furthermore, it highlights not only the importance of morphology, but the implications of how morphology is conceptualized, taught and measured. To structure the discussion, this chapter is divided into three main sections according to the topics addressed in the thesis: 1) the role of morphological knowledge and its conceptualization in language learning and assessment, 2) the role of implicit learning in morphological training, and 3) the role of response behavior in app-based implicit learning. Finally, I summarize the discussion and provide recommendations for research and education in the last section of this chapter.

6.1. The Role of Morphological Knowledge in Language Learning

Several areas of language, such as vocabulary, reading comprehension and text generation, are unconstrained in terms of the vastness of their contents. While these are skills that children typically do acquire over time, it is difficult to target them directly in interventions, since an intervention is time-limited and can necessarily only cover a small subset of the content. Additionally, interventions focusing on specific subsets are unlikely to provide transfer effects to other content. For example, in vocabulary interventions focusing on specific individual words, children's knowledge gains are unlikely to generalize to other untrained words. Furthermore, if an intervention targets general vocabulary, chances are that the children will eventually learn the words even without the intervention. This contributes to what is known as fade-out effects, where the control group catches up to the experimental group over time, levelling out the immediate effects of the intervention. Bailey et al. (2017) argue that effective interventions need to target trifecta skills, skills that are fundamental, malleable and would not otherwise be obtained. Furthermore, the emphasis should be on "generic" skills rather than specific knowledge (Bailey et al., 2020). Morphology represents a fundamental and malleable set of skills with a constrained content area. Furthermore, the lack of mention of morphology in the Norwegian curriculum means that it is unlikely that teachers put much focus on morphological knowledge in their lessons. Thus, it is unlikely that children will obtain these skills in a desirable degree, at least in primary and middle school. This can explain the robust and lasting effects found in Article 1.

The Kaptein Morf app focuses mainly on morphological analysis, or meaning-based skills, and specifically on receptive knowledge. Yet, Article 1 provides evidence of significant gains in both morphological analysis and morphological decoding (form-based skills) as measured at the immediate post-test. Thus, it is clear that morphological training focusing on one skillset can generalize to different skillsets. Furthermore, the strongest effect was on productive knowledge, indicating that receptive understanding of morphemes can increase the ability to explain them. At the follow-up, six months after the intervention, we found indirect effects on meaning-based knowledge of unexposed words, i.e. words containing affixes used in the app, in combination with base words not encountered in the intervention. This finding indicates that children can use their knowledge of affixes to understand previously unknown morphologically complex words. Again, this points to morphological skills as generalizable, with transfer effects to new contexts.

There was, however, a fade out of the effects on word reading efficiency and spelling, indicating that the generalization to morphological decoding is less robust. However, this could be due to word reading efficiency and spelling also relying on phonological awareness, phonological decoding, sight word recognition etc., which receive intensive classroom instruction in second and third grade, and were likely still developing in our sample. A possible implication is that the app could benefit from including tasks that focus more on morphological decoding, thus developing a stronger tie between decoding and analysis.

The Morphological Pathways Framework suggest that morphological knowledge consists of three separate but related dimensions that relate differentially to other language skills such as word reading, spelling and vocabulary (Levesque et al., 2021). While we did not find any evidence of the app producing gains in general vocabulary, this finding may be nuanced, both in light of other relations in the model presented in Article 1, and through the findings of Article 2. In Article 1, all performance measures are regressed on group membership to examine the effects of the intervention. In the case of meaning-based knowledge of unexposed words at the follow-up, there is no significant direct effect. However, there is a significant indirect effect through exposed words at the immediate post-test. This indicates that those who had higher gains in exposed words immediately after the intervention were more likely to generalize this knowledge to unexposed words over the following six months. In the case of general vocabulary, we only included direct effects in the model. The results of Paper 2, however, provide evidence of a strong relationship between morphological analysis and vocabulary. Thus, it seems likely that increasing morphological analysis skills could be beneficial for general vocabulary as well. In light of this, we should

have investigated whether a potential indirect effect on vocabulary at the follow-up exists, due to generalization of the meaning-based knowledge of exposed words at the immediate post-test. Another possible explanation is that the high correlations between vocabulary and morphological analysis, and particularly productive knowledge of morphologically complex words, may have masked a potential effect. In Article 2, the correlated traits model indicates that productive word knowledge is the only morphological factor related to general vocabulary, likely due to the effect of high correlations between morphological factors, especially between receptive and productive knowledge. Similarly, any effect of the intervention on general vocabulary could have been masked by the correlation between meaning-based knowledge and vocabulary.

The measure of general vocabulary did however not focus specifically on morphologically complex words. Rather, most of the earlier items in the test are morphologically basic, consisting only of a base word. Thus, affix knowledge is not relevant in comprehending these words. Any intervention effects would therefore appear in later items, and due to stopping rules, only children with high vocabulary scores would be tested on these items. Either way, Nagy and Anderson (1984) found that the majority of words in printed school English are morphologically complex. Thus, base word knowledge alone does not cover the necessary range of vocabulary to master school texts (although it does constitute an important component).

6.2 The Role of Implicit Learning in Morphological Training

While morphology can support the development of language and literacy, certain types of morphological knowledge can be difficult to teach explicitly to young children, e.g. those that require knowledge about word classes. Since affixes and compounding patterns occur in many different combinations, implicit learning offers another approach. Article 1 shows that implicit learning is indeed effective, and that even short daily sessions can lead to learning that generalizes to new words. Furthermore, the receptive training generalizes to increase children's productive knowledge, or their ability to explain the meaning of words. Thus, implicit learning seems to be a promising approach to support children's development of morphological knowledge.

While our results indicate that implicit learning of morphology can provide positive effects on children's word knowledge, we did not compare it to explicit teaching. Furthermore, I believe that combining implicit learning with explicit instruction where possible would likely have led to even larger gains. Another caveat is pointed out in Articles 3

and 4. The implicit acquisition of morphological knowledge relies on continuous accumulation of information from input with high variability. Thus, if students are not engaged and paying attention to the content of the app, they may not receive the necessary amount of relevant input. On the other hand, if they pay attention while repeating mistakes, the amount of incorrect information could make the wrong content the most salient part of the input. Hence, the positive effects of implicit morphological training rely on how the children interact with the app, which is the focus of the following section.

6.3 The Role of Child-App Interactions in Implicit Language Learning

Process data can provide an important source of information about children's use of educational software and potential sources of individual differences in progression and learning. Certain interaction patterns can have a negative impact on learning, for example gaming the system (Baker et al., 2004; Pardos et al., 2013) and wheel-spinning (Beck & Gong, 2013). In Articles 3 and 4, we chose to examine repeated mistakes. When a child repeats the same mistake a large number of times within a single task, it is clear that they are either not paying attention to what they are doing, or they are for some reason not correcting their answers according to the feedback. The presence of repeated mistakes is of course dependent upon app features. For example, apps providing corrective feedback, i.e. showing the correct answer after a mistake are less likely to elicit repeated mistakes. Similarly, apps where the user proceeds to the next task regardless of the correctness of the answer naturally excludes mistake repetition, unless tasks reappear at a later stage. Thus, the pattern of mistake repetition relates mainly to apps or other software based on some degree of trial and error.

In our context, it is clear that repeated mistakes represent a behavioral pattern that is easy to detect, and that should be addressed. Results from Articles 3 and 4 nuance to the findings in Article 1, illustrating that the app, while overall effective, might not have been equally beneficial for all the children.

In Article 3, we found that children with lower prior knowledge had a higher average propensity to repeat mistakes. The propensity to repeat mistakes was also associated with lower learning gains, and mediated the autoregressive relationship between knowledge at pre-test and post-test. Importantly, children with a higher propensity to repeat mistakes had lower expected post-test scores than low-propensity children with the same pre-test scores.

The lower learning gains may relate to two different sides of implicit learning, referred to in Article 3 as the "dual threat" of repeated mistakes. On the one hand, implicit learning necessitates continuous accumulation of information from the output. Thus, if children

disengage from the content of the app, they are not likely to notice the target patterns. On the other hand, if they are engaged with the content, but still make repeated mistakes, they are subjected to an inordinate amount of incorrect input. This could lead to incorrect combinations becoming salient features, resulting in learning the wrong patterns. Future research should examine whether repeated mistakes made in specific tasks during training sessions are related to the mistakes children make in specific post-test items (e.g., do they make mistakes relating to the same affixes or compounding patterns?).

In addition to prior knowledge, the results in paper 4 showed an association between non-verbal ability and repeated mistakes. This means that children who are already at a disadvantage are more likely to repeat mistakes. Thus, the intervention could contribute to uphold, or in the worst case increase, the gap between high and low achieving children. This highlights the need to identify children with a high propensity to repeat mistakes and provide them with the support necessary to learn from the app. Since children with lower prior knowledge and non-verbal ability are at more risk, we should monitor their progress closely to reduce potential negative effects on their learning.

The interpretation of repeated mistakes depends heavily on the supposed underlying causes of this behavior. While the work presented in this thesis did not delve into these causal analyses, I will discuss two possible perspectives relating to the "dual threat". On one hand, repeated mistakes could represent gaming the system, or more specifically, systematic guessing. Since response options are reshuffled in most task types, such a strategy will fail unless the children pay close attention to which options they have attempted and which options remain untested. Paying close attention to their answers is, however, in direct opposition to the goal of gaming the system, i.e. completing tasks without interacting with tasks (e.g., Baker et al., 2004). Thus, systematic guessing would necessarily lead to repeated mistakes as a consequence of response options changing place. This is congruent with our findings in article 4, where tasks with reshuffling and larger numbers of response options are related to higher levels of mistake repetition. Furthermore, this view is supported by the relation between prior knowledge and repeated mistakes, which is in line with results from research on gaming the system (Baker et al., 2004). If we conceptualize repeated mistakes as an indicator of gaming the system, this means that children who repeat mistakes are disengaged from the app content.

On the other hand, it is also possible that children maintain attention and engagement, and still make repeated mistakes. This could be an indication of wheel-spinning. Wheel-spinning is also associated with lower prior knowledge, as well as with the difficulty of tasks.

Since wheel-spinning involves continued persistence without understanding or learning from tasks, it would also likely lead to repetition of mistakes. In Article 4, we found a negative association between task position in the session and repeated mistakes which could indicate that the mistake repetition is related to wheel-spinning rather that repeated mistakes. The lower counts of repeated mistakes in later tasks could signal that children are confused at first, leading to wheel-spinning, but break out of the pattern in later tasks when the discover "the word of the day", i.e. the target affix or compounding pattern. While more research is needed to support this claim, we argue that it is a viable explanation. This calls for the development of further support structures to help the children who get stuck. Their progress could be scaffolded in several ways, for example by increased feedback (e.g., correctional feedback) or hints within the app, by implementing adaptive task selection, or by allowing multiplayer features where children can collaborate during problem solving. The effect of position is quite small, however, which could be an indication that it does not apply to all children or in all cases. Thus, an important aim for future research is to examine whether repeated mistakes represent different underlying mechanisms, and how different mechanisms may be intervened upon.

6.3 Summary and Recommendations for Practice and Research

While the extent of morphological instruction for Norwegian primary school children is unclear, it is unlikely that teachers put much focus on morphology in their lessons. The work presented in this thesis shows that children could benefit from morphological instruction, and that even short, self-sustained sessions of app-based learning can be beneficial. Thus, a clear recommendation to schools and teachers is to introduce morphology into their curriculum. As for policy-makers, specific morphological learning goals should be included in the national curriculum to make schools and teachers more aware of this important area of language learning.

Additionally, there is a need to develop standardized and normed measures of morphological knowledge in Norwegian. This would contribute to more accessible and comparable information on how morphological knowledge develops, and which areas need strengthening through teaching and intervention programs. In line with Article 2, it seems like a good place to start would be to develop a battery of tests measuring morphological awareness, morphological analysis and morphological decoding. The four tests developed as part of the VLC project can be seen as a first step in this direction.

Regarding the effects of the app, it is likely that the benefits would be even greater with the addition of other activities besides the app. For example, in one classroom (not participating in the RCT), the teacher let the children work individually with the first four sessions of the week. Then, in the fifth session which covers tasks from the previous four days, they solved tasks collectively on a whiteboard. The children acted as teachers, showing their teacher how to solve the different tasks. While we do not have any data from this classroom, the anecdotal evidence at least points to the children being engaged and finding this activity rewarding. Along a similar vein, an ongoing project is currently developing a new multiplayer version of the Kaptein Morf app that facilitates collaborative problem solving (Falck & Torkildsen, 2022). Also, combining the app with more traditional approaches may prove beneficial, for example by discussing the contents of a session in class. Finally, the principles of implicit learning with high variability could be incorporated into short stories, with each story including a target affix or compounding process presented in at least 24 different contexts. Such stories could contribute both to meaning-based knowledge through richer contexts than those presented in the app, and to form-based knowledge by focusing on reading morphologically complex words containing target affixes or compounding patterns. Potential routes to morphological learning should be further examined in future research.

Finally, it is exceedingly clear that we need to monitor how children interact with educational apps. Articles 3 and 4 show that children who are already disadvantaged are at risk of gaining less than their high-achieving peers when working with educational apps, at least apps that are not adaptive. While the focus of this thesis is on repeated mistakes, this seems to be a common theme in research on negative interaction patterns, for example gaming the system and wheel-spinning. Finding ways to intervene and support these children should therefore be an imperative goal for research and development in the field of educational software.

6.4 Limitations

The work presented in this thesis focuses on Kaptein Morf being used in isolation. The large RCT study was conducted to examine whether the app contributes to learning in and of itself. I believe, however, that including the app in a more holistic framework for morphological instruction would be beneficial for the learning process. It is also possible that group sessions or discussions of the app content could help reduce children’s propensity to repeat mistakes.

Furthermore, the app, due in parts to the research design, and in parts to a lack of previous data on the difficulty of different morphemes and task types, is not adaptive. The extensive data collected in the VLC project could, however, facilitate the process of implementing adaptivity in a future iteration of the app.

Regarding tests, the measures of morphological knowledge provided good coverage of the construct, although a second test of morphological awareness would have been desirable. Other areas of language were covered to a lesser extent, and it would have been interesting to include measures of reading comprehension, as well as spelling of base words, to examine relations to morphological factors and potential intervention effects. Additionally, adding another measure of general vocabulary could have contributed to broader coverage of that construct. The test battery we did include was already very comprehensive, though, so it was not possible to include everything we might have wished to add.

The work in Article 2 supports a multidimensional view of morphological knowledge in line with the Morphological Pathways Framework, but more research is needed to solidify claims about the exact structure of morphological knowledge and its associations to other areas of language. The Morphological Pathways Framework is also based largely on English, so in addition to the supporting evidence from Norwegian provided here, research is needed across different languages to examine whether these dimensions are “universal”.

Finally, repeated mistakes represent a new approach to the study of child-app interaction. While I believe the studies presented here make important contributions, they represent only the first small steps towards understanding this behavior and what causes it.

References

Apel, K. (2014). A comprehensive definition of morphological awareness: Implications for assessment. *Topics in Language Disorders, 34*(3), 197-209. https://doi.org/10.1097/TLD.0000000000000019

Apel, K., Henbest, V. S., & Petscher, Y. (2023). Effects of Affix Type and Base Word Transparency on Students' Performance on Different Morphological Awareness Measures. *Journal of Speech, Language, and Hearing Research, 66*(1), 239-256. https://doi.org/10.1044/2022_JSLHR-22-00195

Backe-Hansen, E. (2009). *Barn*. Retrieved from https://www.etikkom.no/FBIB/Temaer/Forskning-pa-bestemte-grupper/Barn/

Bailey, D., Duncan, G.J., Odgers, C.L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness, 10*(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest, 21*(2), 55-97.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students" game the system". In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 383-390). https://doi.org/10.1145/985692.985741

Baker, R. S., De Carvalho, A. M. J. A., Raspat, J., Aleven, V., Corbett, A. T., & Koedinger, K. R. (2009). Educational software features that encourage and discourage “gaming the system”. In *Proceedings of the 14th international conference on artificial intelligence in education* (Vol. 14, pp. 475-482). https://doi.org/10.3233/978-1-60750-028-5-475

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223-241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Beauchamp, T.L., & Childress, J.F. (2001). *Principles of Biomedical Ethics. Fifth Edition.* New York: Oxford University Press, Inc.

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16* (pp. 431-440). Springer Berlin Heidelberg.

Beck, J., Rodrigo, M.M.T. (2014). Understanding Wheel Spinning in the Context of Affective Factors. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.). *Intelligent Tutoring Systems. ITS 2014. Lecture Notes in Computer Science, vol 8474*, (pp. 162-167). Springer, Cham. https://doi.org/10.1007/978-3-319-07221-0_20

Berthiaume, R., Bourcier, A., & Daigle, D. (2018). Morphological Processing Tasks and Measurement Issues. In R. Berthiaume, D. Daigle, & A. Desrochers (Eds.), *Morphological Processing and Literacy Development* (pp. 48-87). Routledge. https://doi.org/10.4324/9781315229140

Boeve, S., Zhou, H., & Bogaerts, L. (2023). *A meta-analysis of 97 studies reveals that statistical learning and language ability are only weakly correlated*. PsyArXiv. https://doi.org/10.31234/osf.io/s8mwv

Borsboom, D., Mellenbergh, G.J., & van Heerden, J. (2003). The Theoretical Status of Latent Variables. *Psychological Review, 110*(2), 203-19.

Bowers, P. N., Kirby, J. R., & Deacon, S. H. (2010). The effects of morphological instruction on literacy skills: A systematic review of the literature. *Review of Educational Research, 80*(2), 144–179. https://doi.org/10.3102/0034654309359353

Callaghan, M. N., & Reich, S. M. (2018). Are educational preschool apps designed to teach? An analysis of the app market. *Learning, Media and Technology, 43*(3), 280–293. https://doi.org/10.1080/17439884.2018.1498355

Carlisle, J. F. (2010). Effects of instruction in morphological awareness on literacy achievement: An integrative review. *Reading Research Quarterly, 45*(4), 464-487. https://doi.org/10.1598/RRQ.45.4.5

Crosson, A. C., McKeown, M. G., Moore, D. W., & Ye, F. (2019). Extending the bounds of morphology instruction: Teaching Latin roots facilitates academic word learning for English learner adolescents. Reading and Writing, 32(3), 689–727. https://doi.org/10.1007/s11145-018-9885-y

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. Developmental Review, 37, 66-108.

Falck, A. & Torkildsen, J.v.K. (2022). The leader learns it all? Using the “Kaptein Morf” tablet game to examine how different roles in joint problem solving affect learning. *Proceedings of the 17th SWECOG conference*, (pp. 7-8). http://swecog.se/files/SweCog_2022_Proceedings.pdf

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research, 74*(1), 59-109.

Gonnerman, L. M. (2018). A linguistic analysis of word morphology. In R. Berthiaume, D. Daigle, & A. Desrochers (Eds.), *Morphological processing and literacy development*, (pp. 3–15). Routledge. https://doi.org/10.4324/9781315229140

Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia, 60*, 183-208. https://doi.org/10.1007/s11881-010-0041-x

Goodwin, A. P., Petscher, Y., & Reynolds, D. (2022) Unraveling Adolescent Language & Reading Comprehension: The Monster’s Data, *Scientific Studies of Reading, 26*(4), 305-326. https://doi.org/10.1080/10888438.2021.1989437

Gómez, R. L. (2002). Variability and detection of invariant structure. Psychological Science, 13(5), 431-436. https://doi.org/10.1111/1467-9280.00476

Han, B., Koh, P. W., Zhang, S., Joshi, R. M., & Li, H. (2022). The relative contributions of facets of morphological awareness to vocabulary development in Chinese: A longitudinal study in grades one to three. *Contemporary Educational Psychology*, *69*, 102063. https://doi.org/10.1016/j.cedpsych.2022.102063

Hulme, C., Snowling, M., West, G., Lervåg, A., & Melby-Lervåg, M. (2020). Children’s language skills can be improved: Lessons from psychological science for educational policy. *Current Directions in Psychological Science*, *29*(4), 372–377. https://doi.org/10.1177/0963721420923684

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? Journal of Quantitative Linguistics, 21(3), 223-245. https://doi.org/10.1080/09296174.2014.911506

Levesque, K. C., Breadmore, H. L., & Deacon, S. H. (2021). How morphology impacts reading and spelling: Advancing the role of morphology in models of literacy development. *Journal of Research in Reading, 44*(1), 10–26. https://doi.org/10.1111/1467-9817.12313

Levesque, K. C., & Deacon, S. H. (2022). Clarifying links to literacy: How does morphological awareness support children's word reading development?. *Applied Psycholinguistics, 43*(4), 921-943. https://doi.org/10.1017/S0142716422000194

Lyster, S. A. H., Lervåg, A. O., & Hulme, C. (2016). Preschool morphological training produces long-term improvements in reading comprehension. *Reading and Writing, 29*(6), 1269-1288. https://doi.org/10.1007/s11145-016-9636-x

Montazami, A., Pearson, H. A., Dube, A. K., Kacmaz, G., Wen, R., & Alam, S. S. (2022). Why this app? How educators choose a good educational app. *Computers & Education*, 184, https://doi.org/10.1016/j.compedu.2022.104513

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English?. *Reading research quarterly*, 304-330.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. Reading Research Quarterly, 47(1), 91–108. https://doi.org/10.1002/RRQ.011

Nemmi, F., Helander, E., Helenius, O., Almeida, R., Hassler, M., Räsänen, P., & Klingberg, T. (2016). Behavior and neuroimaging at baseline predict individual response to combined mathematical and working memory training in children. Developmental Cognitive Neuroscience, 20, 43–51. https://doi.org/10.1016/j.dcn.2016.06.004

Nikolayev, M., Reich, S. M., Muskat, T., Tadjbakhsh, N., & Callaghan, M. N. (2021). Review of feedback in edutainment games for preschoolers in the USA. *Journal of Children and Media, 15*(3), 358–375. https://doi.org/10.1080/17482798.2020.1815227

Owen, V. E., Roy, M. H., Thai, K. P., Burnett, V., Jacobs, D., Keylor, E., & Baker, R. S. (2019). Detecting wheel-spinning and productive persistence in educational games. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.). *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, (pp. 378-383). Montreal: IEDMS. https://files.eric.ed.gov/fulltext/ED599202.pdf

Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 117-124).

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*(2), 184–202. https://doi.org/10.1598/RRQ.40.2.3

Plante, E., & Gómez, R. L. (2018). Learning without trying: The clinical relevance of statistical learning. *Language, speech, and hearing services in schools, 49*(3S), 710-722. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0131

Reed, D. K. (2008). A synthesis of morphology interventions and effects on Reading outcomes for students in grades K–12. *Learning Disabilities Research & Practice, 23*(1), 36-49. https://doi.org/10.1111/j.1540-5826.2007.00261.x

Ricketts, J., Lervåg, A., Dawson, N., Taylor, L. A., & Hulme, C. (2020). Reading and oral vocabulary development in early adolescence. *Scientific Studies of Reading, 24*(5), 380–396. https://doi.org/10.1080/10888438.2019.1689244

Rodrigo, M. M. T., Baker, R. S., Lagud, M. C., Lim, S. A. L., Macapanpan, A. F., Pascua, S. A. M. S., Santillano, J. Q., Sevilla, L. R. S., Sugay, J. O., Tep, S., & Viehland, N. J. (2007). Affect and usage choices in simulation problem solving environments. In R. Luckin, K. R. Koedinger, & J. Greer. (Eds.), *Artificial Intelligence in Education. Building Technology Rich Learning Contexts That Work.* (145-152). IOS Press.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. Wiley Interdisciplinary Reviews: Cognitive Science, 1(6), 906-914.

Shen, Y., & Crosson, A. C. (2023). Chinese adolescents learning to read in English: How do different types of morphological awareness contribute to vocabulary knowledge and comprehension?. *Reading and Writing, 36*(1), 51-76. https://doi.org/10.1007/s11145-022-10292-4

Snow, C. E., & Matthews, T. J. (2016). Reading and language in the early grades. *The Future of Children, 26*(2), 57–74. https://doi.org/10.1353/foc.2016.0012

Stanford, K. (2017). Underdetermination of Scientific Theory. In E.N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/

The Norwegian Directorate for Education and Training (2020). *Curriculum for Norwegian (NOR01-06)*. https://www.udir.no/lk20/nor01-06?lang=eng

The Norwegian National Committees for Research Ethics (2016). *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology.* Retrieved from https://www.etikkom.no/en/ethical-guidelines-for-research/guidelines-for-research-ethics-in-the-social-sciences--humanities-law-and-theology/

Torkildsen, J. v. K., Dailey, N. S., Aguilar, J. M., Gómez, R., & Plante, E. (2013). Exemplar variability facilitates rapid learning of an otherwise unlearnable grammar by

individuals with language-based learning disability. *Journal of Speech, Language, and Hearing Research, 56*, 618-629. https://doi.org/10.1044/1092-4388(2012/11-0125)

Tärning, B. (2018). Review of feedback in digital applications–does the feedback they provide support learning? *Journal of Information Technology Education: Research, 17*, 247-283. https://doi.org/10.28945/4104

Varga, S., Pásztor, A., & Stekács, J. (2022). Online Assessment of Morphological Awareness in Grades 2–4: Its Development and Relation to Reading Comprehension. *Journal of Intelligence*, *10*(3), 47. https://doi.org/10.3390/jintelligence10030047

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*(1), 3–32. https://doi.org/10.1080/10888430709336632

Wagner, R. K., & Quinn, J. M. (2019). The Co-Development of Vocabulary and Reading Comprehension. In J. Horst & J. v. K. Torkildsen (Eds.), *International Handbook of Language Acquisition* (pp. 504-519). Routledge.

Wang, T., & Zhang, H. (2023). Examining the dimensionality of morphological knowledge and morphological awareness and their effects on second language vocabulary knowledge. *Frontiers in Psychology*, *14*, 1207854. https://doi.org/10.3389/fpsyg.2023.1207854

Wechsler, D. (2009). *WISC-IV norsk versjon. Manual del 1*. NCS Pearson, Inc.

Winther, R.G. (2021). The Structure of Scientific Theories. In E.N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/spr2021/entries/structure-scientific-theories/.

Part II: Research papers

Article 1

Torkildsen, J. V. K., Bratlie, S. S., Kristensen, J. K., Gustafsson, J.-E., Lyster, S.-A. H., Snow, C., Hulme, C., Mononen, R.-M., Næss, K.-A. B., López-Pedersen, A., Wie, O. B., & Hagtvet, B. (2022). App-based morphological training produces lasting effects on word knowledge in primary school children: A randomized controlled trial. *Journal of Educational Psychology*, *114*(4), 833–854. https://doi.org/10.1037/edu0000688

Article 2

Kristensen, J. K., Andersson, B., Bratlie, S. S., & Torkildsen, J. V. K. (2023). Dimensionality of Morphological Knowledge—Evidence from Norwegian Third Graders. *Reading Research Quarterly, 58*(3), 406-424. https://doi.org/10.1002/rrq.497

INTERNATIONAL LITERACY ASSOCIATION READING RESEARCH QUARTERLY

Dimensionality of Morphological Knowledge—Evidence from Norwegian Third Graders

Jarl K. Kristensen

Björn Andersson

Centre for Educational Measurement, University of Oslo, Oslo, Norway

Siri S. Bratlie

Department of Education, University of Oslo, Oslo, Norway

Janne V. K. Torkildsen

Department of Special Needs Education, University of Oslo, Oslo, Norway

ABSTRACT

This study aimed to determine the dimensionality of morphological knowledge by examining different sources of variance. According to the Morphological Pathways Framework (*Levesque et al., Journal of Research in Reading, 44, 10-26, 2021*), morphological awareness, morphological analysis and morphological decoding are related, but distinct dimensions of morphological knowledge. However, multidimensionality might also stem from construct-irrelevant variance due to methodological artifacts. We assessed 612 Norwegian third graders on five measures of morphological knowledge and one measure of general vocabulary. Fitting a series of confirmatory factor analysis (CFA) models, we evaluated the dimensionality of morphological knowledge both within and across the five tests. Furthermore, we fitted three structural equation models (SEMs) to explore how different conceptualizations affect the relationship between morphological knowledge and general vocabulary: a five-factor model, a bifactor model, and a higher-order model representing morphological awareness, morphological analysis and morphological decoding. CFAs supported a multidimensional view of morphological knowledge and highlighted the need to account for construct-irrelevant variance. SEM analyses further illustrated that construct-irrelevant variance introduces a confounding element to the relations between morphological knowledge and vocabulary in the test-specific five-factor model, as only the bifactor and higher-order models separate between construct-relevant variance and variance due to methodological artifacts. The bifactor model is useful for separating sources of variance, especially during test development. For research purposes, however, we recommend conceptualizing morphological knowledge in line with *Levesque et al., Journal of Research in Reading, 44, 10-26, 2021*, to increase knowledge of morphological dimensions and their relations to other areas of literacy.

Introduction

In this study, we investigate the dimensionality of morphological knowledge in Norwegian third graders. More specifically, we examine whether tests measuring morphological awareness, morphological analysis and morphological decoding represent a single underlying construct or different dimensions of morphological knowledge. Understanding the dimensionality of the construct is crucial to advance research on morphological knowledge and its relations to other language skills. If the construct is multidimensional, we need to take this into account when comparing results from studies using different measures of morphological knowledge. While it is not a target of the current study, dimensionality may also have implications for the design and evaluation of morphological assessments and interventions.

INTERNATIONAL LITERACY ASSOCIATION

Reading Research Quarterly, 58(3)
pp. 406–424 | doi:10.1002/rrq.497

Morphological knowledge is the ability to recognize, understand, manipulate and produce spoken and written morphemes, the smallest meaning-bearing units of language. It requires knowledge of both the form and meaning of morphemes, as well as the processes through which they can be combined (Nagy et al., 2014). In addition to the term morphological knowledge, the two related terms morphological awareness (e.g., Carlisle, 2010) and morphological processing (e.g., Verhoeven & Perfetti, 2011) are widely used. Morphological awareness refers to explicit morphological knowledge, as it requires conscious reflection on and manipulation of morphemes (Levesque et al., 2021). Morphological processing, on the other hand, refers to the implicit use of morphological knowledge, which may happen at a subconscious level (Bowers et al., 2010; Nagy et al., 2014).

Levesque et al. (2021) introduce the Morphological Pathways Framework. The framework provides a theoretical foundation for morphological knowledge, in which the authors conceptualize it as a multidimensional construct. However, the findings in extant empirical research are mixed. Some studies report evidence of a single dimension of morphological knowledge (e.g., James et al., 2021; Spencer et al., 2015), whereas others propose different dimensions of morphological knowledge such as oral versus written or receptive versus productive (e.g., Jong & Jung, 2015; Tibi & Kirby, 2017). Thus, it is unclear if the theoretical framework suggested by Levesque et al. (2021) is generally applicable across different populations and settings.

Additionally, most studies on the dimensionality of morphological knowledge to date have been conducted in English. Of the 13 studies reviewed in this paper, eight featured English-speaking participants (see Table S1 and the section on empirical studies of dimensionality for more information). However, in their study of English and Korean, Jong and Jung (2015) found evidence of cross-linguistic differences. In English, they found one receptive and one productive dimension, whereas in Korean they found one receptive and two productive dimensions. While this points to possible cross-linguistic differences in dimensionality, it is not clear whether these differences relate to morphological knowledge or a construct-irrelevant source of variance. It is also unclear if such differences exist between languages more closely related than English and Korean.

Dimensionality studies in other areas of language have suggested a developmental trend moving from a single factor that captures language competence in preschool to multidimensional representations in older children (Tomblin & Zhang, 2006). While the existing studies span age ranges from preschool to adulthood, there is a need for further examination of the dimensionality of morphological knowledge in younger children. For morphological tasks which rely on written language, the development from a single factor to multidimensional representations may be affected by the orthographic transparency of the language. Specifically, morphological decoding and analysis may be distinguishable earlier in more orthographically transparent languages where decoding skills place a severe constraint on analysis for only a limited developmental period. Considering the potential impacts of language and age on dimensionality, we aim to add to the current knowledge by examining the construct in Norwegian third graders.

Norwegian Language and Morphology

In many languages, morphology plays an important role in word formation through inflection, derivation and compounding (Gonnerman, 2018), as is also the case in Norwegian. Inflection modifies a word's grammatical features, such as tense (hoppe—hopp*et*, "jump—jump*ed*"), number (blomst—blomst*er*, "flower—flower*s*") or grammatical gender (et fin*t* hus, "a-*neuter* nice-*neuter* house"). Derivation, on the other hand, creates entirely new dictionary words (lexemes), which can change a word's part of speech (spise—spise*lig*, "eat—edi*ble*") and often result in a derived word with a completely different meaning than the base word (tanke—tanke*løs*, "thought—thought*less*"). Compounding also creates new words but does so by joining two bases (sove*rom*, "bedroom"; korrektur*lese*, "proofread") rather than joining a base and a derivational affix.

Norwegian is a Germanic language with a simple verbal morphology (no subject–verb agreement), but a more complex nominal morphology, including three grammatical genders and noun–adjective agreement. Both compounding and the compilation of derivational affixes are widely used as means of word formation in Norwegian. Compounding is a highly productive process in Norwegian, and, thus, closed compounds that can consist of three, four or even more base words are common (e.g., *menneskerettighetsorganisasjon* = human rights organization). Words with three or more derivational affixes are also common (e.g., *u-be-hjelpe-lig* = helpless). A number of the Norwegian derivational affixes are similar to those found in English (e.g., "over-" and "mis-"). Many of the Norwegian derivational affixes and compounds are typical of written language and are, thus, particularly relevant for comprehending and producing academic texts. The Norwegian orthography is morpho-phonetic, and the phoneme–grapheme relationships are more transparent than in English (Seymour et al., 2003). There is a persistent influence of morphology on Norwegian orthography (Lyster, 2002), and morphological features determine the spellings of many words, along with phoneme–grapheme correspondence. Additionally, many high-frequency inflectional and derivational suffixes contain silent letters (e.g., the neuter definiteness marker "-et" /e/ and the common derivation "-lig" /li:/).

The literature on morphological development in Norwegian is scarce and focuses on acquisition of inflections in

preschool children (for an overview, see Ribu et al., 2019). One study of past tense acquisition included children up to early primary school age and found that the overwhelming majority of children have reached ceiling performance by age eight (Ragnarsdóttir et al., 1999). Derivational knowledge has only been examined in one study, which showed that for 5-year-olds the mean performance in a derivational task was substantially lower than the performance in similar tasks measuring inflectional knowledge (Grande, 2018). This result supports the common pattern found in studies of many Indo-European languages that inflectional knowledge is typically acquired earlier than derivational knowledge (Kuo & Anderson, 2006).

In sum, there is a need for studies of the acquisition of derivations and compounds in Norwegian. It is especially important to study these morphological skills in children from third grade. Most Norwegian children are skilled decoders by that age (Hagtvet et al., 2006), and consequently, curriculum texts become more complex, including advanced vocabulary with derived and compounded words. The current study, thus, focused on derivational and compound knowledge in Norwegian third graders. Measures of inflectional knowledge were not included, as previous studies indicate near-ceiling performance for nominal inflections before age 3 and for verbal inflections by age 8.

Dimensionality of Morphological Knowledge

Theoretically, the dimensionality of morphological knowledge depends on whether construct-relevant variance relates to one or more morphological skills. However, construct-irrelevant variance might also be a source of multidimensionality, which stems from specific task requirements, formats or content (e.g., Deacon et al., 2008). Hence, tests of morphological knowledge may measure a multidimensional construct of which only one dimension relates to morphology. In the following sections, we review the Morphological Pathways Framework as a theoretical foundation for understanding construct-relevant dimensionality, present potential construct-irrelevant sources of variance, and summarize findings from previous empirical studies on the dimensionality of morphological knowledge.

The Morphological Pathways Framework

A large body of research has shown that morphological knowledge predicts vocabulary, reading fluency and reading comprehension in many languages (James et al., 2021; Manolitsis et al., 2019; McBride-Chang et al., 2005), and that morphological instruction can enhance children's word knowledge and reading development (e.g., Bowers et al., 2010; Carlisle, 2010; Goodwin & Ahn, 2010; Lyster et al., 2016; Reed, 2008; Torkildsen et al., 2022). The Morphological Pathways Framework introduced by Levesque et al. (2021) provides a theoretical model of the mechanisms behind the influence of morphology on other areas of literacy. Furthermore, it provides a theoretical base for viewing morphological knowledge as multidimensional.

In this framework, the authors present three dimensions of morphological knowledge which influence reading comprehension and writing: morphological awareness, morphological decoding and morphological analysis. Morphological awareness is viewed as a metalinguistic skill that involves the conscious reflection on and manipulation of morphemes. Morphological decoding relates to morpho-orthographic segmentation, that is, the recognition of separate morphemes in written words. This is also referred to as form-based skills, as they operate at the level of orthography, or word form (Levesque et al., 2021; Nagy et al., 2014; Torkildsen et al., 2022). Morphological analysis is a morpho-semantic process and involves the recognition of the meaning of separate morphemes within words. This process operates at the level of semantics and is also referred to as meaning-based skills (Levesque et al., 2021; Nagy et al., 2014; Torkildsen et al., 2022). The Morphological Pathways Framework posits reciprocal relations among morphological awareness, morphological decoding and morphological analysis. These three skills represent related, yet distinct, dimensions of the overarching construct of morphological knowledge.

The framework involves different pathways between morphological awareness, morphological analysis and morphological decoding, and other areas of language including text comprehension and generation. Along these paths, we also find connections to word reading, spelling and word knowledge. Morphological awareness is related to knowledge of word meanings through morphological analysis, thus affecting general vocabulary. Specifically, morphological analysis can support inferences about the meanings of morphologically complex words through the meanings of their constituent morphemes (Levesque et al., 2019). Morphological decoding forms the bond between morphological awareness and word reading by enabling letter-sound mapping at the level of morphemes rather than graphemes (Levesque et al., 2021). The relation to spelling is still somewhat unclear, as little research exists in this area. According to Levesque et al. (2021), it is possible that both morphological decoding and morphological analysis are involved.

While the Morphological Pathways Framework provides a theoretical basis for the multidimensionality of morphological knowledge, it is also evident from the literature that researchers measure morphological knowledge with a large number of different tasks with different task requirements (Berthiaume et al., 2018). These requirements are methodological artifacts that introduce construct-irrelevant variance. Hence, they represent confounding factors in research.

Methodological Artifacts

Morphological tasks vary in input and output modality, content, task type, as well as demands on information processing and prior knowledge (Berthiaume et al., 2018; Deacon et al., 2008). Input and output modality concerns whether tasks presentation (input) or responses (output) are oral or written. Berthiaume et al. (2018) describe 10 different task types that are commonly used to measure morphological knowledge: decomposition, definition, lexical decision, derivation, morphological relation judgment, naming, plausibility judgment, spelling, suffix choice, and word analogy. These involve different knowledge demands. Knowledge demands relate to the distinction between awareness and processing. Some tasks, like word analogies, require explicit morphological awareness. Other tasks may rely on implicit morphological processing, for example, word explanations, where morphological analysis may operate at a subconscious level. Finally, tasks may tap into different additional skills such as phonological decoding or general vocabulary, giving rise to construct-irrelevant variance in item responses.

To sum up, multidimensionality in morphological measures can stem from "true" multidimensionality in morphological knowledge. On the other hand, it may also stem from construct-irrelevant variance due to methodological artifacts.

Empirical Studies on the Dimensionality of Morphological Knowledge

Many studies on language and language development utilize measures of morphological knowledge, but few have investigated the dimensionality of the construct explicitly (Goodwin et al., 2017). For the current study, a literature review yielded 13 papers that examined the dimensionality of morphological knowledge (see Table S1 in the Supplementary material for a detailed overview). To evaluate the dimensionality of morphological knowledge, these studies implement a range of statistical models, including single-factor models, correlated traits models, and bifactor models. A single-factor model implies that a single skillset of morphological knowledge underlies test performance. In a correlated traits model, subsets of items or indicators tap into different, correlated factors. Finally, a bifactor model implies that a general factor of morphological knowledge explains the correlation among all items in a test, while there are also specific uncorrelated factors that account for residual correlations among the item scores in separate subtests, beyond what the general factor can explain.

Some previous studies found evidence supporting a unidimensional conceptualization of morphological knowledge (James et al., 2021; Muse, 2005; Spencer et al., 2015; Tibi, 2016; Tibi & Kirby, 2017). Note that Spencer et al. (2015) report analyses of the same data as Muse (2005), and Tibi and Kirby (2017) report on the same data as Tibi (2016). Findings from these studies suggest that morphological knowledge is best represented as a single skillset. Although contrary to the Morphological Pathways Framework at first glance, these findings could relate to the reciprocal nature of morphological awareness, morphological analysis and morphological decoding. Some measures of morphological knowledge may not sufficiently distinguish between the three skills, and different models may provide acceptable fit to the data. For example, the written tasks in Muse (2005) and Spencer et al. (2015) were read aloud by the test administrator and did not require written responses. Thus, they did not test morphological decoding specifically. The written tests of Tibi (2016) did require participants to read, and in some tasks write the answer, thus measuring morphological decoding. Accordingly, Tibi and Kirby (2017), in an extension of the analyses of the data from Tibi (2016), found that a two-factor model also represented the data well. The two factors were related to the oral and written tests, respectively, thus aligning with the theoretical constructs of morphological analysis and morphological decoding.

Other studies have found support for a multidimensional structure of morphological knowledge (González-Sánchez et al., 2018; Jong & Jung, 2015; Levesque et al., 2017; Tighe & Schatschneider, 2015, 2016; Zhang, 2017). Both González-Sánchez et al. (2018) and Jong and Jung (2015) reported separate dimensions of receptive and productive morphological knowledge in their studies. Their studies targeted Spanish children in the last year of preschool (González-Sánchez et al., 2018) and Korean fifth and sixth graders (Jong & Jung, 2015). Tighe and Schatschneider (2015, 2016) studied morphological knowledge in English-speaking Adult Basic Education students and found evidence that a two-factor model separating real words and pseudowords fit the data best. A common finding in all these studies is that response format is a source of multidimensionality. This is not related to morphological knowledge as such, but rather to how we measure it and the additional skills required to respond. This might indicate that potential differences relating to age and language stem from construct-irrelevant sources rather than differences in morphological knowledge.

Levesque et al. (2017) examined morphological knowledge in English-speaking third graders. They measured the theoretically founded skills of morphological awareness, morphological decoding, and morphological analysis. Comparing unidimensional and multidimensional models, they found that a model representing each skill as a separate factor fit the data best. Zhang (2017) found similar results in a study of the morphological knowledge of Singaporean fourth graders speaking both Chinese and English. A two-dimensional model aligning with the theoretical constructs of morphological analysis

and morphological decoding fit the data well. These studies, along with Tibi and Kirby (2017), provide support for the theoretical dimensions introduced in the Morphological Pathways Framework.

Goodwin et al. (2017) administered seven morphological tasks to English-speaking seventh and eighth graders. The authors found evidence that a bifactor model performed best, meaning that the tasks measured a general factor of morphological knowledge, as well as seven specific factors related to each of the seven types of tasks. This bifactor model of morphological knowledge was further explored by Goodwin et al. (2021), in which they reported that morphological knowledge was best represented by four skill-related (general) factors as well as task-specific factors. The four general factors align with morphological awareness, morphological analysis and morphological decoding, with the addition of a factor representing morphological-syntactic knowledge. Following an inherent assumption in bifactor models, however, the four factors representing morphological skills are uncorrelated, not taking into account the relations posited in the Morphological Pathways Framework.

When considering a structural model where general vocabulary and reading comprehension were regressed on each factor in the bifactor model for morphology, Goodwin et al. (2017) found that the general factor of morphological knowledge explained most of the variance in both vocabulary and reading comprehension. However, additional variance in reading comprehension was explained by the specific factors of morphological meaning (positive), and morphological spelling and word reading (negative). Additional variance in vocabulary was explained by morphological meaning and word generation (positive), and spelling (negative). These results suggest that general and task-specific morphological skills may have a distinct involvement in different literacy tasks.

While there are differences between studies, there are no consistent patterns relating to language or age. Some differences relate to test format, and in the studies that support a unidimensional view, the tests do not necessarily separate between theoretically founded dimensions. Importantly, the differences underline the importance of separating construct-relevant variance from variance that does not relate to morphological knowledge.

Current Study

The purpose of our study is to investigate the dimensionality of morphological knowledge in Norwegian third graders. We examine whether a unitary construct of morphological knowledge underlies five tests that measure different aspects of the participants' knowledge of morphologically complex words: receptive word knowledge, productive word knowledge, word analogies, spelling and word reading fluency. Furthermore, we examine how different conceptualizations affect the relation between morphological factors and general vocabulary.

This study builds on data from a randomized controlled trial (RCT) of a morphological intervention with Norwegian second graders who were followed until third grade (Torkildsen et al., 2022). Participating students were randomly assigned to an eight-week digital morphology program or an active control group. The program consisted of 40 training sessions targeting derivational morphology (26 common derivational morphemes) and compounding processes in Norwegian. The training targeted both morphological decoding and morphological analysis. In line with this, we developed our five outcome measures to tap both of these constructs, in addition to morphological awareness. Specifically, the tests of receptive and productive word knowledge measure morphological analysis, while the spelling and word reading fluency tests measure morphological decoding. The word analogy test measures morphological awareness. For more information, see the test descriptions in the methods section.

As previous studies have provided evidence both for unidimensionality and multidimensionality, we compare several different models that may represent the construct based on these previous findings. As a part of this investigation, the bifactor analyses of Goodwin et al. (2017) are considered for a new age group and a new language. We also include a higher-order model to examine the dimensional structure suggested by Levesque et al. (2021), including a mediation model similar to those examined by Levesque et al. (2017). Both the bifactor framework and the Morphological Pathways Framework hold promise of producing a deeper understanding of this complex area of language, yet few studies have implemented them to date. Hence, we examine these frameworks in the context of our study, to provide further evidence on their applicability when measuring and analyzing morphological knowledge.

Our study was guided by the following three research questions:

1. Do the five tests of morphological knowledge each measure a unidimensional construct?
2. Is morphological knowledge best represented as a unidimensional or multidimensional construct across the five different tests?
3. How do different models affect the relation between morphological knowledge and general vocabulary?

For research question 1, we hypothesized that each test captures a unidimensional facet of morphological knowledge. Some of the tests, however, include items with features that may influence the measured construct. The test of receptive word knowledge measures morphological knowledge in context as well as in isolation and consists of three different item types (see the methods section for

details). Tighe and Schatschneider (2015) examined context versus no context as potential dimensions of morphological knowledge in adults. Although their results did not support this dimensional dichotomy, these might constitute separate dimensions in children. Additionally, the different item types might pose different task demands, thus reflecting different dimensions. The Test of Productive Word Knowledge measures morphological knowledge with real words and pseudowords. Although Jong and Jung (2015) did not find evidence of a real word versus pseudoword division in children, Tighe and Schatschneider (2015, 2016) did find evidence for separate dimensions in adults. Lastly, both the spelling test and the tests of productive and receptive word knowledge measure each specific affix in more than one task. Thus, there is a possibility that the tests of receptive and productive word knowledge may be best represented as multidimensional. Although not representing theoretical dimensions of morphological knowledge, the affix-specific knowledge may cause dependence among items beyond the common variance due to morphological knowledge.

Regarding research question 2, we hypothesized that a common construct of morphological knowledge underlies item responses across all five tests. This could align with studies that support morphological knowledge as a unidimensional construct (James et al., 2021; Muse, 2005; Spencer et al., 2015; Tibi, 2016; Tibi & Kirby, 2017). However, the tests also differ in task demands (e.g., comprehension, production, analogies, reading fluency, and writing). Thus, we hypothesized that the tests may measure other test-specific skills as well as the common factor of morphological knowledge. Additionally, two of our morphological tests measure morphological decoding, two tests measure morphological analysis skills, and one test measures morphological awareness. Hence, we examine whether morphological awareness, morphological decoding and morphological analysis are separate dimensions of morphological knowledge, in line with Levesque et al. (2021).

Finally, with regard to research question 3, the literature points towards strong relationships between morphological knowledge and vocabulary (e.g., McBride-Chang et al., 2005; Nagy et al., 2006). Hence, we hypothesized that potential dimensions of morphological knowledge should have significant positive relations to general vocabulary. However, the results of Goodwin et al. (2017) suggested that these relations are different if morphological knowledge is accounted for in a general factor (i.e., in a bifactor model). Thus, for the bifactor model, we hypothesized that general morphological knowledge, as well as specific receptive and productive word knowledge (morphological analysis) have a significant and positive relation to general vocabulary (which is measured by a meaning-based definition task), whereas the specific skills related to word analogies, reading fluency and spelling (morphological awareness and decoding) have non-significant or negative relationships to general vocabulary, in accordance with Goodwin et al. (2017). In line with Levesque et al. (2021), we expected a significant and positive relation between morphological analysis and vocabulary, as well as an indirect effect of morphological awareness through analysis, in the mediation model.

Methods

Design

The participants in the intervention study were tested before starting the program (pre-test), directly after the program (post-test), and at follow-up, which was approximately 9 months after the pre-test. The current study analyzes data from the follow-up, which was administered during the participants' first semester in the third grade. The decision to use the data from third grade was made to include a word analogy test, which was only administered at this grade level, as a measure of morphological awareness. This enabled us to examine as many potential theoretical and empirical dimensions of morphological knowledge as possible.

Participants

The participants were 612 third graders (n = 325 girls, n = 286 boys, and n = 1 with missing information) from 12 schools in the eastern part of Norway. The approximate mean age was 8.34 years (SD = 0.3). All students in each classroom were invited to participate, with a positive response rate of 93%. Schools were recruited by municipality officials, who were instructed to select schools with different characteristics (average SES and proportion of language minority students) to help make the sample representative of schools in the area. The morphology training program required that schools had access to iPads for all children participating; hence the schools were not randomly selected. Across schools, the proportion of mothers with a university education ranged from 28.3% to 95.7% (mean for the whole sample = 72.9%), and the proportion of students with a language minority background ranged from 2.8% to 93.6% (mean for the whole sample = 28.8%). Ethical approval to conduct the study was granted by the Norwegian Centre for Research Data.

Measures

All measures were administered as part of a larger test battery, either individually or in groups (full classes). As there is a lack of standardized tests of morphological knowledge in Norwegian, these five tests of morphological knowledge were developed within the project. All tests were piloted in several rounds with approximately 200 children who did not participate in the current study. As mentioned, these tests were selected to measure learning outcomes in the

intervention study, not primarily to assess dimensionality. However, we include information on how the measures relate to theoretical and empirical perspectives on dimensionality in the description of each test. Table 1 provides an overview of the measures, including Cronbach's α (ranging from .80 to .96).

The derivations targeted in the intervention program were selected based on frequency information from language corpora, utility and familiarity from a pilot study of 100 fourth graders. The fourth graders' knowledge of 96 derivations was rated on a scale from 0–2 where 0 indicated no knowledge, 1 indicated some knowledge (often highly specific), and 2 indicated more advanced general knowledge. To ensure that the derivations were not only already mastered by most second graders but also not too advanced for them, we selected 26 derivations in which 40–70% of fourth graders demonstrated at least some knowledge. For more information on morpheme selection, see Torkildsen et al. (2022). All the words used in the morphological measures were multimorphemic (e.g., consisting of an affix and a base word). Half of the test items in measures 2–5 contained *exposed words* (i.e., words that were included in the app training sessions) and half of the test items contained *unexposed words* (i.e., words that were not included in the training, but which contained trained affixes). The word analogy test did not contain any exposed words.

Morphological Awareness

Morphological awareness was measured with the word analogy test, in line with Levesque et al. (2017). The test focuses on extracting the bases of derived words, that is, words which are made up of a derivational affix and a base (for an example, see measure 1 in Table 1). This requires knowledge of morpheme boundaries and segmentation. As both presentation and response are given orally, the test does not rely on morphological decoding. The test, adapted from Brinchmann et al. (2016) and Bryant et al. (1997), consists of 15 items. The test administrators first presented a derived word containing a given affix and extracted the base from the derived word. Then another derived word containing the same affix was presented, and the children were prompted to extract the base. The test was administered individually, and item scores were binary (0, 1).

Morphological Analysis

Morphological analysis was measured with two tests that focus on the meaning of words, similarly to Levesque et al. (2017) and Goodwin et al. (2021). The test of receptive word knowledge (see measure 2 in Table 1) measures comprehension of morphologically complex words and consists of 48 multiple choice items covering 20 affixes (each appearing in 2 tasks with different base words), 6 compound words and 2 words with multiple affixes. The

TABLE 1
Overview of the Tests of Morphological Knowledge and General Vocabulary

Measure	Task example(s)	Items (final)	Cronbach's α
1) Word analogy test	"I say the word *typical*, then change it to *type*. We can also change the word *magical* to …"	15 (14)	.80
2) Test of receptive word knowledge	*Word combination*: "Which part can you put after *de-* to make a real word? [*sit, pict, shake, song*] *Cloze tasks*: "Janne wanted to stay in the student council. She hoped for a (…)*election*. [*re, new, well, after*]" *Picture tasks*: "Press the picture that shows *overexertion*."	48 (26)	.85
3) Test of productive word knowledge	*Real words*: "What does *machinist* mean?" *Pseudowords*: "What could *busist* have meant, if it were/had been a real word?"	18 (13)	.80
4) Spelling test	"It was a happy *reunion*. Write *reunion*."	24 (24)	.92
5) Word reading efficiency test	Timed reading of randomized lists of morphologically complex words (i.e., not sorted by difficulty). 30 second time limit.	4 (3)*	.96
6) WISC-IV vocabulary subscale	"What is a thief?"	3 (3)**	.84

Note. All examples are translated from Norwegian. Cronbach's α reported for final item sets. *Four lists of 48 words each (sum scores), all four used in the within-test model, three of the lists retained in the across-tests models. **Three parcels of 12 items each (sum scores).

test was administered digitally. Tasks were presented orally and in writing, in a multiple-choice format with one correct option and three distractor options. The tasks were divided into three different types: morpheme choice tasks, cloze tasks and picture tasks. In the morpheme choice tasks participants were asked to match an affix with a base to form a real word. In the picture selection tasks, participants identified the most appropriate picture in response to a morphologically complex word. The cloze tasks required participants to select an affix to fill a blank in a sentence. Examples of the three task types are given in Table 1. While cloze and picture tasks provided context through the sentences and pictures, the morpheme choice tasks did not. The test was administered in group sessions. Item scores were binary (0, 1).

The second measure of morphological analysis was the Test of Productive Word Knowledge (see measure 3 in Table 1). The test measures the ability to define morphologically complex words and pseudowords. The test covers six affixes, with three items for each affix and, thus, a total of 18 items. Each affix was presented as part of a real word in two of the tasks and as part of a pseudoword in the third task. Pseudowords were created by adding an affix to a regular Norwegian base, creating a nonexistent but plausible word (e.g., *bussist* = busist, which could mean "a person who drives/rides/likes buses"). The test was administered individually, with oral presentation and oral responses. Partial scoring in three categories was used (0, 1, 2). Two points were awarded for synonyms or precise explanations of the meaning of a word and one point was awarded for definitions that reflected only vague knowledge of the word's meaning. Pseudoword explanations were scored for knowledge of what an affix does to the meaning of a word.

Morphological Decoding

Morphological decoding was measured with two tests focusing on the written form of words, in line with Levesque et al. (2017) and Goodwin et al. (2021). The spelling test (see measure 4 in Table 1) measures the ability to spell morphologically complex words with nontransparent spelling patterns. The test consists of 24 morphologically complex words covering 11 derivational affixes, each included in two items, and two items with compound words. The words were first presented in the context of a sentence and then repeated in isolation. The children were then asked to write the target word of each item on a sheet of paper. The test was administered in groups and partial scoring was used. 0–3 points were given for words with derivations (1 point, respectively, for the correct spelling of the affix, the correct spelling of the base word, and writing the morphemes together with no space between, following Norwegian orthographic rules) and 0–2 points for compound words (1 point respectively for correct spelling of the base words and writing of the compounds together with no space between).

The second measure of morphological decoding was the word reading efficiency test (see measure 5 in Table 1), which measures word reading fluency and accuracy. It consists of four lists, each containing 48 morphologically complex words, covering both derivations and compound words. The children were asked to read as many words aloud from each list as they could in 30 seconds. Children were instructed to read the words in the order they were presented, but if unable to read an attempted word, children could skip to the next word on the list. The test was administered individually and sum scores from each list were used for analyses. Note that while there are 192 items across the four lists, we only have four sum score indicators. No items were excluded in the process.

General Vocabulary (Word Definitions)

Using word definitions as a proxy for general vocabulary, we measured this construct with the Vocabulary subtest from the Norwegian 2009 version Wechsler Intelligence Scale for Children® Fourth Edition (WISC-IV; Wechsler, 2009). See measure 6 in Table 1 for an example. This test measures the ability to explain the meaning of words. It consists of 36 items. The test was administered individually, according to the manual. As specified in the manual, the test was discontinued after five consecutive errors. Items were parceled into three sum scores which were used as indicators of general vocabulary in the analyses.

Analyses

Our analyses consisted of three distinct parts, which reflected research questions 1, 2, and 3, respectively. All analyses were conducted in R (R Core Team, 2020), using the packages psych (Revelle, 2021) for descriptive statistics and lavaan (Rosseel, 2012) for the factor analyses and structural equation modeling. The proportion of missing data ranged from 5% to 9% for the models within tests. For the models across tests, the proportion was 4%. Models were estimated based on the observed pairwise information between pairs of variables to minimize the loss of information due to missing responses. We compared this procedure to listwise deletion, and there was no impact on any conclusions of the study.

Descriptive Statistics

Table 2 provides descriptive statistics for total scores on each test. As the data come from an RCT study, we show the statistics for the experimental and control groups separately. The table reports on both the pre-test in second grade and the follow-up in third grade, which was the measurement point in focus in the current study. For item-level statistics, see Tables S2–S7 in the supplementary materials. The patterns of means, standard deviances,

TABLE 2
Descriptive Statistics (Sum Scores) for All Measures at the Pre-Test in Second Grade and Follow-Up in Third Grade

Test	Group	*n* P/F	*M* P/F	*SD* P/F	Range P/F	Skewness P/F	Kurtosis P/F
Word analogy	E	NA/290	NA/7.97	NA/3.52	NA/14	NA/−0.68	NA/−0.27
	C	NA/292	NA/7.89	NA/3.31	NA/15	NA/−0.48	NA/−0.27
Receptive knowledge	E	308/276	17.38/22.98	5.30/8.31	33/41	0.69/0.07	0.62/−0.66
	C	282/278	16.80/21.43	5.54/7.54	33/46	0.78/−0.23	0.73/−0.29
Productive knowledge	E	298/291	11.83/17.74	6.10/6.51	28/30	0.16/−0.25	−0.80/−0.54
	C	287/291	11.53/15.91	5.97/6.39	29/28	0.24/−0.01	−0.28/−071
Spelling	E	306/293	47.95/54.37	11.14/9.21	68/68	−1.34/−1.30	3.27/3.33
	C	297/288	47.88/52.59	11.17/10.20	68/70	−1.60/−1.51	4.22/4.20
Word reading	E	298/291	21.43/39.28	15.61/25.07	97/152	1.35/0.94	2.82/1.03
	C	286/290	22.34/37.65	18.28/25.63	156/183	2.39/1.30	11.29/3.20
Vocabulary	E	298/291	16.83/19.25	5.14/5.71	31/33	0.18/0.37	0.12/−0.08
	C	287/290	17.24/18.87	5.28/5.74	36/36	0.38/0.33	0.98/0.22

Note. C = control group, E = experimental group, F = follow-up, P = pre-test. MCWA was administered at follow-up only.

skewness and kurtosis are similar across the groups, with the exception of the kurtosis of the word reading efficiency test in the control group at pre-test. The experimental group had larger increases in means overall than the control group from the pre-test to the follow-up.

We tested measurement invariance between the experimental and control groups on all exposed items of the tests. These are the test items that contain morpheme combinations that the experimental group has experienced through the tasks in the intervention. As mentioned, the word analogy test does not contain any exposed items. We used chi-square difference tests to test the null hypotheses of measurement invariance against lack of measurement invariance with a Bonferroni-corrected significance level of 0.0125. The results indicated that all exposed items functioned equally for participants in the experimental group and the control group (see Table 3).

Research Question 1

Research question 1 concerned the overall item quality and dimensionality in each of the tests separately. This study provides the first in-depth psychometric evaluation of the tests. Hence, we went through several steps before arriving at the final models. In the first step, correlations between item scores and total scores for each test were calculated, and items with $r < .3$ were excluded from further analyses (Nunnally & Bernstein, 1994). This resulted in the exclusion of five of the 48 items in the test of receptive word knowledge and one of the 15 items in the word analogy test. These items represented noise, likely due to too high difficulty and unintended item features. For example, one item in the test of receptive word knowledge had the target "løsbart". This word can mean either "solvable" (correct response) or "false mustache" in Norwegian, and the only difference lies in the pronunciation. The response options were pictures, of which one could be mistaken to depict a false mustache. Hence, a large number of children chose the confounding distractor. Removing these items did not change the substantive interpretation of the underlying constructs, nor did they change the possible dimensional structures of either test that were evaluated in the subsequent analyses.

TABLE 3
Measurement Invariance Tests

Model	x^2	*Df*	Δx^2	Δdf	*p*
Receptive (p)	505.663	618			
Receptive (f)	615.547	648	47.431	30	.023
Productive (p)	131.710	134			
Productive (f)	169.073	149	26.843	15	.030
Spelling (p)	719.084	526			
Spelling (f)	826.630	572	60.488	46	.074
Word reading (p)	17.739	8			
Word reading (f)	17.750	10	0.011	2	.994

Note. (p) = partial invariance, exposed items free to vary. (f) = full invariance, all items restricted.

In the second step, we considered unidimensional confirmatory factor analysis (CFA) models for each test and evaluated the model fit. Note that for the spelling test and the test of productive word knowledge, the models were specified with correlated residuals between items containing the same affix. We used polychoric correlations with the diagonally weighted least squares (DWLS) estimator for the ordinal data and the ML estimator for continuous data. All the unidimensional models fit the data well, indicating that a single construct underlies responses to each test. As there was a very large amount of indicators across the tests, we decided to exclude items with standardized factor loadings that were below .4 from further analyses. This choice was made to reduce complexity and facilitate the analyses across tests. Note that this item exclusion was carried out after establishing unidimensionality for each test. As cutoff values for considering a factor loading salient vary in the literature (e.g., Brown, 2015), a strict cutoff was deliberately chosen to reduce the vast amount of indicators in the final models containing all tests. This second analysis step resulted in the exclusion of another 17 items from the test of receptive word knowledge (in addition to the five items excluded in step 1), retaining 26 items. From the Test of Productive Word Knowledge, we removed five items, keeping 13 items. We did not exclude any further items from the word analogy test. The spelling test and word reading efficiency test were also kept intact, as there was no factor loading < .4 in the models for these tests. Note, however, that we excluded one indicator from the word reading efficiency test at a later stage, outlined in the next section. The number of items, original and final, are reported in Table 1. The exclusion of items, though substantial, did not affect the substantive or statistical interpretations of the constructs. It did, however, increase the fit indices and coefficient alphas to some extent. For the sake of brevity and continuity, we present the results for the final models based on the reduced item sets in the next chapter, as these are the item sets we use in the subsequent analyses across tests.

Research Question 2

To address research question 2, models containing the retained items from all tests were evaluated to examine dimensionality across the measures. We fit a series of nested CFA models: one-factor (morphological knowledge); three-factor (morphological awareness, morphological analysis, and morphological decoding); five-factor (test-specific); and higher-order (morphological awareness, morphological analysis and morphological decoding). Note that in the higher-order model, morphological awareness is represented as a first-order factor, since we tested this construct with a single test. Figures S1–S4 show conceptual illustrations of these models. The observed indicators from the word reading efficiency test were very highly correlated (ranging from .86 to .90), which caused empirical underidentification in the initial analyses across tests. Thus, to estimate the models, we removed one variable. Because of the high correlations, this did not change the substantive interpretation of the Word Reading factor, nor its contribution to the models across tests. We used chi-square difference tests to select among the models, with a significance level of 0.05. In the last step of the measurement models, we fit a bifactor model to unravel the common and specific variance of the measures (conceptual illustration in Figure S5). We used polyserial correlations and the DWLS estimator in estimation since we had a combination or ordinal and continuous item scores (Olsson et al., 1982). When assessing model fit, we focus primarily on the SRMR. Most data in our analyses are ordinal, and recent studies have suggested that the SRMR is more appropriate to use than fit statistics such as the RMSEA when analyzing ordinal observed variables (e.g., Shi et al., 2020). A value of the SRMR lower than 0.08 indicated a good model fit (Hu & Bentler, 1999). For the bifactor model, we assessed the dominance of the general factor by the explained common variance (Rodriguez et al., 2016).

Research Question 3

To address research question 3, we fitted three structural equation models (SEMs). The first was based on the five-factor model, with general vocabulary regressed on each of the factors. The second model was a mediation model based on the higher-order model. In line with the Morphological Pathways Framework, we specified direct paths from morphological awareness to morphological analysis, morphological decoding and vocabulary, as well as indirect paths from morphological awareness to vocabulary through analysis and decoding. The third model was based on the bifactor model where vocabulary was regressed on the general factor and each of the specific factors.

Results

Individual Test Models (Research Question 1)

Item-level descriptive statistics are listed in Tables S2–S7 in the Supplementary materials. Table 4 displays the fit statistics of the individual unidimensional models for each test. The chi-square tests of model fit were significant for all models except Productive Word Knowledge, but this was not unexpected given the large sample. The SRMR values were all < .08, indicating a good model fit and supporting the hypotheses of unidimensionality within the tests. Recall, however, that we specified the Productive Word Knowledge and spelling models with residual correlations

TABLE 4
Model Fit for Individual Models of Morphological Tests

	n	x^2 (robust)	*Df*	*p*	SRMR	RMSEA	CFI	TLI
Word analogy	580	89.669	64	.019	.055	.026	.990	.986
Receptive knowledge	554	363.697	299	.006	.058	.020	.984	.982
Productive knowledge	582	62.919	56	.245	.036	.015	.997	.996
Spelling	586	621.186	240	< .001	.055	.052	.956	.949
Word reading	583	17.451	2	< .001	.006	.115	.995	.986

among items containing the same affix. This indicates that there is some multidimensionality in the form of shared affix-specific variance within these models.

Models Across Tests (Research Question 2)

Fit statistics for the models that included all morphological tests are reported in Table 5. All models showed significant chi-square values. Again, this was not unexpected due to the large sample size. The unidimensional model provided the least good fit to the data, with SRMR = .083 exceeding the recommended cut-off value of .08. The less restricted models all provided a good fit (see Table 5). To compare the model fit further, we conducted chi-square difference tests for the one-factor, three-factor, five-factor, and higher-order models. The one-factor model is the only model that can be compared directly with the bifactor model using a chi-square difference test, as the other models are not nested in the bifactor model (e.g., Mansolf & Reise, 2017). Hence, we conducted a separate chi-square difference test for the one-factor and bifactor models. The hypothesis testing procedure showed that the five-factor model had a superior fit compared with the other models and that the bifactor model was preferred to the one-factor model. The results of all chi-square difference tests are reported in Table 5 (see the last three columns).

The five-factor model fit the data well (SRMR = .061, see Table 5 for further fit indices). The standardized factor loadings in the five-factor model ranged from .373 to .987. For a list of all standardized factor loadings for the five-factor model, see Table S8. Although the five-factor model pointed to a multidimensional construct, the factor correlations shown in Table 6 were quite high overall. This may indicate that the tests capture common variance across all the tests as well as the specific variance related to each separate test.

While the five-factor model fit significantly better in the model comparison, the higher-order model fit the data equally well in terms of SRMR (SRMR = .061, additional fit statistics in Table 5). The standardized factor loadings for the first-order factors ranged from .372 to .987 (see Table S9 for a complete list). For the second-order factors, the standardized loadings on analysis were .918 (productive) and .919 (receptive), and the loadings on decoding were .717 (word reading) and .968 (spelling). The

TABLE 5
Fit Statistics and Model Comparisons for Models across Tests

CFA	x^2 (robust)	Df	p(x^2)	SRMR	RMSEA	CFI	TLI	Δ x^2	Δdf	p (Δ x^2)
5-factor	3645.271	3037	< .001	.061	.018	.971	.970			
HO	3659.751	3040	< .001	.061	.019	.970	.969	11.468	3	< .01
3-factor	3836.307	3044	< .001	.063	.021	.962	.960	84.082	4	< .001
1-factor	5298.151	3047	< .001	.083	.035	.892	.888	295.470	3	< .001
Bifactor	3828.046	2967	< .001	.062	.022	.959	.956			
1-factor								1222.400	80	< .001
SEM										
5-factor	3916.832	3272	< .001	.060	.018	.970	.969			
HO	4017.891	3278	< .001	.061	.020	.966	.964			
Bifactor	4082.400	3201	< .001	.061	.022	.959	.957			

Note. Δ x2 is based on standard x2 values, not robust.

TABLE 6
Factor Correlations

	Receptive	Productive	Analogy	Spelling
Five-factor model				
Receptive	1			
Productive	.844	1		
Analogy	.646	.687	1	
Spelling	.629	.622	.658	1
Word reading	.457	.404	.542	.694
Higher-order model				
	Awareness	**Analysis**		
Awareness	1			
Analysis	.721	1		
Decoding	.693	.696		

Note. All correlations are significant at $p < .001$.

correlations among awareness, analysis, and decoding were medium to high, in line with the reciprocal relations suggested in the Morphological Pathways Framework (see Table 6).

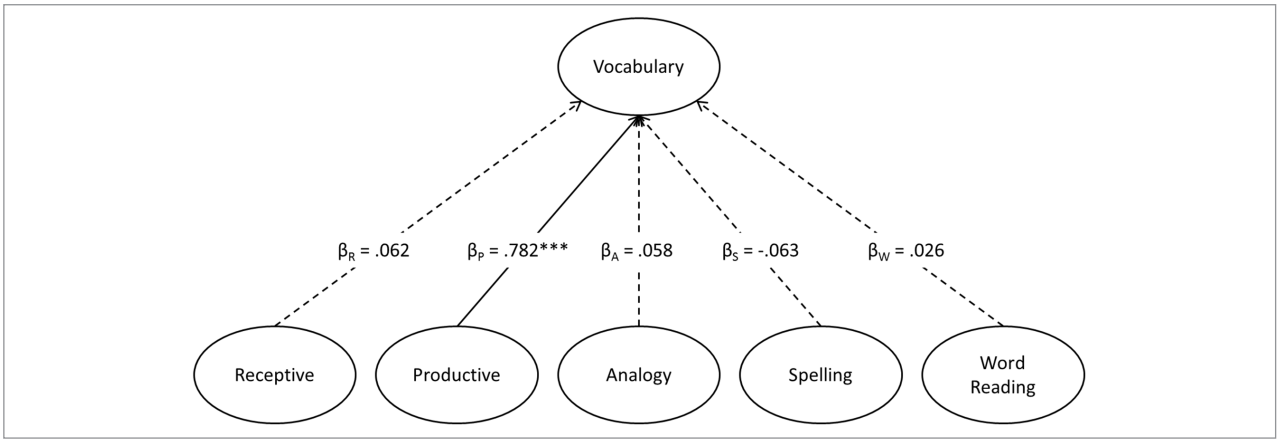
Within the framework of these models, however, it is not possible to examine the common and specific variances of the tests simultaneously. Hence, in accordance with Goodwin et al. (2017), we proceeded with a bifactor model to illuminate the construct of morphological knowledge further. The bifactor model fit the data well (SRMR = .062, see Table 5 for all fit indices). The chi-square difference test showed that the bifactor model fit significantly better than the one-factor model. All factor loadings on the general factor were significant, ranging between .254 and .689. For the specific factors, however, two indicators showed non-significant factor loadings. For a complete list of factor loadings for the bifactor model, see Table S10 in the supplementary material. Seventy-nine percent of the factor loadings on the general factor were ≥ .4, indicating a high degree of overlap between items from the different tests. This overlap was also spread out among the tests, so no test showed less overlap than others. The estimated explained common variance was .63. This value indicates that both the general and specific factors contribute to explaining the variance in the indicators (Rodriguez et al., 2016).

In sum, the results of our analyses clearly favored a multidimensional view of morphological knowledge. However, there was substantial ambiguity regarding how this multidimensionality should be represented. While the five-factor model provided the best fit among the nested models, the factors may be contaminated by substantial amounts of construct-irrelevant variance. The higher-order and bifactor models also provide an excellent fit and can help us separate the construct-irrelevant variance from variance related to morphological knowledge. To further disentangle dimensionality of morphological knowledge, we chose to proceed with all three models in the final part of our analyses.

Structural Equation Models (Research Question 3)

In the final part of our analyses, we expanded each model to a SEM. In these models, general vocabulary, as measured by the WISC-IV Vocabulary subtest, was regressed on each of the morphological factors. The goal of these analyses was not to investigate the relationship between morphological knowledge and vocabulary per se, but rather to demonstrate what kind of information the measurement models can provide. The five-factor SEM (Figure 1) fit the data well (SRMR = .060, see Table 5 for other

FIGURE 1
Five-Factor Structural Equation Model



Note. Indicators are left out for readability. ***$p < .001$

fit statistics). Inspecting the standardized regression coefficients, Productive Word Knowledge was the only factor with a significant relation to general vocabulary ($\beta_p = .782$, $p < .001$). Note that the predictors in this model were substantially correlated, which inflates the standard errors of the estimated regression coefficients. Thus, the model provides little information concerning the relations of morphological factors to vocabulary.

The higher-order SEM (Figure 2) fit the data well (SRMR = .061, see Table 5). In line with the Morphological Pathways Framework, morphological awareness directly affected both morphological analysis ($\beta_{a1} = .845$, $p < .001$) and morphological decoding ($\beta_{b1} = .799$, $p < .001$). Morphological analysis was also related to vocabulary ($\beta_{a2} = 1.002$, $p < .001$). There were no direct effects of morphological decoding or morphological awareness on vocabulary, nor any indirect effect of awareness through decoding. There was, however, a significant indirect effect of awareness through analysis ($\beta_{a1*a2} = .846$, $p < .001$). Since there was no direct effect of morphological awareness on vocabulary, the relation between them was fully mediated through morphological analysis. This provides additional support for the theoretical relations of the Morphological Pathways Framework.

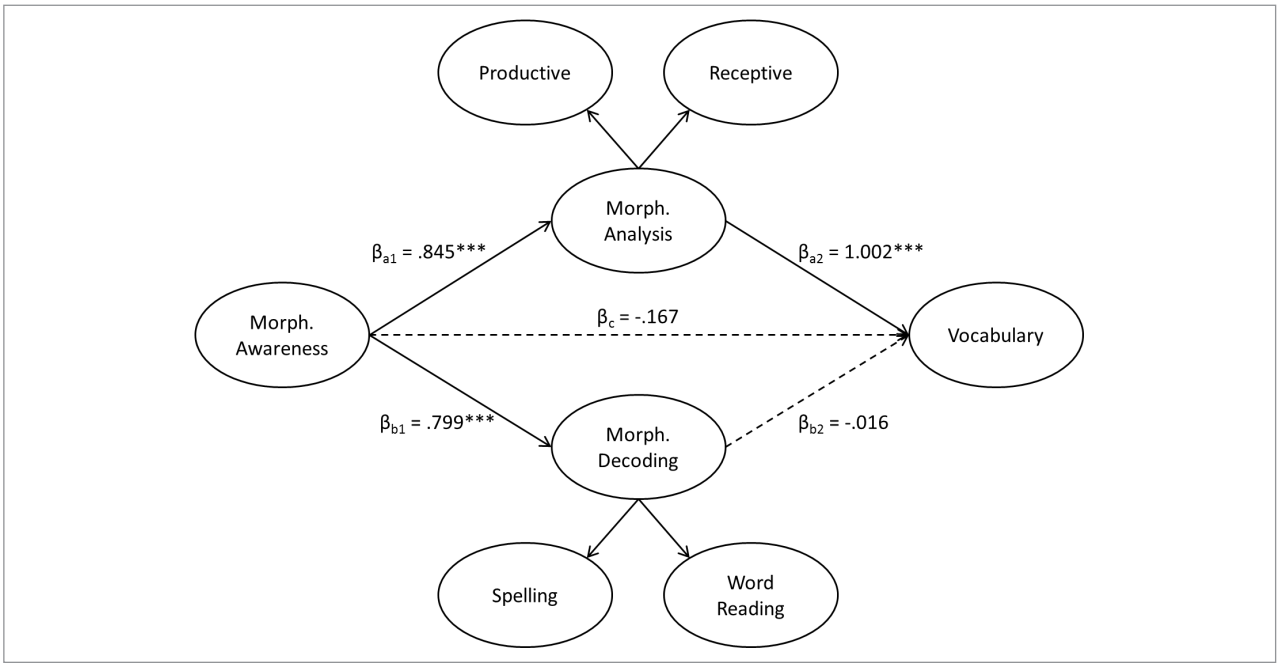
Finally, the bifactor SEM (Figure 3) also fit the data well (SRMR = .061, see Table 5). In this model, the general factor of morphological knowledge had the strongest relation to general vocabulary ($\beta_G = .664$, $p < .001$), followed by the specific productive factor ($\beta_{SP} = .527$, $p = .001$; see Figure 3). The specific receptive factor ($\beta_{SR} = .339$, $p = .003$) was also positively associated with general vocabulary, whereas the relation to the specific word reading factor ($\beta_{SW} = -.136$, $p = .010$) was negative. The specific factors of spelling ($\beta_{SS} = -.049$, $p = .280$) and analogy ($\beta_{SA} = .121$, $p = .077$) were not significantly related to general vocabulary. The results were similar to those found by Goodwin et al. (2017) with the exception of reading and spelling. In their study, reading was not significantly related to vocabulary, whereas spelling had a negative relation.

To sum up, the five-factor model, while empirically sound, provided little information about morphological knowledge and its relation to vocabulary. The higher-order model provided more information, particularly about the relations between the morphological constructs. It does not, however, allow us to investigate the specific variance of the first-order factors related to morphological decoding and analysis. Finally, the bifactor model provided information about construct-relevant and construct-irrelevant variance of morphological knowledge and allows us to investigate the specific variance within tests, as well as the common variance related to morphological knowledge.

Discussion

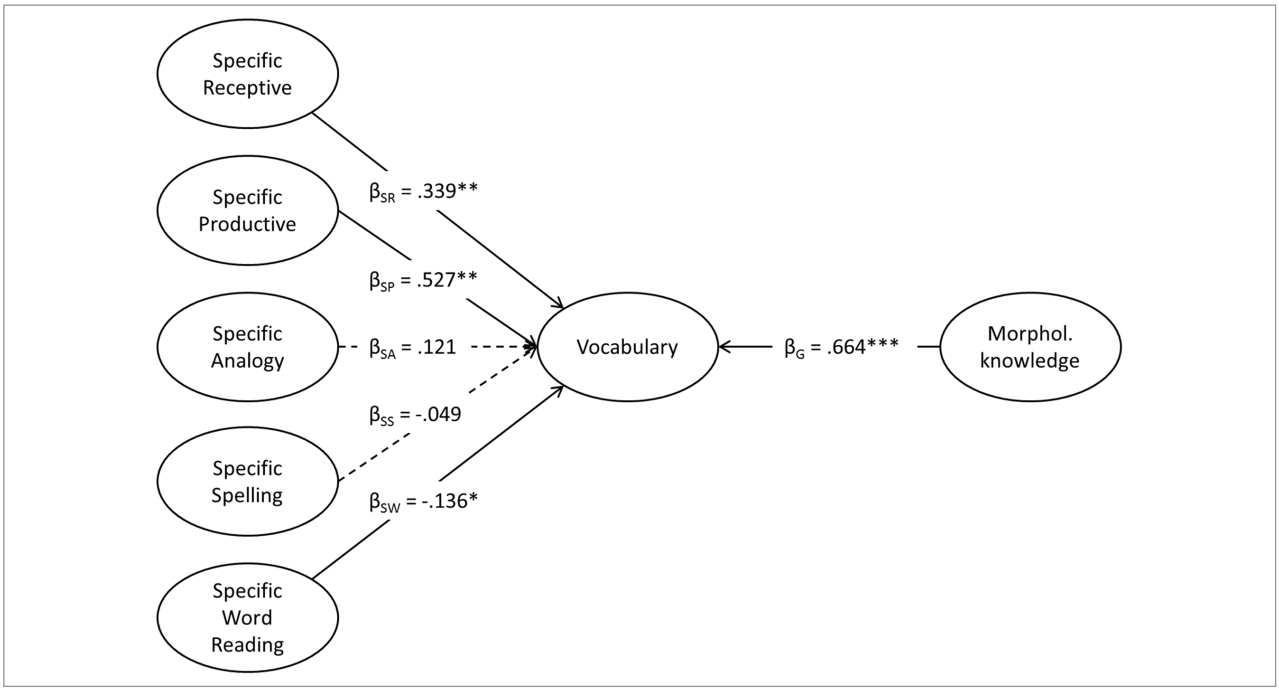
The present study evaluated the dimensionality of morphological knowledge in Norwegian third graders, both within and between tests that require different skills in addition to morphological knowledge. Moreover, the study investigated how different conceptualizations affect the

FIGURE 2
Higher-Order Structural Equation Model



Note. Indicators are left out for readability. ***$p < .001$

FIGURE 3
Bifactor Structural Equation Model



Note. Indicators are left out for readability. *$p < .05$, **$p < .01$, ***$p < .001$

relations between morphological dimensions and general vocabulary. We examined test-related dimensions, theoretical dimensions (morphological awareness, morphological analysis and morphological decoding), and general and specific dimensions relating to construct-relevant and construct-irrelevant variance. Our results show that each of the five tests measures a unidimensional construct. When analyzed together, the tests are best represented as multidimensional. The findings from the measurement models alone, however, are ambiguous as to whether a five-factor, higher-order, or bifactor model is most appropriate. When general vocabulary is regressed on each factor in the models, the different models imply different relationships between the morphological factors and general vocabulary. Below we discuss the findings related to each of our three research questions in turn.

RQ1: Are the Constructs Measured by the Separate Tests Unidimensional?

The individual test models support unidimensionality within tests. This means that the potential dimensions related to the lexical status, contextual cues and item types are not supported by our analyses within tests. Although previous studies investigated these dimensions across different tests, there was the possibility that subsets of items in our test could function as different subtests. The test of receptive word knowledge contains both tasks with context and tasks without context. In line with Tighe and Schatschneider (2015), we found no evidence for separate contextual dimensions in our sample, indicating similarities across age groups and languages in this regard. Neither did we find any evidence for separate dimensions relating to the three item types in the receptive test. In the Test of Productive Word Knowledge, we found no evidence of dimensionality relating to real words versus pseudowords. This is contrary to the findings of Tighe and Schatschneider (2015, 2016), but in line with Jong and Jung (2015), perhaps indicating that children are more inclined to accept pseudowords in line with real words than adults. The test of productive word knowledge and the spelling test did however require correlated residuals of items measuring the same affix, indicating some dimensionality related to specific affix knowledge. In the productive test, the children are asked to provide definitions of words. For example, if a child knows that *alveaktig* (elflike) means "similar to an elf", they would also likely be able to infer that the pseudoword *høneaktig* (henlike) could mean "similar to a hen". Hence, it is not surprising that the residuals are correlated for items containing the same affix. Similarly, in the spelling test, if a child knows that *endelig* (*final*) is spelled with a silent g at the end, they are probably more likely to remember the silent g in *fredelig* (peaceful).

It is perhaps not surprising that each test measures a unidimensional construct, with the exception of the correlated residuals, in the productive test and the spelling test. Although the items in our analyses vary in lexical status (words vs. nonwords) in the productive test, and

contextual cues (the presence vs. absence of a linguistic/image context) and item types in the receptive test, the same specific task demands are posed within each test. This might point to the task demands having a greater impact on dimensionality than item characteristics within a test. Nevertheless, this step was important to establish unidimensionality within tests and avoid potential confounding in further analyses.

RQ2: Is Morphological Knowledge Best Represented as a Unidimensional or Multidimensional Construct Across the Different Tests?

Considering the models incorporating all tests, the results of this study do not support a strictly unidimensional construct of performance on different tests of morphological knowledge. This indicates that using a single measure of morphological knowledge, whether in assessment or research, could impart an incomplete picture of children's morphological skills, at least in Norwegian. Moreover, morphological knowledge may be confounded with other skills such as decoding or general vocabulary, making claims of the effect of morphological knowledge uncertain. Hence, morphological knowledge should be measured across different tests that allow us to separate the common variance attributable to morphological knowledge from the specific variance due to other skill requirements inherent in the tests. This can provide a deeper understanding of the morphological knowledge and enhance comparisons across studies.

Our results indicate that a five-factor model fits the data very well, and significantly better than the three-factor and higher-order models, similar to the findings of Goodwin et al. (2017). This provides evidence of similarities in English and Norwegian, and across primary and middle school. Our finding that receptive and productive knowledge make up two of these factors is also in line with González-Sánchez et al. (2018) and Jong and Jung (2015), indicating similarities with Spanish children in preschool as well as with Korean fifth and sixth graders. This separation of receptive and productive knowledge may, however, represent construct-irrelevant variance due to differences between general comprehension and language production, rather than separate dimensions of morphological knowledge. A critical drawback of the five-factor model is that it does not separate construct-relevant and irrelevant variance. Thus, multidimensionality could be a consequence of tests measuring other skills in addition to morphological knowledge. While more research is needed to strengthen any conclusions, one potential source of multidimensionality is the methodological artifacts inherent in the set of tests. This could explain some of the differences found across studies thus far, since different tasks may pose different demands of both morphological knowledge and other linguistic skills. Furthermore, these demands may vary across age groups and languages, potentially explaining why some studies find evidence of unidimensionality and others of multidimensionality. Another drawback is that a test-specific conceptualization of morphological knowledge implies that every test measures a separate morphological dimension, thus disabling comparisons of results from studies using different measures of morphological knowledge. These drawbacks make a correlated traits model like the five-factor model an ill-advised choice for research.

Although the five-factor model provides a closer fit than the higher-order model in terms of chi-square difference, the latter also fits the data very well with an equal value of SRMR. This model is theoretically founded in the Morphological Pathways Framework (Levesque et al., 2021), and provides similar results to those found by Levesque et al. (2017), Zhang (2017), and in parts by Tibi and Kirby (2017). This provides further evidence of similarities rather than discrepancies across languages and age groups. A key benefit of using a higher-order model rather than a three-factor (correlated factors) model to represent the morphological dimensions is that it allows us to separate out the construct-relevant variance in the second-order factors. In this respect, the addition of a second measure of morphological awareness would have strengthened our model. One drawback of the model is that we cannot investigate the construct-irrelevant variance directly to assess additional sources of variance within tests. It does, however, enable us to examine the relations between morphological dimensions according to the Morphological Pathways Framework.

In the bifactor model, 79% of the factor loadings on the general factor were of a magnitude indicating overlap in the variance of items across the tests. Furthermore, most indicators have significant positive loadings on their respective specific factors (see Table S10). Along with an estimated explained common variance of .63, this means that the tests measure unique skillsets in addition to the common factor, and the bifactor model allows us to separate the common and specific variance of the tests. Similarly to the higher-order model, the extraction of construct-relevant variance is a crucial point if we wish to compare findings across studies, as the interpretation of relationships between morphological knowledge and other skills in language and literacy tasks depends on what causes the variance in morphological knowledge. Although we tested a model with a single general factor in our study, a bifactor model could also incorporate multiple general factors to further account for multidimensionality (e.g., Goodwin et al., 2021). The assumption that factors are uncorrelated is a drawback of the bifactor model, however. For example, the Morphological Pathways Framework cannot be tested in a bifactor model, since it does not allow for relations between morphological dimensions.

RQ3: How do Different Models Affect the Relation Between Morphological Knowledge and General Vocabulary?

The results from the three structural models imply very different relations between morphological knowledge and general vocabulary. The five-factor SEM in our study suggests that productive word knowledge is the only dimension of morphological knowledge that is related to general vocabulary, as measured by a word definition test. The lack of any relationship among general vocabulary and the other morphological factors might be due to the factor correlations disguising the unique contributions of each factor. This makes the five-factor model less informative, and further strengthens our claim that a correlated factor model is not suited for research on morphological knowledge and the relations of its facets to other areas of language and literacy.

The higher-order SEM is more informative, as we are able to examine relations among the morphological factors, as well as their relations to vocabulary. Supporting the theoretical pathways posited by Levesque et al. (2021), we found that morphological analysis was strongly related to vocabulary, whereas morphological decoding had no direct relation. The relation between morphological awareness and vocabulary was fully mediated through analysis, which is also in line with the theory. Thus, we found evidence of specific mechanisms within dimensions of morphological knowledge that influence the relations to other linguistic skills, exemplified with vocabulary in our study.

The results from our bifactor SEM analysis closely resemble those found by Goodwin et al. (2017). While we cannot draw any firm conclusions, the similarity in relations between the morphological factors and vocabulary provides support for the interpretation that these are generalizable patterns that apply to different languages and age groups. A model with three general factors, in line with Levesque et al. (2021), or four, as in Goodwin et al. (2021), might have been even more informative, but this was beyond the scope of our study. Even if such a model had been possible, the assumption of uncorrelated factors would prohibit an investigation of potential relations between the general factors. Still, the bifactor model provides the opportunity to examine the relations of specific factors to other linguistic skills. This is of importance when developing assessments, as it provides information on which skills we are measuring in addition to morphological knowledge.

In sum, our results show that morphological knowledge in Norwegian third graders is a multidimensional construct and that we need to account for construct-irrelevant variance due to methodological artifacts to get a clear representation of the construct. The five-factor model cannot separate construct-relevant and construct-irrelevant variance. Hence, it does not provide a clear view of whether the separate factors are due to different dimensions of morphological knowledge, or due to methodological artifacts such as tests measuring other language skills in addition to morphological knowledge. The bifactor model is well suited to separate construct-relevant and construct-irrelevant variance and accounts for multidimensionality as a methodological artifact. Thus, it provides an excellent framework for examining the overarching construct of morphological knowledge. A substantial drawback is that all factors in a bifactor model are uncorrelated, so in a theoretical model with three general factors representing morphological awareness, morphological analysis and morphological decoding, we would have to assume that these dimensions are unrelated. This assumption does not align with theory. Our results support the theoretical structure proposed by Levesque et al. (2021). To represent this structure, a higher-order model provides the best alternative, allowing us to remove construct-irrelevant variance while still enabling relations among the different factors.

Implications for Assessment and Research

It is clear from the findings of the present study that the associations between morphological knowledge and other language and literacy skills depend on how morphological knowledge is conceptualized. There is no doubt about the major differences in interpretation when comparing the five-factor, higher-order and bifactor SEMs in the current study. This implies that the use of different measures and different models may lead to confusion or misinterpretation if we are not careful in how we interpret results. Furthermore, the bifactor model might remove some of the confounding factors by separating the construct-relevant variance from that which is irrelevant. If we aim to investigate general morphological knowledge, it would be favorable to remove the variance related to other constructs, whether these represent specific morphological skills or other linguistic or task-related abilities. On the other hand, if our aim is to examine the relations among different morphological skills, we should turn to a higher-order model to enable relations between these factors. The bifactor model might be especially informative in test development. To further our understanding of morphological knowledge, however, we recommend representing the three dimensions of morphological awareness, morphological analysis and morphological decoding, in line with Levesque et al. (2021).

Our results indicate that in Norwegian, at least, morphological knowledge can be differentiated into morphological awareness, morphological analysis and morphological decoding from a relatively early age. Morphological decoding does require that the children have mastered basic word

reading and spelling skills. Given that this assumption is met, we recommend measuring all three constructs to get a complete picture of children's morphological knowledge. To separate potential confounding information, we should use a model that separates construct-relevant variance from variance attributable to sources other than morphological knowledge, for example, a higher-order model.

Regarding the construction of interventions, we should take into account that morphological awareness, morphological analysis and morphological decoding might require different supporting skills, such as general vocabulary (base word knowledge) and decoding or spelling skills. According to the Morphological Pathways Framework, growth in morphological awareness will impact both morphological analysis and morphological decoding. Furthermore, improving morphological analysis will increase word knowledge and comprehension, whereas morphological decoding can enhance word reading and spelling. Thus, morphological interventions can aim to enhance language development broadly, or be tailored to affect specific skills, for example, reading or spelling.

Limitations and Future Research

To help us understand the source of variance in different dimensions of morphological knowledge and shed further light on the interpretation of factors, future research should aim to investigate the relationship between morphological factors and a wide variety of linguistic skills, such as reading comprehension, reading fluency, spelling, and listening comprehension. One particular limitation of the current study is the lack of a reading comprehension measure. Including a measure of reading comprehension would have provided additional context for factor interpretation, especially in the case of the specific word reading and spelling factors of the bifactor model, as well as the morphological decoding factor of the higher-order model. Additionally, general vocabulary was measured only with the vocabulary subtest of WISC-IV, a word definition test. A broader construct of vocabulary, for example, including a test of receptive vocabulary, would have been preferable.

Another limitation of the current study relates to the extensive exclusion of items, particularly from the test of receptive word knowledge. While the item exclusion did not change the substantive or statistical interpretations of the constructs measured, the analyses should be replicated in an independent sample to examine the generalizability of our models and results. This would also help refine the measures we developed in the project for use in future studies. Reducing the number of items will decrease the effort required from the children, as well as the time needed for testing, provided that validity holds for the intended use of the test scores.

The study supports the conclusion of Goodwin et al. (2017) that the bifactor model can help separate between construct-relevant and construct-irrelevant variance. Since the bifactor model also represents task-specific variance explicitly, it can contribute information about what we are measuring in addition to morphological knowledge. Future research should investigate how such specific factors related to general measures of skills such as word reading, spelling, and reading comprehension, as this could be informative for test development. Our results also support the Morphological Pathways Framework of Levesque et al. (2021). This provides preliminary evidence that the skills underlying the three theoretical constructs of this framework emerge relatively early in Norwegian, and perhaps in other alphabetic languages such as English. To strengthen the generalizability of the findings, future research should investigate whether the framework can be extended to similar languages as well as languages with different writing systems or distributions of morphemes (derivations, compounds and inflections), such as Chinese or Hebrew. Future research should also include children in preschool and early primary school to shed further light on the age of onset for the different morphological skills.

Acknowledgments

The authors thank the students, teachers, school leaders, and municipalities who participated. They thank Engagelab, especially Richard Nesnass, Ole Smørdal, Siri Jønnum and Jesús Gonzáles Torres, for their contributions to app development. They also thank Bente E. Hagtvet, Sol Lyster, Anita Lopez-Pedersen, Kari-Anne B. Næss, Riikka-Maija Mononen, Ona Bø Wie, Charles Hulme, and Catherine Snow for their valuable contributions to the design and implementation of the RCT. They are grateful to Terje Ulv Throndsen, Stein Malmo, Germán García Grande, and Nina Melsom Kristensen for research assistance, and master's students at the Department of Special Needs Education, University of Oslo, for their help with data collection and scoring. This work was supported by the Research Council of Norway, Grant no. 24033.

Conflicts of Interest

The authors have no known conflict of interest to disclose.

REFERENCES

Berthiaume, R., Bourcier, A., & Daigle, D. (2018). Morphological processing tasks and measurement issues. In R. Berthiaume, D. Daigle, & A. Desrochers (Eds.), *Morphological processing and literacy development* (pp. 48–87). Routledge.

Bowers, P. N., Kirby, J. R., & Deacon, S. H. (2010). The effects of morphological instruction on literacy skills: A systematic review of the

literature. *Review of Educational Research*, *80*(2), 144–179. https://doi.org/10.3102/0034654309359353
Brinchmann, E. I., Hjetland, H. N., & Lyster, S. A. H. (2016). Lexical quality matters: Effects of word knowledge instruction on the language and literacy skills of third-and fourth-grade poor readers. *Reading Research Quarterly*, *51*(2), 165–180. https://doi.org/10.1002/rrq.128
Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.
Bryant, P., Nunes, T., & Bindman, M. (1997). Backward readers' awareness of language: Strengths and weaknesses. *European Journal of Psychology of Education*, *12*(4), 357–372. https://doi.org/10.1007/BF03172798
Carlisle, J. F. (2010). Effects of instruction in morphological awareness on literacy achievement: An integrative review. *Reading Research Quarterly*, *45*(4), 464–487. https://doi.org/10.1598/RRQ.45.4.5
Deacon, S., Parrila, R., & Kirby, J. (2008). A review of the evidence on morphological processing in dyslexics and poor readers: A strength or weakness? In G. Reid, A. J. Fawcett, & F. Manis (Eds.), *The SAGE handbook of dyslexia* (pp. 212–238). SAGE Publications Ltd.
Gonnerman, L. M. (2018). A linguistic analysis of word morphology. In R. Berthiaume, D. Daigle, & A. Desrochers (Eds.), *Morphological processing and literacy development* (pp. 3–15). Routledge.
González-Sánchez, L., García, T., Areces, D., Fernández, E., Arias-Gundin, O., & Rodríguez, C. (2018). Validación del Instrumento de Evaluación de la Conciencia Morfológica Oral (IECMO). *European Journal of Education and Psychology*, *11*(2), 107–122. https://doi.org/10.30552/ejep.v11i2.225
Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia*, *60*, 183–208. https://doi.org/10.1007/s11881-010-0041-x
Goodwin, A. P., Petscher, Y., Carlisle, J. F., & Mitchell, A. M. (2017). Exploring the dimensionality of morphological knowledge for adolescent readers. *Journal of Research in Reading*, *40*(1), 91–117. https://doi.org/10.1111/1467-9817.12064
Goodwin, A. P., Petscher, Y., & Tock, J. (2021). Multidimensional morphological assessment for middle school students. *Journal of Research in Reading*, *44*(1), 70–89. https://doi.org/10.1111/1467-9817.12335
Grande, G. G. (2018). *Morphological awareness in Norwegian preschoolers. [Master's thesis, University of Oslo]*. http://urn.nb.no/URN:NBN:no-66745
Hagtvet, B. E., Helland, T., & Lyster, S. A. H. (2006). Literacy acquisition in Norwegian. In R. M. Joshi & P. G. Aaron (Eds.), *Handbook of orthography and literacy* (pp. 15–30). Routledge.
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118
James, E., Currie, N. K., Tong, S. X., & Cain, K. (2021). The relations between morphological awareness and reading comprehension in beginner readers to young adolescents. *Journal of Research in Reading*, *44*(1), 110–130. https://doi.org/10.1111/1467-9817.12316
Jong, Y. O., & Jung, C. K. (2015). Pedagogical significance of morphological awareness in Korean and English. *English Language Teaching*, *8*(8), 79–93. https://doi.org/10.5539/elt.v8n8p79
Kuo, L. J., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist*, *41*(3), 161–180. https://doi.org/10.1207/s15326985ep4103_3
Levesque, K. C., Breadmore, H. L., & Deacon, S. H. (2021). How morphology impacts reading and spelling: Advancing the role of morphology in models of literacy development. *Journal of Research in Reading*, *44*(1), 10–26. https://doi.org/10.1111/1467-9817.12313
Levesque, K. C., Kieffer, M. J., & Deacon, S. H. (2017). Morphological awareness and reading comprehension: Examining mediating factors. *Journal of Experimental Child Psychology*, *160*, 1–20. https://doi.org/10.1016/j.jecp.2017.02.015
Levesque, K. C., Kieffer, M. J., & Deacon, S. H. (2019). Inferring meaning from meaningful parts: The contributions of morphological skills to the development of children's reading comprehension. *Reading Research Quarterly*, *54*(1), 63–80. https://doi.org/10.1002/rrq.219
Lyster, S. A. H. (2002). The effects of morphological versus phonological awareness training in kindergarten on Reading development. *Reading & Writing*, *15*(3–4), 261–294. https://doi.org/10.1023/A:1015272516220
Lyster, S. A. H., Lervåg, A. O., & Hulme, C. (2016). Preschool morphological training produces long-term improvements in reading comprehension. *Reading and Writing*, *29*(6), 1269–1288. https://doi.org/10.1007/s11145-016-9636-x
Manolitsis, G., Georgiou, G. K., Inoue, T., & Parrila, R. (2019). Are morphological awareness and literacy skills reciprocally related? Evidence from a cross-linguistic study. *Journal of Educational Psychology*, *111*(8), 1362–1381. https://doi.org/10.1037/edu0000354
Mansolf, M., & Reise, S. P. (2017). When and why the second-order and bifactor models are distinguishable. *Intelligence*, *61*, 120–129. https://doi.org/10.1016/j.intell.2017.01.012
McBride-Chang, C., Cho, J. R., Liu, H., Wagner, R. K., Shu, H., Zhou, A., Cheuk, C. S., & Muse, A. (2005). Changing models across cultures: Associations of phonological awareness and morphological structure awareness with vocabulary and word recognition in second graders from Beijing, Hong Kong, Korea, and the United States. *Journal of Experimental Child Psychology*, *92*(2), 140–160. https://doi.org/10.1016/j.jecp.2005.03.009
Muse, A. E. (2005). *The nature of morphological knowledge. [Doctoral dissertation, Florida State University]*. DigiNole: FSU's Digital Repository https://diginole.lib.fsu.edu/islandora/object/fsu%3A180396
Nagy, W., Berninger, V. W., & Abbott, R. D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of Educational Psychology*, *98*(1), 134–147. https://doi.org/10.1037/0022-0663.98.1.134
Nagy, W. E., Carlisle, J. F., & Goodwin, A. P. (2014). Morphological knowledge and literacy acquisition. *Journal of Learning Disabilities*, *47*(1), 3–12. 10.1177%2F0022219413509967
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.
Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, *47*, 337–347. https://doi.org/10.1007/BF02294164
R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/
Ragnarsdóttir, H., Simonsen, H. G., & Plunkett, K. (1999). The acquisition of past tense morphology in Icelandic and Norwegian children: An experimental study. *Journal of Child Language*, *26*(3), 577–618. https://doi.org/10.1017/S0305000999003918
Reed, D. K. (2008). A synthesis of morphology interventions and effects on Reading outcomes for students in grades K–12. *Learning Disabilities Research & Practice*, *23*(1), 36–49. https://doi.org/10.1111/j.1540-5826.2007.00261.x
Revelle, W. (2021). *Psych: Procedures for personality and psychological research*. Northwestern University, Evanston, Illinois, USA https://cran.r-project.org/package=psych
Ribu, I. S., Simonsen, H. G., Løver, M. A., Strand, B.-M. S., & Kristoffersen, K. E. (2019). N-LARSP: A developmental language profile for Norwegian. In M. J. Ball, P. Fletcher, & D. Crystal (Eds.), *Grammatical profiles: Further languages of LARSP* (pp. 1–48). Multilingual Matters.
Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*, 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02
Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*, 143–174. https://doi.org/10.1348/000712603321661859
Shi, D., Maydeu-Olivares, A., & Rosseel, Y. (2020). Assessing fit in ordinal factor analysis models: SRMR vs. RMSEA. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 1–15. https://doi.org/10.1080/10705511.2019.1611434
Spencer, M., Muse, A., Wagner, R. K., Foorman, B., Petscher, Y., Schatschneider, C., Tighe, E. L., & Bishop, M. D. (2015). Examining the underlying dimensions of morphological awareness and vocabulary knowledge. *Reading and Writing.*, *28*, 959–988. https://doi.org/10.1007/s11145-015-9557-0
Tibi, S. (2016). *Cognitive and linguistic factors of reading Arabic: The role of morphological awareness in reading. [Doctoral dissertation, Queens University Kingston]*. QSpace: Queen's Scholarship & Digital Collections https://qspace.library.queensu.ca/bitstream/handle/1974/14674/Tibi_Sana_T_201607_PhD.pdf.pdf?sequence=1
Tibi, S., & Kirby, J. R. (2017). Morphological awareness: Construct and predictive validity in Arabic. *Applied PsychoLinguistics*, *38*, 1019–1043. https://doi.org/10.1017/S0142716417000029
Tighe, E. L., & Schatschneider, C. (2015). Exploring the dimensionality of morphological awareness and its relations to vocabulary knowledge in adult basic education students. *Reading Research Quarterly*, *50*(3), 293–311. https://doi.org/10.1002/rrq.102
Tighe, E. L., & Schatschneider, C. (2016). Modeling the relations among morphological awareness dimensions, vocabulary knowledge, and Reading comprehension in adult basic education students. *Frontiers in Psychology*, *86*(7), 395–409. https://doi.org/10.3389/fpsyg.2016.00086
Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research*, *49*, 1193–1208. https://doi.org/10.1044/1092-4388(2006/086)
Torkildsen, J. V. K., Bratlie, S. S., Kristensen, J. K., Gustafsson, J.-E., Lyster, S.-A. H., Snow, C., Hulme, C., Mononen, R.-M., Næss, K.-A. B., López-Pedersen, A., Wie, O. B., & Hagtvet, B. (2022). App-based morphological training produces lasting effects on word knowledge in primary school children: A randomized controlled trial. *Journal of Educational Psychology*, *114*(4), 833–854. https://doi.org/10.1037/edu0000688
Verhoeven, L., & Perfetti, C. A. (2011). Morphological processing in reading acquisition: A cross-linguistic perspective. *Applied Psycho-Linguistics*, *32*(3), 457–466. https://doi.org/10.1017/S0142716411000154
Wechsler, D. (2009). *WISC-IV norsk versjon. Manual del 1*. NCS Pearson, Inc.
Zhang, D. (2017). Multidimensionality of morphological awareness and text comprehension among young Chinese readers in a multilingual context. *Learning and Individual Differences*, *56*, 13–23. https://doi.org/10.1016/j.lindif.2017.04.009

Submitted January 19, 2022
Final revision received March 1, 2023
Accepted March 5, 2023

JARL K. KRISTENSEN (corresponding author) is a Doctoral Research Fellow at the Centre for Educational Measurement, University of Oslo, Oslo, Norway; email: jarlkkristensen@gmail.com

BJÖRN ANDERSSON is an Associate Professor at the Centre for Educational Measurement, University of Oslo, Oslo, Norway; email: bjorn.andersson@cemo.uio.no

SIRI S. BRATLIE is a Postdoctoral Research Fellow at the Department of Education, University of Oslo, Oslo, Norway; email: s.s.bratlie@iped.uio.no

JANNE V. K. TORKILDSEN is a Professor at the Department of Special Needs Education, University of Oslo, Oslo, Norway; email: janneto@isp.uio.no

Supporting Information

Additional supporting information may be found in the online version of this article on the publisher's website:10.1002/rrq.497/suppinfo

Table S1. Literature Review.
Table S2. Descriptive Statistics—Test of Receptive Word Knowledge.
Table S3. Descriptive Statistics—Test of Productive Word Knowledge.
Table S4. Descriptive Statistics—Word Analogy Test.
Table S5. Descriptive Statistics—Spelling Test.
Table S6. Descriptive Statistics—Word Reading Efficiency Test.
Table S7. Descriptive Statistics—WISC-IV Vocabulary.
Table S8. Factor Loadings, Five-Factor Model.
Table S9. Factor Loadings, Higher Order Model.
Table S10. Factor Loadings, Bifactor Model.
Figure S1. One-Factor Model.
Figure S2. Three-Factor Model.
Figure S3. Five-Factor Model.
Figure S4. Higher Order Model.
Figure S5. Bifactor Model.

Article 3

Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2024). Repeated Mistakes in App-Based Language Learning: Persistence and Relation to Learning Gains. *Computers & Education*. https://doi.org/10.1016/j.compedu.2023.104966

Computers & Education 210 (2024) 104966

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Repeated mistakes in app-based language learning: Persistence and relation to learning gains

Jarl K. Kristensen [a,*], Janne v. K. Torkildsen [b], Björn Andersson [a]

[a] Centre for Educational Measurement, University of Oslo, Oslo, Norway
[b] Department of Special Needs Education, University of Oslo, Oslo, Norway

A B S T R A C T

Over the past decade, there has been an enormous upsurge in the use of educational apps in primary schools. However, few studies have examined how children interact with these apps and how their interaction patterns relate to learning outcomes. An interaction pattern that is potentially detrimental to learning is repeated mistakes, defined as making the same mistake more than once when answering a task. With interaction data from an eight-week digital vocabulary intervention, we examined 1) whether the propensity to make repeated mistakes changes across app sessions, and 2) how repeated mistakes relate to children's prior knowledge and their learning gains from the intervention. Our sample consisted of 363 Norwegian second graders who worked with the vocabulary app in a randomized controlled trial. Using growth curve modeling and confirmatory factor analyses, we found that the propensity to repeat mistakes remained stable over time. Furthermore, a structural equation model showed that repeated mistakes related negatively to both pre-test and post-test scores. A substantial proportion of the total effect of prior knowledge on learning gains was mediated by the propensity to repeat mistakes. Children who made more repeated mistakes had lower expected learning gains across all levels of prior knowledge. We suggest that the propensity to repeat mistakes may pose a double threat to learning by diminishing exposure to relevant content, and amplifying the exposure to incorrect input. Considering the stability of mistake repetition, it is crucial to identify students with a high propensity to repeat mistakes and help them break the pattern to support learning. App developers can help this process by implementing automatic detection and feedback.

1. Introduction

Recent years have seen an avalanche of educational apps designed for primary school children, coupled with a sharp increase in use (e.g, Montazami et al., 2022). There is a critical need to examine how schoolchildren interact with these apps, and how their interaction patterns relate to their learning outcomes. In educational apps, children's engagement with task content is critical to promote learning. When children disengage from the content, they suspend the learning process. One pattern of disengagement shown to affect learning negatively is gaming the system (Baker et al., 2004). This includes both guessing and hint abuse, behaviors that aim to complete tasks without engaging with the content. While hint abuse is more system-dependent, as it requires a help function that allows progression without solving tasks, guessing is more independent of individual system features.

Rapid guessing, i.e. providing a response in less time than it would take to read and understand a task, is frequently studied in the context of assessments, where it poses a threat to the validity of test results by introducing construct-irrelevant variance to the test scores (e.g. Wise, 2017). In assessment contexts, researchers typically use response time and accuracy to identify rapid guessing. This approach is straightforward in traditional multiple-choice settings since only a single response is required. In educational apps, however, the number and types of responses needed vary depending on the content and format of the tasks. Furthermore, guessing might represent an appropriate solution strategy when tasks provide feedback on the correctness of responses, while giving little

* Corresponding author.
E-mail address: jarlkk@uio.no (J.K. Kristensen).

https://doi.org/10.1016/j.compedu.2023.104966
Received 13 July 2023; Received in revised form 10 November 2023; Accepted 20 November 2023
Available online 29 November 2023

explicit instruction.

To learn from their guesses, however, children need to pay attention to the responses they choose and the feedback from the app. When children fail to attend to their responses and the feedback they receive, they are more likely to repeat mistakes. We define repeated mistakes as any erroneous answer given more than once within a task. Mistake repetition can potentially affect learning negatively in at least two ways. First, it can signal a lack of attention to task content that means that children distance themselves from the relevant input from the app. Second, repeated mistakes increase the exposure to incorrect input, potentially causing children to learn the wrong thing (e.g., Plante & Gómez, 2018).

The present study examines repeated mistakes in the context of an eight-week app-based intervention designed to promote implicit learning of morphological knowledge (Torkildsen et al., 2022). First, we examine whether the propensity to make repeated mistakes changes across sessions in the app. Do some children, for example, make more repeated mistakes in later sessions than in earlier ones? Considering the potential negative effects of repeated mistakes, it is important to know whether repeating mistakes is something children do intermittently or whether the propensity to repeat mistakes remains stable over longer periods of time. Furthermore, whether the propensity to repeat mistakes changes has implications for how we can assess its relations to other characteristics. If there are specific patterns of change, these must be accounted for in analyses. Second, we investigate how repeated mistakes relate to children's prior knowledge of morphology and their learning outcomes, i.e. their improvement in morphological knowledge from pre-test to post-test. To our knowledge, this is the first study to address whether children's propensity to make repeated mistakes in app-based language learning changes over time, and how it relates to learning outcomes.

1.1. Educational apps for language learning

Vocabulary is an important target for educational apps, as vocabulary knowledge is key to reading comprehension and educational success in all school subjects (Ash & Baumann, 2017; Milton & Treffers-Daller, 2013). In line with this, the majority of educational apps for language learning focus on vocabulary (Dehghanzadeh et al., 2021; Heil et al., 2016). However, vocabulary is difficult to teach due to its vast problem space. Specifically, school texts may contain close to a hundred thousand different words, many with complex meanings (Nagy & Anderson, 1984). Thus, vocabulary is often considered an unconstrained skill in the sense that interventions can only cover small parts of the content space (Paris, 2005; Snow & Matthews, 2016). There is an acute need for teaching approaches that promote generalization to untaught words, but this has proven difficult to obtain with traditional vocabulary instruction (Cervetti et al., 2023).

Considering these issues relating to vocabulary interventions, it is problematic that many apps focus on vocabulary in isolation (Heil et al., 2016). However, there is an increasing focus on teaching words in various contexts, through different modalities such as listening, reading, writing and speech. A well known example is the Duolingo language app, where tasks range from recognizing isolated words to highly contextualized dialogues, and responses are multimodal, e.g. selecting among response options, writing or speaking (Freeman et al., 2023).

1.2. The role of feedback in educational apps

Feedback comes in many forms: positive feedback relating to correct answers and negative feedback in response to incorrect attempts. It also varies in specificity and complexity (e.g., Nikolayev et al., 2021). Simple feedback includes verification and correction, while complex feedback involves elaboration and scaffolding (Nicolayev et al., 2021; Tärning, 2018).

Verification is a non-specific form of feedback that simply shows whether an answer is correct (positive verification) or incorrect (negative verification), while correction is a specific form of negative feedback where the indication of incorrectness is supplemented by the provision of the correct one. Positive verification can also be supplemented by textual or verbal provision of the correct answer, in which case it provides specific feedback (Callaghan & Reich, 2018; Nikolayev et al., 2021). In their review, Nikolayev et al. (2021) found that 85% of the included apps provided positive, non-specific feedback, i.e. positive verification. Positive specific feedback, highlighting the correct answer, was only included in 13% of the apps. Negative feedback showed similar trends with 49% including negative verification and only 13% including correction (specific negative feedback).

According to Tärning (2018), verification feedback allows for trial-and-error strategies that can increase the propensity to game the system, whereas corrective feedback does not allow for trial and error, hence eliminating gaming behavior. However, simply giving the correct answer after an incorrect answer could just as easily lead children to select a random answer, knowing they will proceed in the task anyway, which also constitutes a form of gaming the system. However, as noted by Tärning (2018), the effect of feedback depends on the app design. Specifically, verification can be separated into low-cost, risky, and time-consuming trial-and-error. Low-cost trial-and-error represents an "easy way out" and could promote gaming the system, whereas risky and time-consuming trial-and-error incurs costs, e.g. in terms of points lost or inordinate amounts of time consumed. Thus, while low-cost trial-and-error can increase the amount of gaming the system, risky and time-consuming trial-and-error is more likely to foster beneficial solution behaviors.

Related to feedback is the concept of rewards. Previous research has found that rewards designed to promote extrinsic motivation, such as badges or score boards, can have a negative impact on intrinsic motivation (Deci et al., 2001; Glover, 2013). Deci et al. (2001) argue that educational apps should foster intrinsic motivation, rather than focus on rewards for extrinsic motivation.

1.3. Morphological pathways to word knowledge

While an isolated focus on specific words is unlikely to lead to generalizable knowledge that will transfer to new words,

morphological instruction is a promising approach. Morphology is a constrained area of language that can serve as a gateway to unconstrained areas such as vocabulary and reading comprehension (Bratlie et al., 2022; Torkildsen et al., 2022). Morphemes, such as co-in cooperate and -ist in guitarist, are the smallest meaning-bearing units of language. Since they occur in numerous combinations, they provide generalizable knowledge that transfers to new contexts, e.g., *un*tidy means *not* tidy, so *un*fair must mean *not* fair.

Research suggests that morphology affects word learning through three dimensions: morphological awareness, morphological analysis, and morphological decoding (Levesque et al., 2021). Morphological awareness is the ability to consciously reflect on and manipulate morphemes. Morphological analysis involves knowledge of morpheme meanings, whereas morphological decoding is knowledge about the written forms of morphemes. While this theory is largely based on studies of the English language, there is evidence of this structure in other languages, e.g., Norwegian (Kristensen et al., 2023). Levesque et al. (2021) suggest that the three dimensions of morphological knowledge are reciprocally related. Thus, training one dimension can support development in the other two. Furthermore, Torkildsen et al. (2022) found evidence that training mainly receptive skills (word reading and listening comprehension) provided positive effects on expressive skills (word explanations and spelling). While morphological training can contribute to generalized word knowledge, there is a lack of research on educational apps targeting morphology.

1.4. Implicit learning and educational language apps

A challenge in teaching language, and perhaps especially morphology, is that explicit instruction requires an elevated level of metalinguistic competence from the learners; competence that may be beyond reach for children in early primary school. Some morphemes are easy to explain, such as *un*-in *un*happy, which reverses the meaning of the base word. Other affixes are more difficult to explain explicitly. For example, in Norwegian, the affix *-ende* (*-ing*) in "fly*ende*" (fly*ing*) changes the word class from verb to adjective. Explicit teaching of such content is likely to be too difficult for younger primary school children who lack the prerequisite metalinguistic skills, e.g. explicit knowledge of word classes. Implicit learning offers a different approach where children acquire knowledge of the patterns, forms, and meanings of morphemes without having to engage with metalinguistic descriptions and labels (e.g., Plante & Gómez, 2018).

Theories of implicit statistical learning are based on our ability to register, segment and internalize patterns, or statistical regularities, in our environment. Learning happens implicitly, i.e., there is no direct instruction involved. This ability has been examined in the context of language acquisition, amongst other areas. Extant research provides evidence of implicit statistical language learning in the first year of life (Saffran & Kirkham, 2018) and that this ability is sustained in adulthood (Saffran et al., 1997). The likelihood of pattern learning and retention increases with the amount of input (Plante & Gómez, 2018), and the amount of input needed varies among individuals. For example, Evans et al. (2009) found that children with developmental language disorders needed twice as much input as typically developing children to learn patterns implicitly.

Additionally, the variability of the input also influences the learning process (Torkildsen et al., 2013). If the target of learning is presented many times, with a high variability in non-target elements, the target becomes the most salient feature. For instance, if we want to teach the prefix *mis*, we could teach a couple of words such as 'misunderstand' and 'misuse'. However, the learner would likely just retain the whole-word understanding of these two examples. If, on the other hand, we greatly increase the number of words beginning with *mis*, the prefix becomes the most salient feature, e.g., *mis* means "wrong". Torkildsen et al. (2013) found that as many as 24 different exemplars may be needed to support generalization of the target element.

Educational apps are well suited to deliver large amounts of tailored input with high variability. Tasks can be presented with a minimum of explicit instructions or explanations, and immediate feedback facilitates learning by trial and error. Several educational apps rely on implicit learning to some degree. For example, the Duolingo apps for language, literacy and math all rely on principles of implicit statistical learning as a keystone in their design (Freeman et al., 2023).

Implicit learning relies on continued accumulation of input to identify regularities and statistical patterns. Thus, lapses of attention may be detrimental for implicit learning. For example, Toro et al. (2005) found that implicit learning of speech segmentation is affected by attention. Brosowsky et al. (2021), on the other hand, found that implicit learning in a serial reaction task using visual stimuli did not depend on attention. It is possible that attention plays different roles in implicit learning depending on types of input, e.g., auditory vs. visual stimuli, but this is not clear in the current literature.

Regardless of the role of attention, repeated mistakes can pose a hindrance to learning. If implicit learning happens without attention, repeated mistakes will expose learners to more incorrect input. One of the input principles presented by Plante and Gómez (2018) posits that all input is input in implicit learning. This means that incorrect input, if presented in large quantities, will lead to the learning of incorrect patterns. Thus, repeated mistakes may lead children to learn wrong patterns instead of the intended ones.

1.5. Repeated mistakes in educational games and assessments

In the current study, we define repeated mistakes as incorrect responses given more than once within a task. While there is a lack of studies investigating this construct, a previous study examined a related behavioral pattern. Hou (2015) investigated behavioral patterns when university students played a science education game. One such pattern was to follow up on an incorrect response by providing another incorrect response. Using cluster analysis, they identified three distinct clusters linked to students with low, medium, or high levels of self-reported flow. The author defines flow as "… a person's mental state when he is fully immersed in an activity and filtering out irrelevant emotions" (p. 425). The low-flow group exhibited a lack of transitions from mistakes back to analyzing the problem at hand, and they frequently followed one incorrect response with another. Furthermore, the low-flow group was the only group where students repeatedly responded incorrectly. This indicates that the propensity to give incorrect responses

repeatedly is associated with reduced levels of engagement and immersion. While Hou's (2015) study concerns university students, it seems likely that there is a similar association between disengagement and repetition of mistakes in younger learners as well.

Regarding the stability of the propensity to repeat mistakes, as well as relations to prior knowledge and learning, there is a lack of studies targeting this construct specifically. Hence, for comparison, we present findings regarding other behaviors relating to disengagement in the context of digital educational tools. In assessment settings, studies show that the frequency of rapid guessing increases over time, both within and across tests (Demars, 2007; Lindner et al., 2019). On the other hand, affective states like boredom, which are related to increases in gaming the system, are relatively persistent (Baker et al., 2010). While the study did not focus on the persistence of gaming the system specifically, the persistence of the related affective state of boredom makes it likely that levels of gaming the system are relatively stable over time, at least when students are bored. Regarding the propensity to repeat mistakes, it is unclear whether it is stable like gaming the system, or liable to change similarly to rapid guessing.

Concerning the relation to prior knowledge and learning outcomes, higher levels of affective states and behaviors such as disengagement and gaming the system have been associated with both lower levels prior knowledge and poorer learning outcomes. Baker et al. (2004) found that gaming the system was negatively associated with both pre-test and post-test scores. There is also evidence of long-term associations between gaming the system-behavior in intelligent tutoring systems and lower end-of-year grades (Pardos et al., 2013). It is likely that the same is true for the propensity to repeat mistakes. In implicit learning, repeated mistakes pose a threat not only by suspending the learning process, but also by increasing the exposure to incorrect information. If the students are exposed to more incorrect answers than correct ones, the incorrect information may become the most salient feature of the task content. Hence, when the children recall task content, the incorrect answers may overshadow the correct ones. Thus, there is a dual threat to learning, where children may receive less exposure to correct input, while receiving an inordinate amount of exposure to incorrect input.

2. Current study

The overarching aim of the current study is to examine how persistently children repeat mistakes when working with educational apps, and how the number of repeated mistakes relate to learning outcomes. More specifically, we exemplify the phenomenon using data from a morphology-based app developed to increase children's knowledge of both the meanings and written forms of morphologically complex words. Previous studies show that detrimental behaviors such as rapid guessing and gaming the system differ in persistence. While prior research suggests that rapid guessing increases both within and across tests, affective states associated with gaming the system are more stable (Baker et al., 2010; Demars, 2007; Lindner et al., 2019). These findings, however, related to change over relatively short time spans. In the current study, we examine children's behavior over an eight-week intervention period.

The intervention was effective in improving school children's word knowledge at the group level (Torkildsen et al., 2022) but unstructured observations from the classroom suggested large individual differences in how children interacted with the app. Specifically, some children appeared to answer without paying any apparent attention to which response option they chose or the feedback regarding the correctness of the response. This led to frequent repetitions of erroneous responses, indicating that the children did not learn from their mistakes. Hence, we decided to examine the count of repeated mistakes as a negative indicator of learning. Considering the findings from studies of rapid guessing (Demars, 2007; Lindner et al., 2019), we hypothesized that children might grow tired of the app over time, and start repeating mistakes as a result of disengagement due to boredom or fatigue. However, the propensity to repeat mistakes could also be more stable, as seems to be the case with gaming the system (e.g., Baker et al., 2010). Since there are no studies on the persistence of repeated mistakes, we aimed to uncover whether this behavior changes over time. Furthermore, it seemed likely that initial morphological knowledge affected the propensity to repeat mistakes and that the rates of repeated mistakes throughout the intervention would affect the final learning outcomes. We examined these hypotheses through the following research questions:

1. Does the propensity to repeat mistakes during an app-based language intervention change systematically over training sessions or does it remain stable?
2. How do rates of repeated mistakes relate to initial morphological knowledge and the final learning outcomes after eight weeks of using the app?

3. The morphology app

The app used in the present study was based on research regarding 1) how morphological knowledge supports word learning (Bertram et al., 2000; Bowers & Kirby, 2010; Goodwin & Ahn, 2013) and 2) how variability in non-target elements can support implicit language learning (Plante & Gómez, 2018; Torkildsen et al., 2013). Effects of working with the app for 8 weeks (40 sessions) were tested in a trial where 717 children were randomized to receive either the morphological app or an active control condition (a non-verbal mathematics app). Results showed robust generalization effects to untaught vocabulary containing trained morphemes. These effects were equally large at post-test and at follow-up six months later (Torkildsen et al., 2022).

3.1. Gamification and storyline

The app includes elements of gamification to increase the motivation of children while working (Zainuddin et al., 2020). These include elements targeting both intrinsic and extrinsic motivation. Extrinsic motivation is targeted through rewards, e.g. unlocking new levels (sessions) and advancing the storyline. The main element targeting intrinsic motivation is the inclusion of a storyline to

foster emotional and psychological engagement, as well as cognitive and behavioral involvement.

In the app, we follow the story of Morph, an alien training to become a spaceship captain. The first task given to the children is to help Morph with his final exam. This provides a backdrop for the receptive test of morphological word knowledge which was administered to the current sample before and immediately after the 8 weeks of training, as well as six months after the intervention.

Having passed his final exam and graduated as a captain, Morph embarks on his first journey. He soon encounters problems when he runs out of fuel (stardust) and crash lands on Earth. Here, the children have to help Captain Morph collect stardust by solving different tasks at different locations on the world map. Each completed session is marked by a flag raised at the session's map location and unlocks the next location on the map. In the cockpit of the spaceship, a stardust meter shows the current progress of fuel collection, indicating the proportion of completed sessions. The story is told through short videos and animations which are embedded into the children's work sessions.

3.2. Session structure

The 40 app sessions are structured into eight week plans containing five sessions each, intended to be played every day from Monday through Friday. The first four sessions in a week introduces new material (a new affix or compounding pattern), and the fifth session is a consolidation session composed of a mix of tasks from the preceding four sessions. Each app session consists of 25 tasks which all have to be completed before ending the session. The sessions are presented in a set order.

Following previous research on the effects of non-target variability on language learning and generalization, each morphological learning target is presented in the context of at least 24 root words in the course of a session. For example, in the session focusing on the affix *-ist*, children work with at least 24 different words ending in -ist, for example *guitarist, activist, Buddhist, florist, receptionist, journalist*, and so forth.

3.3. User interface and feedback

Fig. 1 gives an overview of the app's user interface. The app is developed for iPad. Users interact with the app through touch screen, by selecting images, dragging and dropping items, drawing arrows and writing via keyboard (see section 3.4. and Fig. 2 for details). There is audio support for all content in the app. Task instructions are read aloud when each screen is loaded and can be re-read by pressing a button. All words and affixes that children interact with can be read aloud by pressing the word itself. In line with research showing that variability in voices support retention of linguistic material (Richtsmeier et al., 2009), the app uses nine different voices, two adult voices for instruction and seven child voices for the rest of the app content.

Tasks require children to find a varying number of correct answers, shown by the number of star outlines in the top right corner of the screen (see Fig. 1). Every correct response gives immediate feedback through the filling-in of a star outline as well as the correct option being displayed on screen (specific positive feedback). Every incorrect answer gives immediate feedback in that the chosen response disappears and the incorrect response is reshuffled into the remaining response options (non-specific negative feedback). The

Fig. 1. User interface of the app.

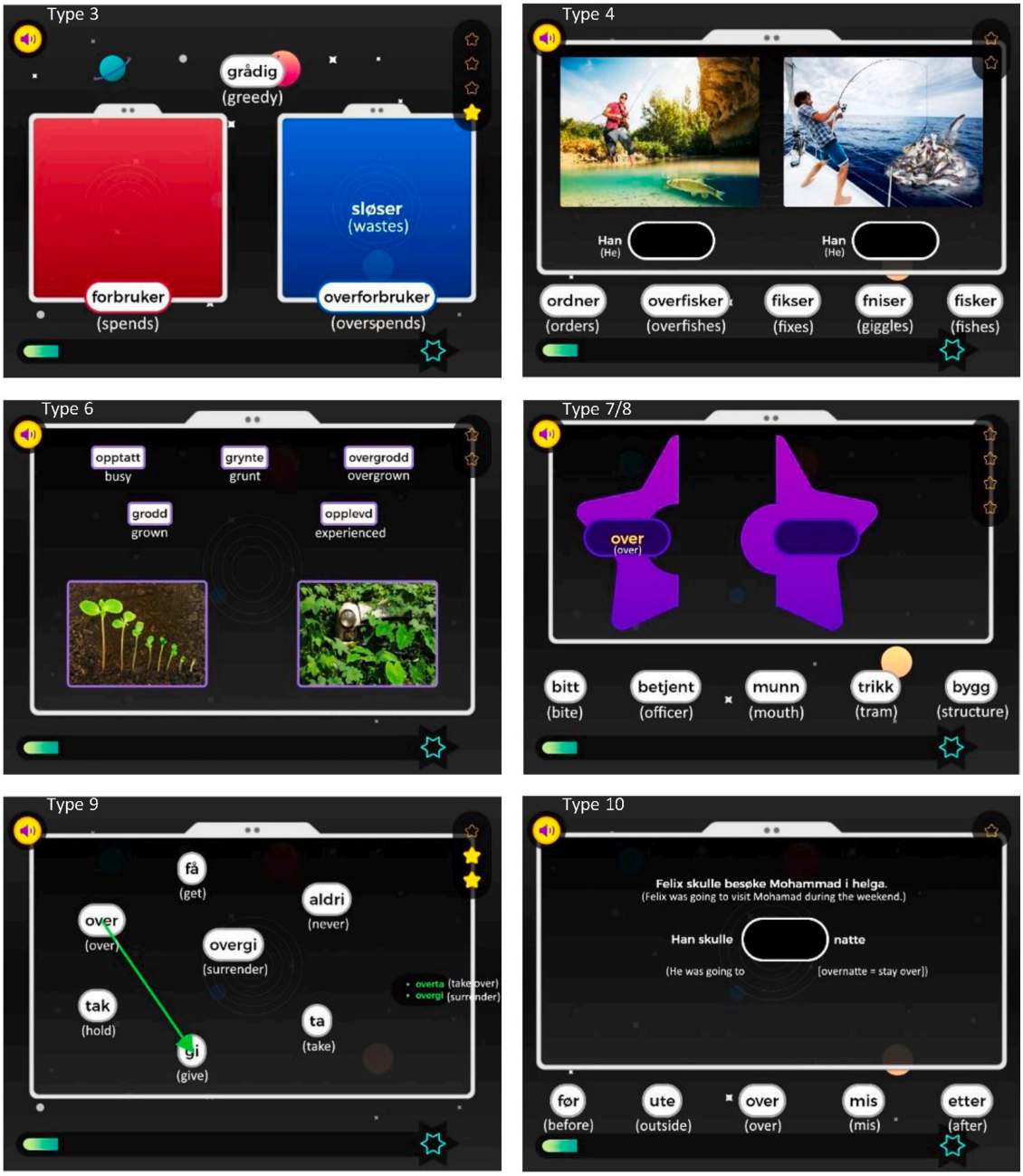


Fig. 2. Examples of the task types included in the analyses.

reshuffling of incorrect responses was implemented to discourage the type of gaming the system where children systematically try responses until they find the correct ones without engaging with the content. Session progress, i.e. proportion of tasks completed, is indicated by the stardust meter at the bottom of the screen. The main reward system is centered around progress, e.g. unlocking of new map locations containing new sessions or "levels" and collecting stardust with the end goal of helping Morph return to his home planet, rather than extrinsic rewards such as badges or scores.

3.4. Tasks

There are twelve different task types in the app (see Fig. 2 for task examples). Each session begins with two type 1 tasks and ends with a type 12 task. The remaining 22 tasks in each session are presented in random order. In accordance with the principles of implicit learning and high variability, all tasks require a certain number of correct answers before continuing on to the next task, and each task must be solved to complete the session.

Here we focus on the seven task types included in our analyses (see section 4.3.1.). For a description of the remaining tasks, see Torkildsen et al. (2022). In type 3 tasks (upper left panel of Fig. 2), the children are asked to sort words into two containers according to

their meaning. In the example, the instruction is "Drag the words that fit with 'spends' to the red box, drag the words that fit with 'overspends' to the blue box.". Task type 4 (upper right panel) presents two pictures with sentences describing the pictures. The sentences are missing a word or morpheme, and the children are asked to drag the correct word/morpheme to the open box in the sentences, e.g. "Drag the correct word to each sentence".In type 6 tasks (middle left panel), the children are asked to draw an arrow between two images and the words that best describe them: "Draw a line between corresponding words and pictures". In task types 7 and 8 (middle right panel), the children build words by dragging morphemes to the empty boxes, with one empty box in type 8 and two in type 7. The instruction for the example task is "Drag the parts that go together with 'over' to the empty space to form new words".

In task type 9 (lower left panel) the children are instructed to "Draw a line between parts that can combine to form a word". Finally, task type 10 (lower right panel) consists of two related sentences, where the second is missing a morpheme. The children are asked to "Drag the correct word part to the sentence".

3.5. Limitations of the app

The app's foundation in implicit learning provides a solid framework for learning, but also carries some limitations. To ensure that all students receive the required exposure to learning targets and variability in non-target elements, all tasks and all sessions had to be completed. This requirement, combined with the lack of information about the difficulty level of different linguistic items, prevented adaption of task difficulty. Also relating to the implicit nature of the app, feedback had to be kept at a simple level. Elaborate feedback would have required high levels of metalinguistic skills (e.g. explicit knowledge about word classes) for explanations to make sense.

4. Methods

4.1. Study design

The current study presents analyses of data collected in a larger project where we developed and evaluated a morphological app (Torkildsen et al., 2022). Morphological knowledge was assessed at three time points: before the intervention (pre-test), within two weeks after the intervention (post-test) and approximately 6–7 months after the intervention (follow-up). The present study uses data from the pre-test and post-test. Additionally, we gathered process data from children's interactions with the app. During the training sessions, the app recorded information such as time stamps, which response options the children attempted, correctness of responses, time between attempts, and use of audio support functions. In the current study, we use process data regarding which response options the children chose to identify repeated mistakes.

The intervention originally spanned 40 sessions over an eight-week period. On average, the children completed 38.16 sessions (*SD* = 5.05), with an average of 12 min and 49 s spent on each session (*SD* = 2 min and 17 s). However, the first two sessions were introductory sessions with much easier content. Additionally, every fifth session was a consolidation session containing tasks from the previous four sessions. Hence, we chose to omit these ten sessions from the analyses in the current study, retaining a total of 30 sessions.

4.2. Participants

The intervention study included 717 Norwegian second graders recruited from 12 schools in the eastern part of Norway. The schools were recruited from areas with varying socioeconomic status and proportions of children with language minority backgrounds. The children were randomly assigned to an experimental group working with the language app or an active control group working with another educational app. In the current study, we analyze data from the language app, which constrains our sample to the experimental group. This group originally consisted of 366 children (52.46 % girls, mean age 7.60). Twenty-six per cent of these children had a language minority background, i.e. neither parent was a native speaker of a Scandinavian language. Six percent of the children received some form of special education. Among the parents, 73% of mothers and 66% of fathers had a college or university degree. Three of the children in the experimental group dropped out during the first week. Hence, our sample consists of the remaining 363 children.

4.3. Measures

4.3.1. Repeated mistakes

In all tasks, the children were required to find a given number of correct answers before proceeding to the next task. While each correct answer was recorded and removed from the pool of response options, incorrect answers were reshuffled into the remaining response options. Thus, the children could select any incorrect option several times during a task. To calculate the number of repeated mistakes, we counted the number of erroneous responses given more than once in each task. Some task types do not allow for repeated mistakes, or do not track them in sufficient detail. Hence, the current analyses are restricted to seven task types: 3, 4, 6, 7, 8, 9 and 10 (see Fig. 2 for examples). In the type 3 task shown in the upper left panel of Fig. 2, the children are asked to sort words into boxes according to their meanings. In this example, if a child tries to put "sløser" (wastes) into the wrong box ("forbruker") three times during the task, this counts as two repeated mistakes. Likewise, if a child puts "sløser" and "grådig" (greedy) into the "forbruker" box twice each, this also counts as two repetitions. In the analyses, we use the mean number of repeated mistakes per task within each session as observed variables.

4.3.2. Test of receptive word knowledge

The test of receptive word knowledge measures children's ability to understand morphologically complex words, i.e. words that consist of two or more morphemes. The test was administered in the app, using a multiple-choice format. We used the binary item scores of 26 items as indicator variables in the analyses. The test is a researcher-developed assessment, described in detail elsewhere (Bratlie et al., 2022; Torkildsen et al., 2022). Kristensen et al. (2023) conducted an in-depth examination of the measurement properties of the test. Results indicated that it measures one dimension of morphological knowledge, namely (receptive) morphological analysis, which is the ability to use meaning-based knowledge of affixes to find the meaning of morphologically complex words. This supports the interpretation of test scores as indicators of meaning-based knowledge of morphologically complex words. Chronbach's alpha, estimated with the R package psych (Revelle, 2023), was 0.69 at pre-test and 0.82 at post-test. The increase in internal consistency between time points is likely due to a decrease in guessing at post-test.

4.4. Analyses

Regarding the first research question, we hypothesized that the propensity to repeat mistakes would change over time. However, we did not have specific hypotheses about the shape of the growth curve. Hence, we fit a nonlinear latent growth curve model to allow for freely estimated growth curves. We also fit a unidimensional confirmatory factor analysis (CFA) model to evaluate the potential stability of the construct over time (i.e., no systematic change).

To answer the second research question, we fit a structural equation model (SEM) where repeated mistakes mediated the relation between receptive morphological knowledge at pre-test and post-test. As the pre-test and post-test are repeated measures, we tested for longitudinal invariance. Our results suggested that there were five non-invariant items in the test (for details, see Appendix A). Hence, we specified a partially invariant model where the parameters of these five items were allowed to vary freely. Furthermore, the model specification depended on the results of RQ1. Should the evidence support repetition of mistakes as a state, we planned to extend the growth curve model into a SEM with both of the latent variables, intercept and slope, as mediators. On the other hand, should the evidence point to stability in the propensity to repeat mistakes, we planned to use the unidimensional representation of repeated mistakes as mediator. This allowed us to investigate the relation between initial knowledge and the propensity to repeat mistakes, as well as the relation between repeated mistakes and learning outcomes, while controlling for initial knowledge.

We conducted all analyses in R (R Core Team, 2021), using the package lavaan (Rosseel, 2012) for CFA and SEM analyses, and psych (Revelle, 2023) for descriptive statistics. For the growth curve model and the unidimensional model of repeated mistakes, we used full information maximum likelihood (FIML) estimation. Savalei and Bentler (2005) found that FIML estimation is robust for highly nonnormal data (skewness [−3.03, 6.67], kurtosis [19.48, 328.81]), with 15% or 30% missing data per variable.

In our data, skewness ranged from −1.61 to 3.73, except for one variable with skewness 8.15. Kurtosis ranged from −2.01 to 64.63, and the proportions of missing data ranged from 0% to 10.5%. As the rates of missing data and nonnormality were generally less severe in our data than in the study by Savalei and Bentler (2005), we proceeded with this approach, using robust standard errors and scaled test statistics. Since the items in the test of receptive word knowledge have binary scores, we used the diagonally weighted least squares estimator (DWLS) and polyserial correlations for the mediation model (Olsson et al., 1982). To minimize the loss of information due to missing responses, we based model estimation on pairwise information between variables.

5. Results

Table 1 shows the descriptive statistics for the total (raw) scores at pre-test and post-test, as well as the mean number of repeated mistakes across sessions. There was a relatively small difference of approximately three points between pre-test and post-test means. However, there was substantial variance in scores at both time points, with an even larger standard deviation at post-test. While the mean number of repeated mistakes per task across sessions and participants is 18.12, the largest amount of repeated mistakes made within a single task is 109. This highlights a substantial difference between children, and also between tasks for individual children.

Table 2 shows the correlations between pre-test, post-test and repeated mistakes. There is a strong positive correlation between pre-test and post-test measures, while there are moderate to strong negative correlations between number of repeated mistakes and pre-/post-test measures.

5.1. Propensity to repeat mistakes

Fig. 3 shows the observed individual growth curves. While there were peaks in some sessions, the overall trend appeared to be stable over time. This was confirmed by the estimated latent growth curve model. The model fit the data well ($\chi^2 = 620.077$, df = 403, $p < 0.001$, CFI = 0.952, TLI = 0.948, RMSEA = 0.043, SRMR = 0.050). Inspecting the factor loadings, however, we found that none of

Table 1
Descriptive statistics.

	Mean	SD	Skewness	Kurtosis	Min/Max
1. Pre-test total score	17.88	5.33	0.64	0.36	5/38
2. Post-test total score	21.10	7.60	0.46	−0.57	7/41
3. Repeated mistakes	18.12	8.83	0.73	0.06	4.17/47.80

Table 2
Correlations.

	1.	2.	3.
1. Pre-test total score	1		
2. Post-test total score	0.64	1	
3. Repeated mistakes	−0.55	−0.61	1

Note. All correlations are significant at $p < 0.001$.

the loadings on the slope factor were significant. This indicated that all the variance in the observed variables was explained by the intercept factor. In essence, there was no evidence of systematic changes over time. This was further confirmed by the results of the unidimensional CFA model, which showed acceptable fit to the data ($\chi^2 = 673.213$, df = 405, $p < 0.001$, CFI = 0.938, TLI = 0.933, RMSEA = 0.050, SRMR = 0.043).

5.2. Relation to prior knowledge and learning outcomes

The mediation model fit the data well ($\chi^2 = 3705.167$, df = 3224, $p < 0.001$, CFI = 0.960, TLI = 0.959, RMSEA = 0.020, SRMR = 0.068). Fig. 4 provides a path diagram showing the standardized regression coefficients.

Children's receptive knowledge at pre-test was negatively associated with the propensity to repeat mistakes ($\beta_a = -0.741$). Repeated mistakes were also negatively associated with learning outcomes at post-test ($\beta_b = 0.285$). The total effect of pre-test scores on post-test scores was 0.806, however, a significant proportion was due to the indirect effect through repeated mistakes ($\beta_a * \beta_b =$

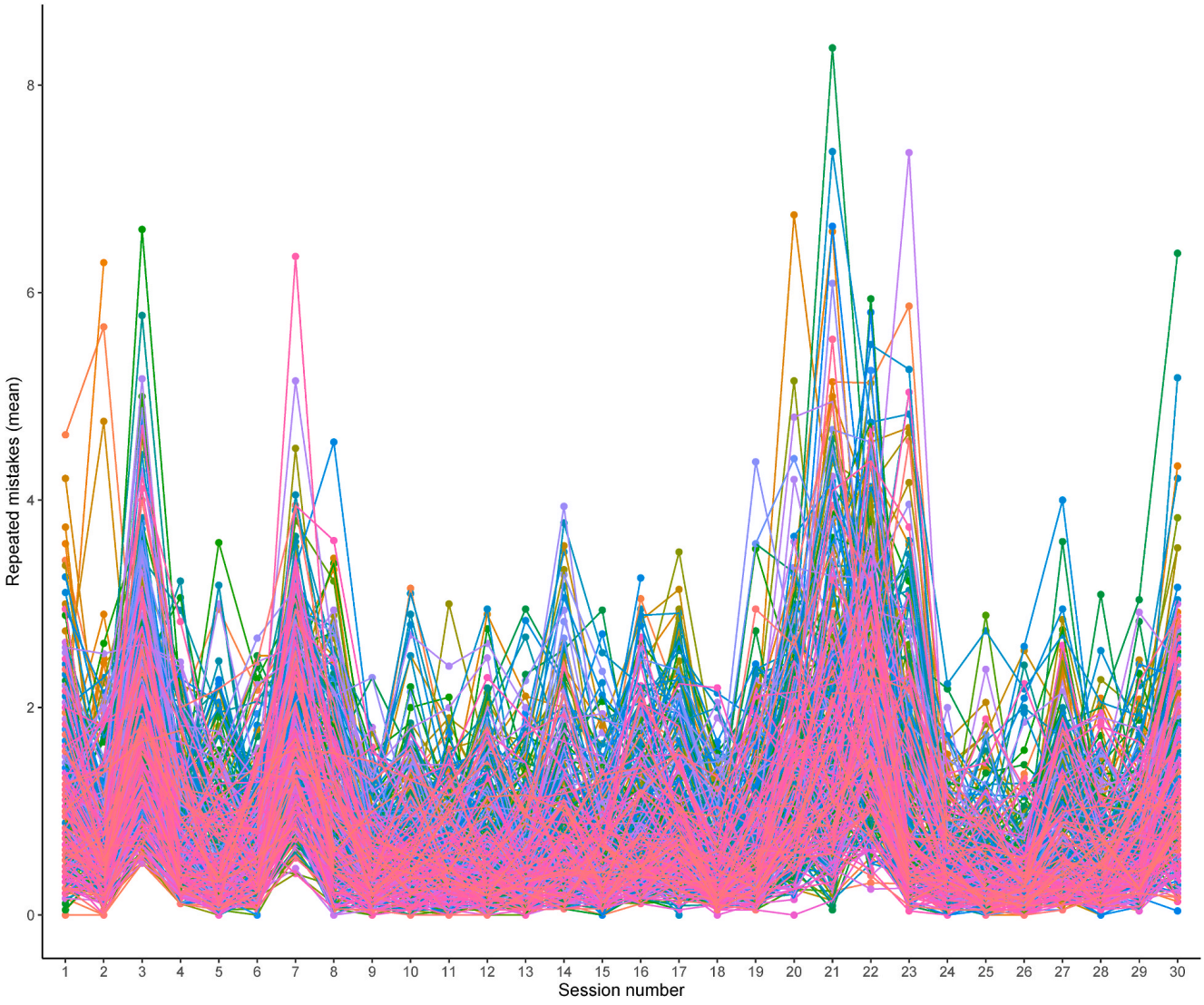


Fig. 3. Observed individual growth curves.

0.211, see Table 3).

To further illustrate how repeated mistakes mediated the relationship between pre-test and post-test, Fig. 5 shows the association between pre-test and post-test scores divided amongst the children with mean repeated mistakes in the lower 50% of the sample, and the children in the upper 50%. The regression lines show that the expected growth from pre-test to post-test was lower for the high group across all values of pre-test scores. For example, an average child in the low repeated mistakes group with a pre-test score of 13 has an expected post-test score of 20. An average child in the high repeated mistakes group with a pre-test score of 13, however, has an expected post-test score of 16. Thus, a pre-test score of 13 is associated with an expected seven-point increase in the low repeated mistakes group and only a three-point increase in the high repeated mistakes group.

6. Discussion

6.1. Stability of repeated mistakes

In line with research on rapid guessing, we hypothesized that the propensity to repeat mistakes might change across the sessions, for example as a result of disengagement due to fatigue (e.g. Lindner et al., 2019). Contrary to our hypothesis, however, the propensity to repeat mistakes remained stable across the eight weeks of training sessions. Thus, mistake repetition resembles gaming the system behavior in terms of persistence. While we implemented reshuffling of incorrect responses specifically to discourage systematic selection of responses without engaging with the content, it is likely that some children still engaged in such behavior. Thus, it is possible that repeated mistakes, at least in some cases, represents gaming the system "gone wrong". Some sessions showed collective spikes of increased repetition of mistakes, probably due to content-specific variation, e.g. difficulty. Yet the overall trend shows a striking consistency, as evidenced by the non-significant loadings on the slope factor in the growth curve model, as well as the good fit of the unidimensional model of repeated mistakes. This finding carries important implications for classroom practices. Since the propensity to repeat mistakes seems to be stable over time, it is unlikely that it will change without some form of intervention. Hence, it becomes important to identify children who are more likely to repeat mistakes and to examine how we might help them break this pattern. While we cannot make any conclusive claims, it also seems likely that the propensity to repeat mistakes is a stable behavioral pattern that affects learning contexts other than our language app. The negative associations with prior knowledge and learning outcomes, discussed in the following sections, makes it imperative to find ways to ameliorate repetition of mistakes.

6.2. Prior knowledge and repeated mistakes

In line with previous research on gaming the system (Baker et al., 2004), there was a strong negative association between prior knowledge and the propensity to repeat mistakes. This could indicate that children repeat mistakes more often when faced with tasks that are difficult relative to the child's current level of knowledge. The underlying mechanism is not entirely clear, however. Frustration or boredom due to difficulties with understanding tasks can lead to disengagement. In such cases, children respond without paying any attention to the responses they give. Along these lines, the lack of attention could explain the negative effect of repeated

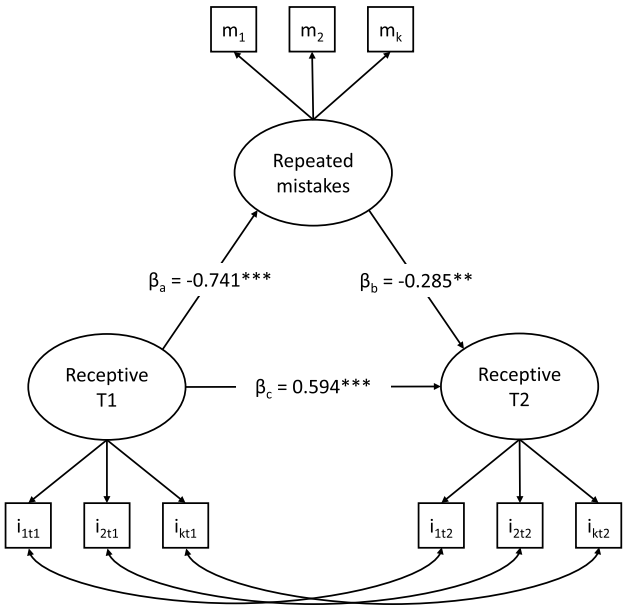


Fig. 4. Structural relation between pre-test and post-test, mediated by repeated mistakes
Note. The model is exemplified with three indicators per factor for readability. The Receptive factors have 26 indicators at each time point, with correlated residuals between same items across time points. The repeated mistakes factor has 30 indicators. **p < 0.01, ***p < 0.001.

Table 3
Direct and indirect effects on post-test scores of receptive word knowledge.

	Estimate	*p*-value
β_a	−0.741	$p < 0.001$
β_b	−0.285	$p = 0.008$
β_c	0.594	$p < 0.001$
Indirect effect ($\beta_a * \beta_b$)	0.211	$p = 0.005$
Total effect ($\beta_a * \beta_b + \beta_c$)	0.806	$p < 0.001$

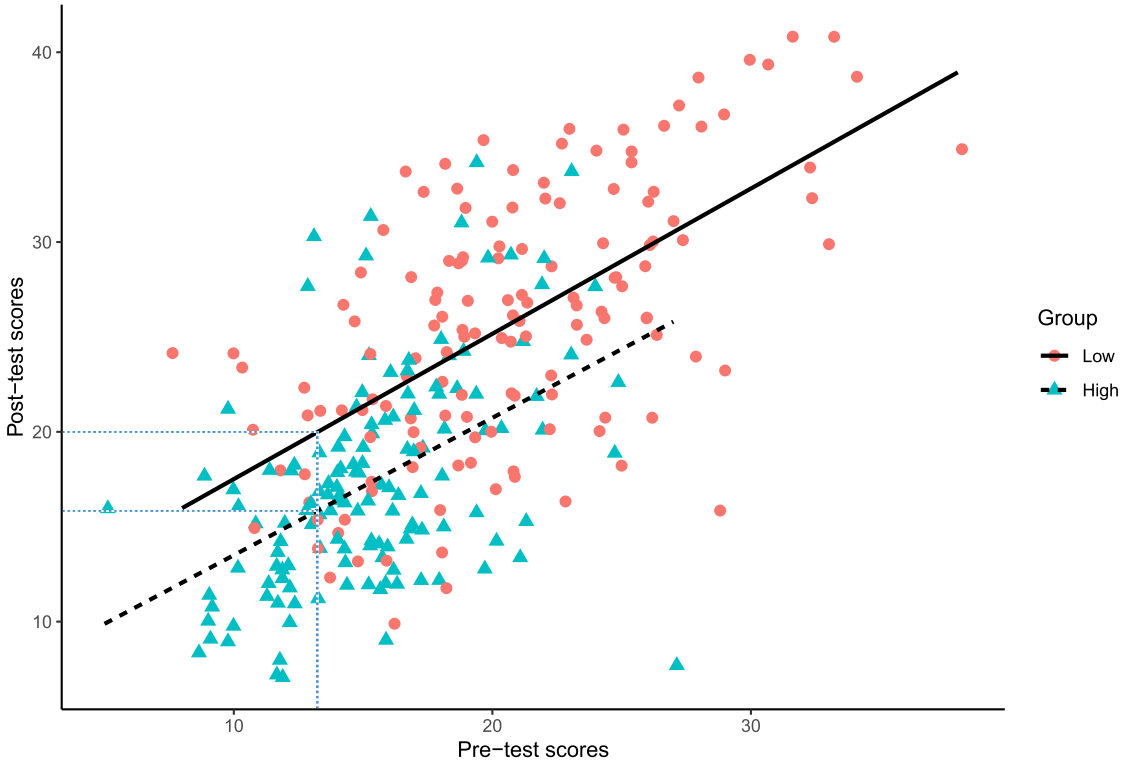


Fig. 5. Association between pre-test and post-test for high and low propensity groups.
Note. Scores are raw score sums at pre-test and post-test. Low group (circles and solid line) = children with mean repeated mistakes in the lower 50% of the sample. High group (triangles and dotted line) = children in the upper 50%. Vertical/horizontal lines show expected post-test values given pre-test values for each group.

mistakes on learning outcomes. Alternatively, higher ratios of repeated mistakes could be the result of misconceptions. It is conceivable that children will be inclined to attempt an incorrect option more than once if they are convinced (wrongly) that the answer is correct.

6.3. Repeated mistakes and learning outcomes

Pre-test scores normally explain a large amount of the variance in post-test scores. This is also true in our results, where the total effect of pre-test knowledge on post-test outcomes was 0.806. However, a substantial proportion of the total effect was due to the mediation through repeated mistakes ($\beta_a * \beta_b = 0.211$). As is shown in Fig. 5, the children who scored relatively high on the pre-test, but made many repeated mistakes, showed less growth in the post-test measure compared to those who made fewer repeated mistakes. Simultaneously, those who had lower scores on the pre-test, yet made fewer repeated mistakes, showed greater growth from the pre-test to the post-test.

To exemplify a potential mechanism underlying this association to repeated mistakes, imagine a task where the child needs to find two correct answers. In the process, the child responds incorrectly more than 100 times before selecting both of the correct answers. The child is then exposed to an enormous proportion of incorrect input. Not only will this reduce the opportunities to learn the correct pattern, it will also increase the probability of learning incorrect ones. Such extreme cases of more than 100 repeated mistakes within a task, while rare, do occur in the data we analyzed. Considering a possible double threat to learning, i.e. less learning of correct patterns combined with increased learning of incorrect ones, it is no wonder that the propensity to repeat mistakes is associated with poorer learning outcomes.

6.4. Implications and limitations

For research purposes, repeated mistakes can represent an important measure of fidelity, since children with a high propensity to repeat mistakes do not use the app in the intended manner. Due to its negative effect on learning outcomes, repeated mistakes may act as a confounding factor when assessing intervention effects. While it is not clear whether the negative effects on learning are due to disengagement or retention of incorrect patterns, it is important to know whether children are behaving unexpectedly and how this behavior relates to learning gains. Thus, when evaluating effects of app-based interventions, researchers should control for measures of unintended behavior such as repeated mistakes. Examination of unintended behavior can elucidate the mechanisms which lead to differences in learning gains. Future studies should thus examine whether children who repeat mistakes retain patterns learned from incorrect input, for example, how repetition of specific mistakes relates to specific errors during post-tests.

The results of our analyses indicate that children's propensity to repeat mistakes is relatively stable over time, thus resembling gaming the system more than rapid guessing in this respect. Given the reshuffling of incorrect answers, it is likely that mistake repetition in some cases represent an "unsuccessful" form of gaming the system. This is an area that needs further examination, for example by having children complete some sessions with reshuffling and some without it. If repeated mistakes are indeed a form of gaming the system, we would expect the children with high propensity to repeat mistakes to also exhibit higher levels of gaming the system more generally.

Furthermore, given the negative impact of repeating mistakes, there is a need to intervene to help children interact with the app in ways that are more constructive. This could be implemented as specific corrective feedback given to the children through the app. While more elaborative feedback could also be beneficial, this is difficult to achieve without making excessive demands on the children's metalinguistic skills. Another possibility is to notify teachers when children repeat mistakes, e.g. through a dashboard function, so that the teachers can intervene. Either way, future studies should consider how to break the negative interaction patterns. A third possibility would be to mark or remove incorrect response options after they have been chosen once. This would, however, open up for the systematic trial-and-error version of gaming the system.

To our knowledge, this study presents the first investigation of the characteristics of repeated mistakes and their relation to learning outcomes in app-based learning. We modeled repeated mistakes as a unidimensional construct at the level of sessions, but it is possible that different patterns of repetition represent different underlying constructs on the item level. For example, there may be differences between repeating the same mistake four times and repeating four mistakes one time each. Differentiating between such patterns was beyond the scope of the current study but should be addressed in future research. On a related note, the inclination to make repeated mistakes was time-invariant across the sessions in the intervention, but we do not know if this was also the case within sessions. Future research should examine whether children are more likely to repeat mistakes towards the end of a session, for example due to fatigue. There is also a need for research on how characteristics of the child, task and session relate to the frequency of mistake repetition. Understanding which children are more likely to repeat mistakes can help us provide the necessary support, whereas knowledge of which tasks and sessions elicit more repeated mistakes can guide future app development.

7. Conclusion

This study investigated the propensity to repeat mistakes in app-based word learning. We examined whether the propensity changes over time, and how it relates to prior knowledge and learning outcomes in an eight-week language intervention. Our results show that the propensity to repeat mistakes was stable over time, and that children with lower levels of prior knowledge were likely to make more repeated mistakes. Furthermore, a higher propensity to repeat mistakes was related to poorer learning outcomes. This could constitute a dual threat to learning. On one hand, children who repeat more mistakes may not register which responses they choose or whether or not their choices are correct. In this case, they will not learn from their mistakes, hence gaining less knowledge from working with the app. On the other hand, following Plante and Gómez (2018), all input is input in implicit learning. This means that children with a high propensity to repeat mistakes are exposed to inordinate amounts of incorrect input, making erroneous patterns more salient than correct ones. In this case, they gain more incorrect knowledge from the app. Either way, it is unlikely that the propensity to repeat mistakes is confined to a specific app. Thus, it is imperative to examine such behavior across different contexts, and to find out which children are more likely to engage in it, as well as how we can help the children break such negative interaction patterns.

Credit author statement

Jarl Kleppe Kristensen: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization. **Janne von Koss Torkildsen:** Conceptualization, Investigation, Writing – Review & Editing. **Björn Andersson:** Conceptualization, Methodology, Formal Analysis, Data Curation, Writing – Review & Editing.

Data availability

The authors do not have permission to share data.

Acknowledgements

We would like to thank all participating students, teachers, schools and municipalities. This work was supported by the Research Council of Norway, Grant no. 24033.

Appendix A

The participants completed the same test of receptive morphological knowledge at pre-test and post-test. Hence, we investigated longitudinal measurement invariance to examine whether the test items measure the same construct at different time points. In the first step, we compared a fully invariant model to a configural baseline model with no invariance restrictions. Since the item scores are binary, we simultaneously restricted thresholds, intercepts, and factor loadings in the invariant model. The fully invariant model fit the data significantly worse than the configural model (see Table A1). Thus, we proceeded to estimate separate models releasing restrictions on each item while keeping all other items invariant. Five items showed significant improvement of model fit when restrictions were released ($p < 0.00192$, using Bonferroni correction for testing 26 individual models). In the final step, we fit a model where these five items were allowed to vary freely while keeping the restrictions on the remaining 21 items. Comparing this partially invariant model to the configural model, the likelihood ratio test showed no significant difference between the models (Table A1). Following these results, we used the partially invariant model when testing for mediating effects of repeated mistakes.

Table A1
Invariance tests for the longitudinal model of receptive morphological knowledge

	χ^2	df	$\Delta\chi^2$	Δdf	p
Baseline	1187.8	1247			
Full invariance	1284.4	1271	62.622	24	<.001
Baseline	1187.8	1247			
Partial invariance	1221.3	1261	17.010	14	0.256

The partially invariant longitudinal model for receptive morphological knowledge at pre-test and post-test fit the data well ($\chi^2 = 1354.350$, $df = 1261$, $p < 0.05$, $CFI = 0.963$, $TLI = 0.961$, $RMSEA = 0.014$, $SRMR = 0.080$). The factors were highly correlated ($r = 0.804$, $p < 0.001$).

References

Ash, G. E., & Baumann, J. F. (2017). Vocabulary and reading comprehension: The nexus of meaning. In S. E. Israel, & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (2nd ed., pp. 347–370). Routledge.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students" game the system". In E. Dykstra-Erickson, & M. Scheligi (Eds.), *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 383–390). https://doi.org/10.1145/985692.985741

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Bertram, R., Laine, M., & Virkkala, M. M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology, 41*(4), 287–296. https://doi.org/10.1111/1467-9450.00201

Bowers, P. N., & Kirby, J. R. (2010). Effects of morphological instruction on vocabulary acquisition. *Reading and Writing, 23*, 515–537. https://doi.org/10.1007/s11145-009-9172-z

Bratlie, S. S., Gustafsson, J. E., & Torkildsen, J. V. K. (2022). Effectiveness of a classroom-implemented, app-based morphology program for language-minority students: Examining latent language-literacy profiles and contextual factors as moderators. *Reading Research Quarterly, 57*(3), 805–829. https://doi.org/10.1002/rrq.447

Brosowsky, N. P., Murray, S., Schooler, J. W., & Seli, P. (2021). Attention need not always apply: Mind wandering impedes explicit but not implicit sequence learning. *Cognition, 209*, 1–14. https://doi.org/10.1016/j.cognition.2020.104530

Callaghan, M. N., & Reich, S. M. (2018). Are educational preschool apps designed to teach? An analysis of the app market. *Learning, Media and Technology, 43*(3), 280–293.

Cervetti, G. N., Fitzgerald, M. S., Hiebert, E. H., & Hebert, M. (2023). Meta-analysis examining the impact of vocabulary instruction on vocabulary knowledge and skill. *Reading Psychology*, 1–38. https://doi.org/10.1080/02702711.2023.2179146

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research, 71*(1), 1–27. https://doi.org/10.3102/00346543071001001

Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaee, E., & Noroozi, O. (2021). Using gamification to support learning English as a second language: A systematic review. *Computer Assisted Language Learning, 34*(7), 934–957.

Demars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23–45. https://doi.org/10.1080/10627190709336946

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing, 52*(2), 321–335. https://doi.org/10.1044/1092-4388(2009/07-0189

Freeman, C., Kittredge, A., Wilson, H., & Pajak, B. (2023). *The Duolingo method for app-based teaching and learning*. Duolingo. https://duolingo-papers.s3.amazonaws.com/reports/duolingo-method-whitepaper.pdf.

Glover, I. (2013). Play as you learn: Gamification as a technique for motivating learners. In *Proceedings of world conference on educational Multimedia*. Hypermedia and Telecommunications, 2013 http://shura.shu.ac.uk/7172/.

Goodwin, A. P., & Ahn, S. (2013). A meta-analysis of morphological interventions in English: Effects on literacy outcomes for school-age children. *Scientific Studies of Reading, 17*(4), 257–285. https://doi.org/10.1080/10888438.2012.689791

Heil, C. R., Wu, J. S., Lee, J. J., & Schmidt, T. (2016). A review of mobile language learning applications: Trends, challenges, and opportunities. *The EuroCALL Review, 24*(2), 32–50.
Hou, H. T. (2015). Integrating cluster and sequential analysis to explore learners' flow and behavioral patterns in a simulation game with situated-learning context for science courses: A video-based process exploration. *Computers in Human Behavior, 48*, 424–435. https://doi.org/10.1016/j.chb.2015.02.010
Kristensen, J. K., Andersson, B., Bratlie, S. S., & Torkildsen, J. V. (2023). Dimensionality of morphological knowledge—evidence from Norwegian third graders. *Reading Research Quarterly*, 406–424. https://doi.org/10.1002/rrq.497
Levesque, K. C., Breadmore, H. L., & Deacon, S. H. (2021). How morphology impacts reading and spelling: Advancing the role of morphology in models of literacy development. *Journal of Research in Reading, 44*(1), 10–26. https://doi.org/10.1111/1467-9817.12313
Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology, 10*, 1–15. https://doi.org/10.3389/fpsyg.2019.01533
Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistics Review, 4*(1), 151–172.
Montazami, A., Pearson, H. A., Dube, A. K., Kacmaz, G., Wen, R., & Alam, S. S. (2022). Why this app? How educators choose a good educational app. *Computers & Education, 184*. https://doi.org/10.1016/j.compedu.2022.104513
Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly, 19*(3), 304–330. https://doi.org/10.2307/747823
Nikolayev, M., Reich, S. M., Muskat, T., Tadjbakhsh, N., & Callaghan, M. N. (2021). Review of feedback in edutainment games for preschoolers in the USA. *Journal of Children and Media, 15*(3), 358–375. https://doi.org/10.1080/17482798.2020.1815227
Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*, 337–347. https://doi.org/10.1007/BF02294164
Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 117–124).
Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*(2), 184–202. https://doi.org/10.1598/RRQ.40.2.3
Plante, E., & Gómez, R. L. (2018). Learning without trying: The clinical relevance of statistical learning. *Language, Speech, and Hearing Services in Schools, 49*(3S), 710–722. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0131
R Core Team. (2021). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.
Revelle, W. (2023). *psych: Procedures for psychological, Psychometric, and Personality research*. Evanston, Illinois: Northwestern University. R package version 2.3.3 https://CRAN.R-project.org/package=psych.
Richtsmeier, P. T., Gerken, L., Goffman, L., & Hogan, T. (2009). Statistical frequency in perception affects children's lexical production. *Cognition, 111*(3), 372–377.
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02
Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology, 69*, 181–203.
Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental Language learning: Listening (and learning) out of the corner of Your ear. *Psychological Science, 8*(2), 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x
Savalei, V., & Bentler, P. M. (2005). A statistically justified pairwise ML method for incomplete nonnormal data: A comparison with direct ML and pairwise ADF. *Structural Equation Modeling, 12*(2), 183–214. https://doi.org/10.1207/s15328007sem1202_1
Snow, C. E., & Matthews, T. J. (2016). Reading and language in the early grades. *The future of children, 26*(2), 57–74. http://www.jstor.org/stable/43940581.
Tärning, B. (2018). Review of feedback in digital applications–does the feedback they provide support learning? *Journal of Information Technology Education: Research, 17*, 247.
Torkildsen, J. V. K., Bratlie, S. S., Kristensen, J. K., Gustafsson, J.-E., Lyster, S.-A. H., Snow, C., et al. (2022). App-based morphological training produces lasting effects on word knowledge in primary school children: A randomized controlled trial. *Journal of Educational Psychology, 114*(4), 833–854. https://doi.org/10.1037/edu0000688
Torkildsen, J.v. K., Dailey, N., Aguilar, J., Gómez, R., & Plante, E. (2013). Exemplar variability facilitates rapid learning of an otherwise unlearnable grammar by individuals with language-based learning disability. *Journal of Speech, Language and Hearing Research, 56*(2), 618–629. https://doi.org/10.1044/1092-4388(2012/11-0125
Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*(2), B25–B34. https://doi.org/10.1016/j.cognition.2005.01.006
Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61. https://doi.org/10.1111/emip.12165
Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review, 30*.

Article 4

Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2023). *Who repeats mistakes and when? Task and child covariates of repeated mistakes in app-based learning* [Manuscript in preparation]. Centre for Educational Measurement, University of Oslo.

Appendix: Errata

Errata List

Name of candidate:

Jarl Kleppe Kristensen

Title of thesis:

Pieces in the Puzzle of Language Learning. On the Roles of Morphological Knowledge, App-Based Implicit Learning and Child-App Interactions.

Abbreviations:

Form: correction of formatting, Ref: correction of reference, Spell: correction of spelling

Page	Original text	Type of correction	Corrected text
IX	3) Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2023).	Ref. (Year changed after printed version was published.)	3) Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2024).
8	"…the ability to recognize, understand, manipulate and produce spoken and written morphemes"	Form.	"…the ability to recognize, understand, manipulate and produce spoken and written morphemes"
9	…, as is discussed in article 2.	Form.	…, as is discussed in Article 2.
9	Morphological awareness, or the explicit knowledge of morphemes and morphological processes relate…	Spell.	Morphological awareness, or the explicit knowledge of morphemes and morphological processes, relates…
12	…the success of such app…	Spell.	…the success of such apps…
25	…research based educational application…	Spell.	…research-based educational application…
25	…the principle of *justice* focus…	Spell.	…the principle of *justice* focuses…
29	The results provides evidence…	Spell.	The results provide evidence…
30	Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2023).	Ref. (Year changed after printed version was published.)	Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2024).
32	…an intervention is time limited…	Spell.	…an intervention is time-limited…
45		Ref. (Reference missing from the reference list.)	Rodrigo, M. M. T., Baker, R. S., Lagud, M. C., Lim, S. A. L., Macapanpan, A. F., Pascua, S. A. M. S., Santillano, J. Q., Sevilla, L. R. S., Sugay, J. O., Tep, S., & Viehland, N. J. (2007). Affect and usage choices in simulation problem solving environments. In R. Luckin, K. R. Koedinger, & J. Greer. (Eds.), *Artificial Intelligence in Education. Building Technology Rich Learning Contexts That Work* (145-152). IOS Press.
Article 4	All references to Kristensen et al. (2023), including reference list.	Ref. (Year changed after printed version was published.)	Kristensen et al. (2024).