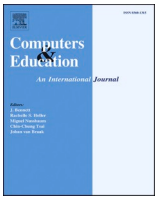
Computers & Education 210 (2024) 104966

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.

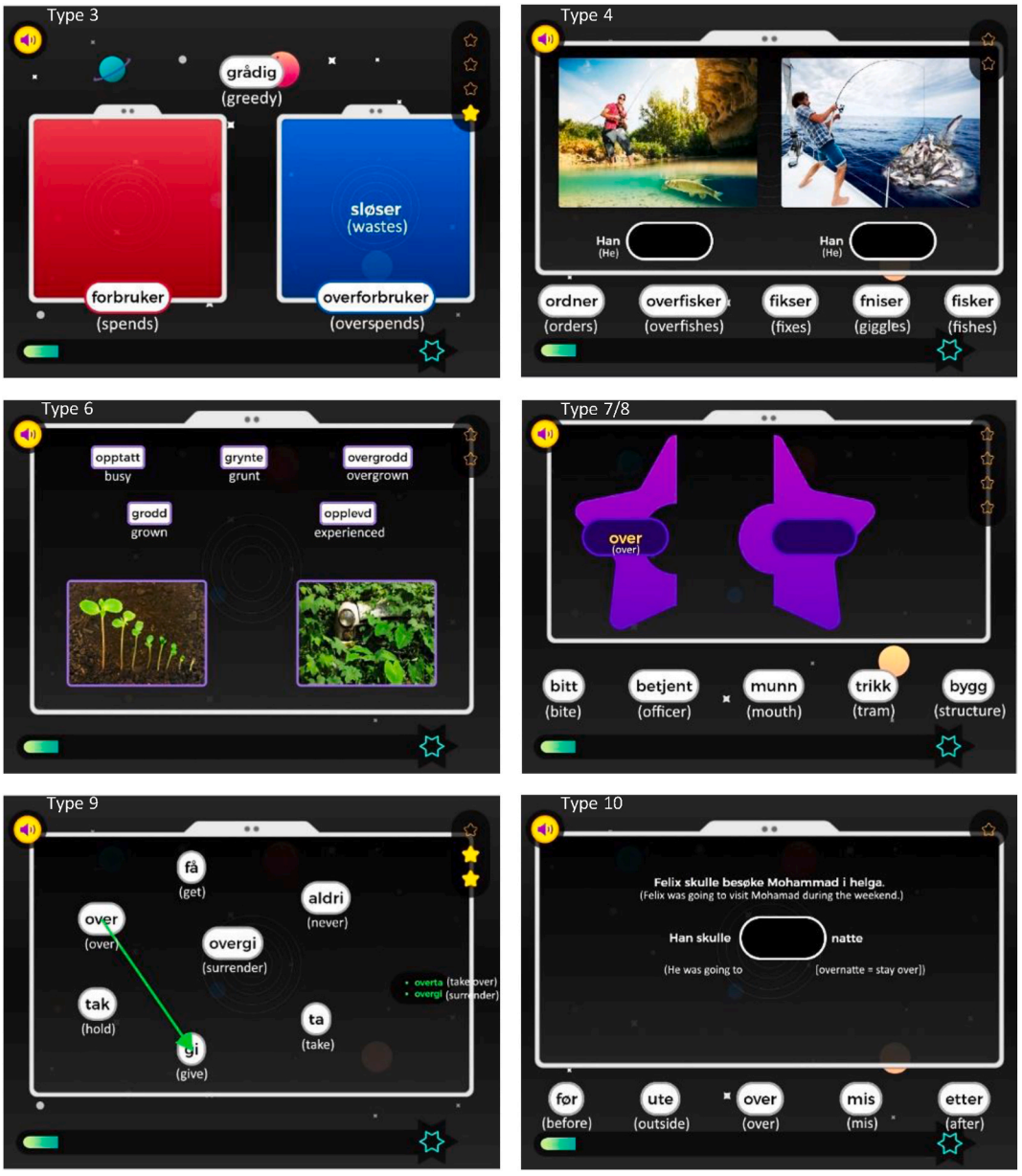


Fig. 2. Examples of the task types included in the analyses.

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 ($\beta_a * \beta_b =$

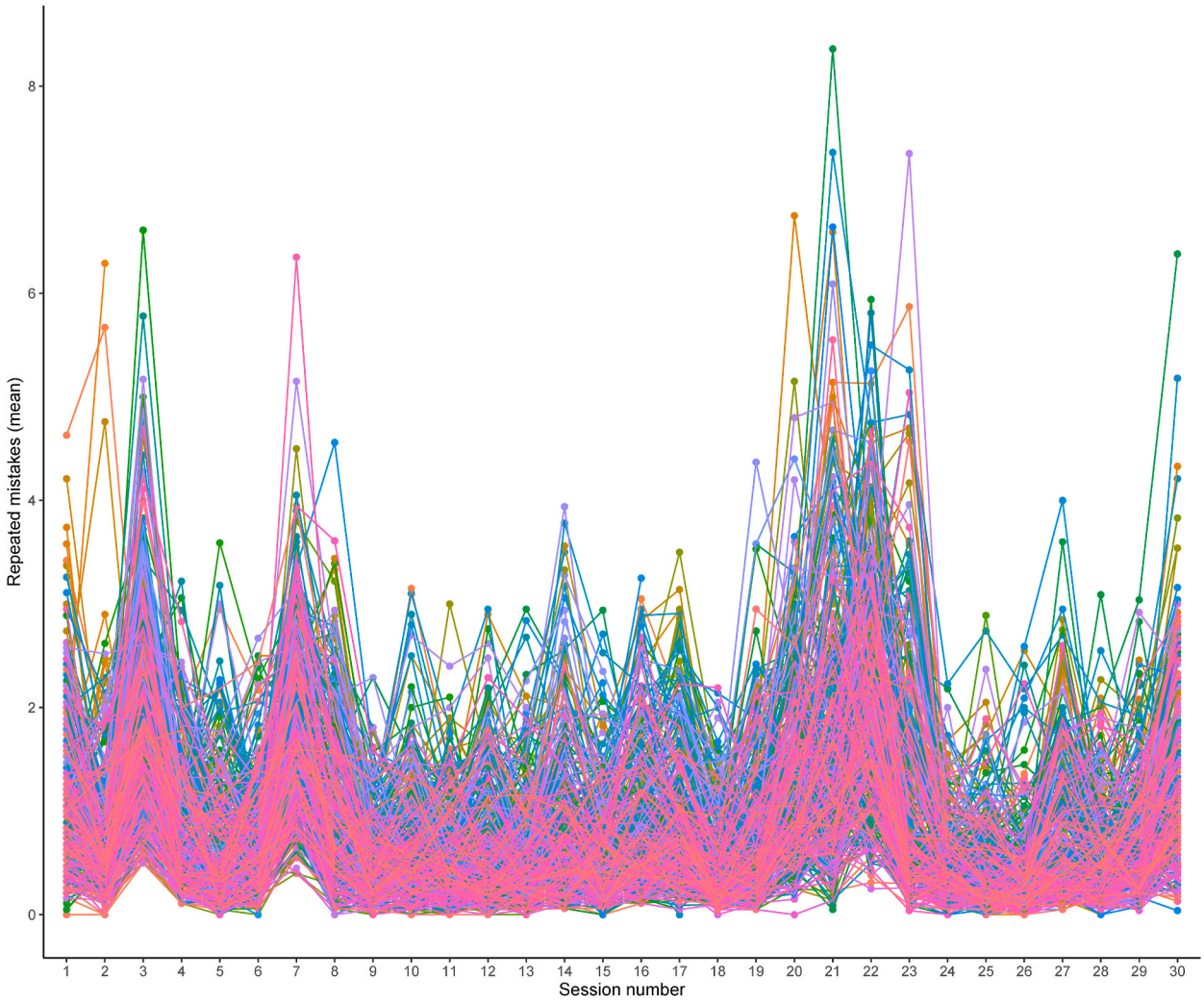


Fig. 3. Observed individual growth curves.

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

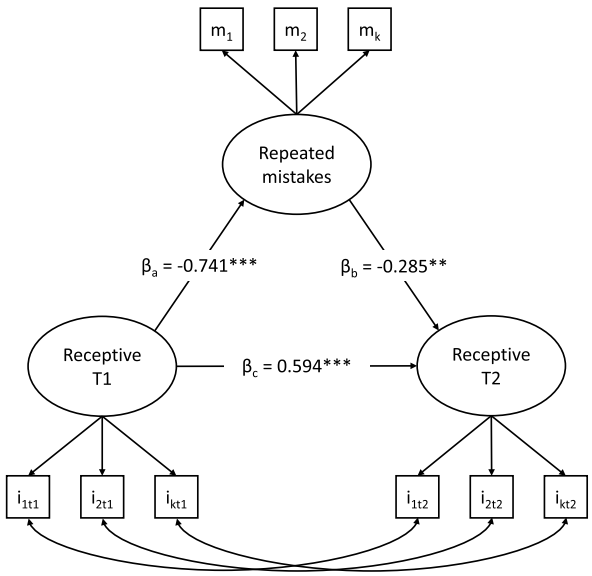


Fig. 4. Structural relation between pre-test and post-test, mediated by repeated mistakes
Note. The model is exemplified with three indicators per factor for readability. The Receptive factors have 26 indicators at each time point, with correlated residuals between same items across time points. The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

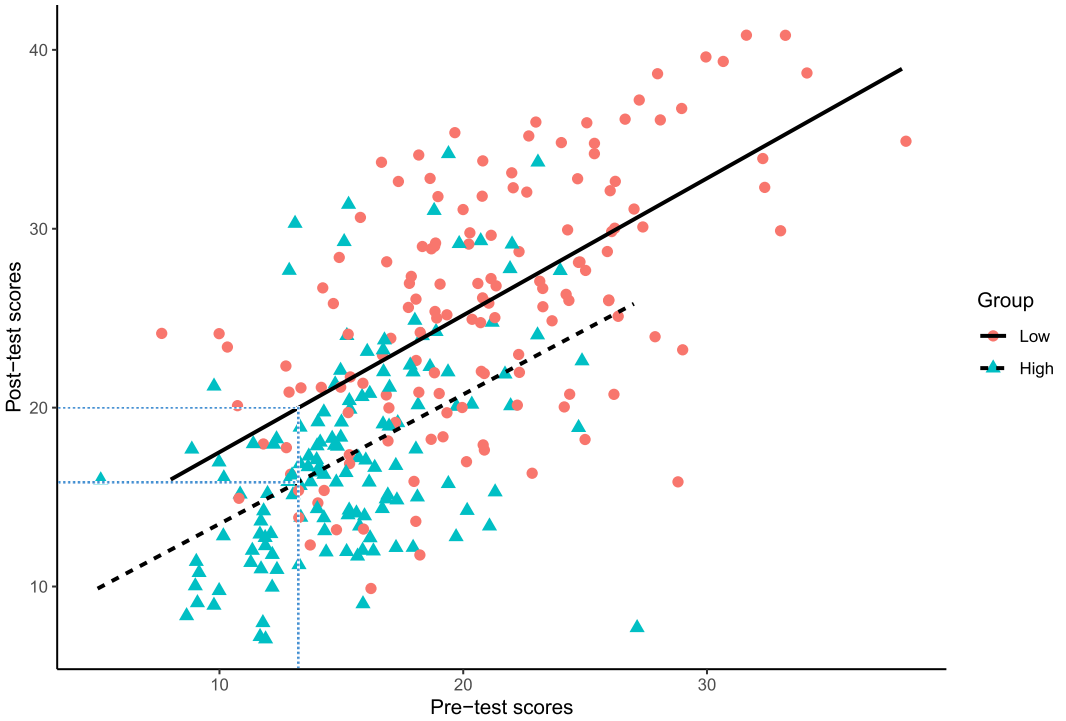


Fig. 5. Association between pre-test and post-test for high and low propensity groups.
Note. Scores are raw score sums at pre-test and post-test. Low group (circles and solid line) = children with mean repeated mistakes in the lower 50% of the sample. High group (triangles and dotted line) = children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