

ORIGINAL ARTICLE OPEN ACCESS

Deceptive Overgeneralization: When Adaptive Learning Enables Systematic Misapplication

Marshall An  | Bruce M. McLaren  | John Stamper 

Human–Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Correspondence: Marshall An (haokanga@andrew.cmu.edu)

Received: 19 February 2026 | Revised: 13 July 2026 | Accepted: 20 July 2026

Keywords: adaptive learning | bayesian knowledge tracing | deceptive overgeneralization | intelligent tutoring systems | mastery learning

ABSTRACT

Background: Adaptive learning systems infer mastery from observed correctness, but this approach may fall short when learners exhibit ‘deceptive overgeneralization’—performing correct actions based on incomplete understanding. This phenomenon can mislead systems into prematurely stopping practice, leaving knowledge gaps unaddressed. Clarifying when deceptive overgeneralization occurs and how to address it can improve the precision of adaptive learning technologies.

Objectives: This research empirically examines whether deceptive overgeneralization occurs in adaptive learning, assesses its consequences for mastery-based stopping rules, and evaluates a systematic procedure for detecting and remedying it.

Methods: We conducted 11 experiments using Intelligent Tutoring Systems (ITSs) for Riichi Mahjong across multiple semesters with university students (total $N=192$). The studies employed a pretest–posttest design to assess ITS effectiveness, and within-subjects designs with adaptive and full-exposure conditions to examine overgeneralization and remediation. Tailored remediation consisted of targeted ‘do-not-act’ problems with feedback that names the missing application constraints.

Results and Conclusions: ITSs improved learning outcomes, but learners frequently misapplied acquired actions in contexts where they should have refrained, indicating deceptive overgeneralization. Under adaptive learning, mastery stopping rules often ended practice before learners encountered such ‘do-not-act’ contexts, leaving the issue undetected. Tailored remediation using targeted ‘do-not-act’ problems with constraint-naming feedback reduced misapplication to near-floor levels across skills and cohorts, highlighting a limitation of correctness-based mastery inference and a practical method to strengthen mastery assessment for knowledge compilation. A secondary analysis of K–12 decimal learning data further illustrates how the framework applies in a traditional educational domain.

1 | Introduction

Computer-assisted learning (CAL) broadly refers to the use of computers to support education or training (Oxford Reference, [n.d.](#)). Within CAL, adaptive learning technologies dynamically adjust instruction based on learner performance, aiming to optimize learning efficiency and mastery (Aleven, McLaughlin, et al. 2016). Their effectiveness is well supported by various theoretical perspectives, including Vygotsky’s zone of proximal development (Vygotsky 1978), cognitive apprenticeship (Collins et al. 1991), the expertise reversal effect (Kalyuga 2009),

the assistance dilemma (Koedinger and Aleven 2007) and mastery learning (Bloom 1968).

Adaptivity is not a binary property, but rather ‘a matter of degree’ (Aleven, McLaughlin, et al. 2016, 523). Intelligent tutoring systems (ITSs) distinguish themselves by adapting across all three major time scales defined by the Adaptivity Grid: step, task, and design (Aleven, McLaughlin, et al. 2016, 525). Within the step-loop, ITSs provide timely and targeted feedback at each problem-solving step. Indeed, timely feedback is critical to enable learners to continuously

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Journal of Computer Assisted Learning* published by John Wiley & Sons Ltd.

Lay Summary

- What is currently known about this topic?
 - Adaptive learning systems often infer mastery from learners' correct actions.
 - Learners may act correctly without knowing when to withhold that action.
 - Action-focused adaptive systems may miss such overgeneralization.
- What does this paper add?
 - It provides evidence of deceptive overgeneralization in Riichi Mahjong experiments.
 - It offers a cross-domain illustration from K-12 decimal learning.
 - It shows action-focused adaptivity can stop practice too early.
 - It shows tailored 'do-not-act' practice reduced misapplication errors.
- Implications for practice or policy
 - Treat streaks of correct actions as insufficient evidence of conditional understanding.
 - Include practice tasks that require learners to *withhold* the learned action.
 - Add these tasks before adaptive systems stop practice based on mastery estimates.
 - Use targeted feedback that names the missing application constraint.

monitor their learning and evaluate their problem-solving strategies and their current understanding (National Research Council 2000). The positive effects of feedback are well supported by a wealth of evidence (Shute 2008). Feedback is most effective when it clearly highlights discrepancies between a learner's current performance and the desired outcome while offering actionable guidance to help learners meet specific target criteria (Ambrose et al. 2010, 139). ITSs embody these best practices of feedback by detecting and diagnosing observable discrepancies between expected and actual actions at each step. With a developed cognitive model, a cognitive tutor employs *model tracing* to compare learner actions at each problem-solving step against the possible actions generated by the cognitive model in order to provide individualized, just-in-time learning support tailored to learners' specific approach to a problem (Koedinger and Corbett 2006, 142). In the task-loop, ITSs use *knowledge tracing* algorithms such as Bayesian Knowledge Tracing (BKT) (Corbett and Anderson 1994) to update mastery estimates from step-level correctness and adapt subsequent practice. When the system reaches a high mastery estimate, often above a predefined threshold such as 95% (Koedinger and Corbett 2006, 144), it stops assigning tasks for that skill, balancing practice time against overpractice. Design-loop adaptivity uses learning data for instructional (re)design before and between ITS iterations (Alevan, McLaughlin, et al. 2016, 526).

However, adaptivity inherently depends on the completeness and accuracy of the performance data (Alevan, McLaughlin, et al. 2016, 523); limitations in those data can therefore make adaptive decisions ill-informed. In this paper, we focus on

task-loop adaptivity, specifically mastery inference and mastery stopping rules based on whether learners take the expected actions. The concern we examine is neither an inherent flaw of adaptive learning nor BKT itself, but a limitation of an assessment design pattern used to generate performance data and drive adaptive decisions: they may present only scenarios where the action is always appropriate, assessing whether learners can execute the expected action, without also testing whether learners know when that action should be withheld. This gap is critical in cases of a learning phenomenon we coin *deceptive overgeneralization*.

The definition and theoretical modelling of deceptive overgeneralization are grounded in Adaptive Control of Thought-Rational (ACT-R) (Anderson 1992; Anderson 1996) and the Knowledge-Learning-Instruction (KLI) framework (Koedinger et al. 2012). Deceptive overgeneralization occurs when a learner acquires a partial understanding of a skill—retaining an incomplete subset of the necessary conditions while omitting critical application constraints—yet still produces correct actions. This learning phenomenon is 'deceptive' because it creates an illusion of mastery—learners appear to perform the expected action, which may foster inflated beliefs about competence despite incomplete understanding (Bjork et al. 2013). For adaptive learning systems that rely on observed actions to infer mastery, deceptive overgeneralization poses a significant challenge: it can mislead adaptive systems into prematurely stopping necessary practice, leaving overgeneralization unaddressed.

1.1 | Theoretical Model: Using ACT-R and KLI to Characterize Deceptive Overgeneralization

1.1.1 | ACT-R

ACT-R, a cognitive architecture for understanding and modelling human cognitive processes, posits that cognitive behaviours are orchestrated by productions (Anderson 1992; Anderson 1996). A production can be represented as a condition-action pair (Anderson 1983, 5), with the *condition* part specifying the circumstances under which the production can apply, and the *action* part specifying what should be done when the production applies (Anderson 1982, 3). Empirical studies underpinning ACT-R have led to a proliferation of ITSs that successfully enhance learning outcomes across diverse educational settings (Alevan, McLaren, et al. 2016; VanLehn 2006; Woolf 2010).

1.1.2 | The KLI Framework

Learning is typically characterized by a gradual and continuous process rather than sudden transformative insights (Anderson and Schunn 2013). The KLI framework views learning as the acquisition of Knowledge Components (KCs), which are acquired units of cognitive functions or structures (Koedinger et al. 2012). The KLI framework identifies *induction and refinement* as one primary type of learning process, particularly for acquiring KCs associated with variable conditions: for KCs with conditions that can vary in form or

value, learners must induce and refine KCs so that the acquired KCs are ‘accurate, appropriately general, and discriminating’ (Koedinger et al. 2012). As we consider the induction and subsequent refinement of a KC as a continuous learning progression, learners may initially acquire an inaccurately generalized version of the target KC. This initial misunderstanding may either be refined into an accurate KC through further practice, or it may persist as inaccurate due to a lack of practice opportunities that support the refinement process.

1.1.3 | Modelling of Deceptive Overgeneralization

According to KLI, a KC links problem features to a corresponding response (Koedinger et al. 2012). A KC is considered accurate, or has ‘high feature validity,’ if it accounts for all relevant features while excluding irrelevant ones (LearnLab 2011). Otherwise, the acquired KC is inaccurate and requires further refinement. Inaccurate generalization can take the form of overgeneralization, undergeneralization, or a nuanced mix of both. Among these, deceptive overgeneralization is the specific phenomenon we investigate in this study.

To model the schema of deceptive overgeneralization, we build upon how production rules are represented in ITSSs. In ITSSs, particularly those developed using the Cognitive Tutor Authoring Tools (CTAT) (Aleven et al. 2006), each production’s condition-action pair is structured as an IF-THEN statement (Koedinger et al. 1997): IF <condition(s)> THEN <action>. Overgeneralization occurs when a learner acquires a production rule with an overly broad IF condition. We identify a schema of production rules prone to deceptive overgeneralization:

```
IF all required conditions are met THEN perform action
```

A learner exhibiting deceptive overgeneralization may omit part of the necessary constraints and instead acquire the following:

```
IF a subset of required conditions are met THEN perform action
```

For example, consider a production rule requiring two conditions to be met simultaneously to warrant an action:

```
IF condition1 AND condition2 are both met THEN perform action
```

A learner exhibiting deceptive overgeneralization may instead acquire:

```
IF condition1 is met THEN perform action
```

This modelling reveals a critical issue: when instructional designs present only scenarios where the action is warranted and assess learners solely on their ability to perform it, they risk leaving deceptive overgeneralization unaddressed. Adaptive systems risk exacerbating this by inferring mastery from correct performance, potentially halting practice prematurely.

1.1.4 | Knowledge Compilation and Deceptive Overgeneralization

A natural question arises: why not treat the verification of each condition as a separate KC? The answer lies in the expectation of task efficiency as learners develop expertise. As learners advance, they must handle increasingly complex tasks with greater fluency, necessitating compiled knowledge structures rather than independently verifying each condition. According to the KLI framework (Koedinger et al. 2012), KCs describe mental processes at approximately the unit task level within Newell’s time scales of human action (Newell 1994), typically taking about 10s. If each condition were treated as an independent KC, decision-making would become fragmented, disrupting fluency and reducing efficiency in complex tasks.

Knowledge compilation, a core mechanism in ACT-R, improves cognitive efficiency by eliminating multiple production rule firings (Anderson 1992, 169). A primary compilation process, known as *composition*, merges a chain of production rules into a single ‘macro-production,’ eliminating intermediate cognitive actions while maintaining the same final output (Anderson 1983, 235). Although composition increases overall efficiency, the composed ‘macro-productions’ tend to grow larger, particularly with an increase in the size of the condition sides (Anderson 1983, 239). With an increasingly complex and composite condition side, it becomes more likely that some conditions will be overlooked, leading to deceptive overgeneralization. Because knowledge compilation is a fundamental process in expertise development, deceptive overgeneralization is likely just as prevalent.

Because our empirical studies focus on Riichi Mahjong, a domain that may be unfamiliar to many readers, we first use an aviation case as an accessible worked example of the same theoretical mechanism. The case requires little domain-specific knowledge to follow, yet it starkly illustrates how an efficient compiled rule can become harmful when a critical application constraint is omitted. Specifically, the Crossair Flight 498 crash demonstrates how deceptive overgeneralization can be formed during expertise development and can lead to severe consequences. The official investigation report identified one human factor probable cause as follows: ‘when interpreting the attitude display instruments under stress, the commander resorted to a reaction pattern (heuristics) which he had learned earlier’ (Aircraft Accident Investigation Bureau 2002, 10).

As illustrated in Figure 1, in Soviet-designed attitude displays, the horizon remains fixed and the aircraft symbol rotates (a ‘third-person’ view); a counter-clockwise rotation indicates the aircraft is rolling left, and the correct stabilization response is to rotate right. This can be represented as a sequence of smaller production rules and then compressed through knowledge compilation, as shown in Algorithm 1, which highlights several benefits of *knowledge compilation*: fewer conditions and actions, fewer variables to track, and the elimination of subgoals. These optimizations enhance the efficiency of the macro-production compared to the original series of separate productions (Anderson 1982, 35).

Importantly, the accurate macro-production preserves the display-type constraint that it applies *only to Soviet displays*:

IF [Current Goal] The goal is to stabilize an airplane

AND [Application Constraint] The display is a Soviet-designed attitude display

AND [Perceptual Cue] The attitude display rotates counter-clockwise

THEN [Single Action] Rotate the airplane right

The overgeneralized version, however, omits this application constraint, allowing a response that was successful for Soviet displays to be misapplied to Western displays. As illustrated in Figure 2, Western attitude displays show a ‘first-person’ view where the horizon rotates relative to a fixed aircraft symbol. Therefore, a counter-clockwise rotation indicates the aircraft is rolling *right*—yet the overgeneralized composed rule would incorrectly instruct a right rotation, exacerbating the roll. This is precisely what occurred in the Crossair Flight 498 crash: after

20 years of flying exclusively Soviet aircraft, the commander had compiled the overgeneralized rule without ever encountering a ‘do-not-act’ practice opportunity that would have revealed the missing constraint (Aircraft Accident Investigation Bureau 2002). Despite over 8000 h of experience, the commander resorted to the overgeneralized heuristic under stress, with catastrophic consequences.

Why did the commander choose the overgeneralized shortcut over the correct sequence of basic productions? ACT-R’s conflict resolution mechanism provides an explanation: production strength reflects the frequency of successful past applications (Anderson 1983, 250). Over decades of flying Soviet aircraft, the overgeneralized macro-production consistently led to correct actions, increasing its strength and making it the preferred choice during conflict resolution (Anderson 1983, 133). The demanding nature of the emergency further favoured the efficient, albeit fatally overgeneralized, macro-production over a slower sequence of basic productions.

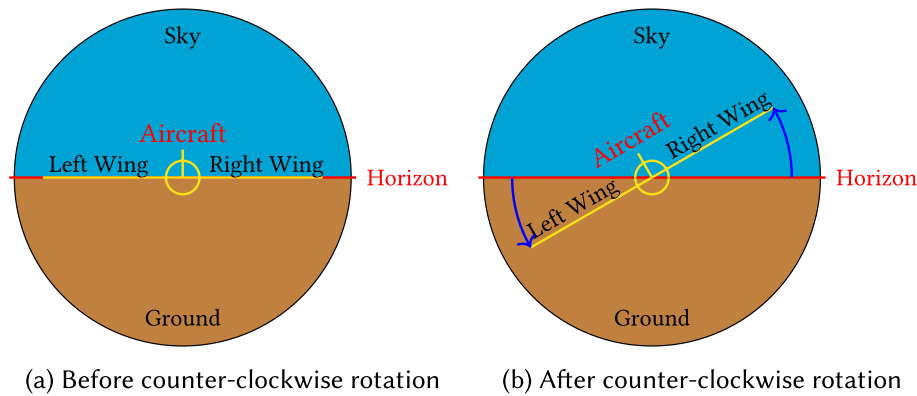


FIGURE 1 | A simplified depiction of a Soviet attitude display. The display reflects a ‘third-person view’, where the horizon stays fixed, and the airplane’s position is shown relative to the horizon. A counter-clockwise rotation (of the airplane relative to the horizon) indicates that the airplane is rolling left.

ALGORITHM 1 | Knowledge Compilation for Interpreting Attitude Display to Stabilize an Airplane.

Rule 1:

Condition: goal == stabilizeAirplane AND rollDirection == unknown

Action: subgoal = identifyRollDirection

Rule 2:

Condition: subgoal == identifyRollDirection AND displayType == Soviet AND displayRotation == counterClockwise

Action: rollDirection = left

Rule 3:

Condition: goal == stabilizeAirplane AND rollDirection != unknown

Action: subgoal = recoverAttitude

Rule 4:

Condition: subgoal == recoverAttitude AND rollDirection == left

Action: rotateAirplane(right)

Compiled macro-production:

Condition: goal == stabilizeAirplane AND displayType == Soviet AND displayRotation == counterClockwise

Action: rotateAirplane(right)

Efficiency gain from knowledge compilation:

Four production firings are reduced to one macro-production, eliminating the need to maintain separate subgoals, intermediate variables, and intermediate cognitive actions.

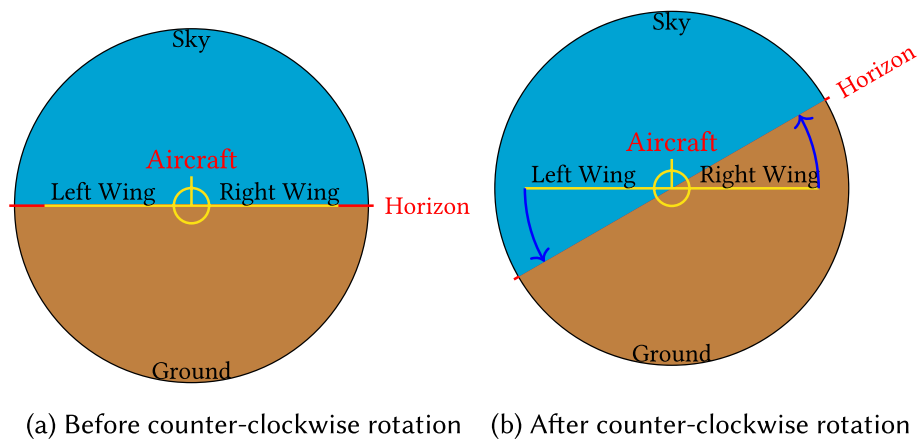


FIGURE 2 | A simplified depiction of a Western attitude display. The display reflects a ‘first-person view’, where the airplane stays fixed and the horizon rotates relative to the airplane. A counter-clockwise rotation (of the horizon relative to the airplane) indicates that the airplane is rolling right.

More broadly, the aviation case illustrates the domain-agnostic nature of the risk of forming deceptive overgeneralization during knowledge compilation: critical application constraints can be omitted as learners compile efficient macro-productions, especially when instruction provides few opportunities to practice discriminating when a rule should *not* be applied. We examine this risk empirically in Riichi Mahjong and later discuss its cross-domain relevance through a retrospective analysis of K–12 decimal learning.

1.2 | Positioning Relative to Related Concepts

1.2.1 | Deceptive Overgeneralization vs. Negative Transfer

Deceptive overgeneralization is related to negative transfer, but the two have distinct instructional implications. Negative transfer refers to cases in which prior learning interferes with performance in a new context (National Research Council 2000; Perkins and Salomon 1992). A classic related example is the Einstellung effect, in which repeated success with a familiar solution procedure biases later problem solving toward reusing that procedure even when a more direct or efficient response is available (Luchins 1942).

In their influential review of transfer, Perkins and Salomon argued that ‘Negative transfer typically causes trouble only in the early stages of learning a new domain. With experience, learners correct for the effects of negative transfer’ (Perkins and Salomon 1992). Our theoretical model, however, suggests that this expectation may not hold when learners acquire an apparently correct but overgeneralized production that continues to compete with later knowledge. In such cases, experience alone may be insufficient unless instruction explicitly refines the production’s application conditions. The Crossair Flight 498 case illustrates this point: the commander had over 8000h of flight experience yet still relied on an overgeneralized rule when the missing application constraint became critical. Thus, experience alone does not guarantee correction. Detecting and remedying

deceptive overgeneralization may require targeted intervention that assesses whether learners know when to withhold an otherwise appropriate action.

1.2.2 | Deceptive Overgeneralization vs. Shallow Problem Features

Furthermore, this study differs from prior research that has primarily focused on distinguishing between shallow and deep features in learning. Shallow features, also known as superficial or surface features, are those that do not contribute to correct solution pathways (Chang 2006; Chi et al. 1981; Ross 1984). For example, a learner might choose to use a pie chart because the data is demographic (shallow and incorrect) rather than categorical (deep and correct) (Chang 2006). In contrast, this research investigates scenarios where learners take correct actions based on a *relevant yet incomplete* set of features. Importantly, unlike shallow features, all the features used by learners exhibiting deceptive overgeneralization are part of the correct solution pathways. As a result, if learners are presented and assessed only on scenarios in which the action is warranted, deceptive overgeneralization may be behaviorally indistinguishable from accurate understanding, even though learners have not acquired the full set of application constraints.

1.3 | Research Questions

The research questions translate the theoretical argument above into an empirical test using ITSs with task-loop adaptivity. Task-loop adaptivity is neither a precondition for deceptive overgeneralization nor the source of its underlying cognitive mechanism. Rather, task-loop adaptivity provides a tractable setting for observing one instructional consequence of the mechanism: when learners repeatedly succeed on practice items that do not test a missing application constraint, correctness-based mastery inference may prematurely deem mastery and stop practice before learners encounter cases in which the practiced action should *not* be applied.

We chose Riichi Mahjong as the study domain, a complex decision-making domain rich in composite-condition skills with single observable actions—a structural configuration that, according to our theoretical modelling, is particularly prone to deceptive overgeneralization. The general effectiveness of ITSs has been well established in prior research (Aleven, McLaren, et al. 2016; VanLehn 2006; Woolf 2010). However, to the best of our knowledge, we are the first to apply ITSs to teach Riichi Mahjong. Accordingly, rather than revisiting the broader claim that ITSs can be effective across domains, we first establish a study-internal baseline showing that the specific ITSs used in this study produced measurable learning gains. This baseline is necessary before interpreting misapplication patterns as evidence of deceptive overgeneralization, and as evidence that correctness-based adaptive stopping rules may prematurely end practice rather than as a consequence of general instructional inadequacy.

Building on this foundation, we then examine whether deceptive overgeneralization emerges when learners acquire complex skills with composite conditions, whether correctness-based adaptive stopping rules risk leaving such overgeneralization undetected, and whether a targeted detection and remediation procedure can reduce systematic misapplication.

To this end, we formulate the following research questions:

RQ1. As a study-internal baseline, do our Riichi Mahjong ITSs produce measurable learning gains on the targeted complex skills?

RQ2. Does deceptive overgeneralization occur when learning complex skills with composite conditions?

RQ3. Does adaptive learning risk prematurely stopping practice, leaving deceptive overgeneralization unaddressed?

RQ4. Can tailored ‘do-not-act’ practice designed via our proposed procedure effectively address deceptive overgeneralization?

1.4 | Overview and Contributions

To investigate deceptive overgeneralization, we conducted a series of experiments using ITSs developed with the CTAT, designed to teach Riichi Mahjong, a domain rich in composite-condition, single-action skills.

Experiment 1 evaluated whether ITSs improve learning outcomes for complex skills in this domain. Experiment 2 tested whether learners exhibit deceptive overgeneralization by misapplying an acquired action on the first ‘do-not-act’ detector item, where one or more required application constraints were violated. Experiments 3–11 replicated this observation under task-loop adaptivity across multiple skills and cohorts, and tested whether correctness-based mastery stopping rules can exacerbate deceptive overgeneralization by terminating practice before learners encounter detector items that require refraining from the learned action. Within Experiments 3–11, a subset of experiments incorporated tailored remediation (Experiments 6, 7, 9, 10

and 11) by introducing a short sequence of targeted ‘do-not-act’ practice problems immediately after the initial misapplication, designed in accordance with our systematic detection/remediation procedure.

This article makes two contributions: (1) empirical evidence of deceptive overgeneralization and its implications for correctness-based task-loop adaptivity, and (2) an instructional design procedure to systematically detect and remedy this phenomenon. Across multiple skills and cohorts over multiple semesters, we show that learners can appear competent during practice yet later misapply learned actions systematically in inappropriate contexts, and that adaptive systems relying solely on correctness risk stopping practice prematurely. We then propose and validate a procedure that helps educators detect and remedy deceptive overgeneralization when teaching complex skills. To demonstrate how the framework can be applied beyond the study domain, we also use an exploratory secondary analysis of large-scale K–12 decimal-learning data from *Decimal Point* to provide a mechanism-level account of whole-number bias, a well-documented misconception in decimal learning.

Together, these contributions position deceptive overgeneralization as a theoretically grounded framework for identifying cases in which observed correctness masks incomplete conditional understanding. The framework is also practically actionable and empirically testable: it is operationalized through a clearly defined four-step design procedure for creating concrete ‘do-not-act’ detector items and targeted remediation practices, and we evaluate this procedure across multiple skills and cohorts, yielding evidence that it can reveal systematic misapplication and support refined understanding through targeted remediation.

2 | Materials & Methods

2.1 | A Procedure to Systematically Detect and Remedy Deceptive Overgeneralization

Addressing deceptive overgeneralization requires designing tailored activities that test whether students have generalized their learning appropriately. Grounded in our theoretical modelling, we propose and validate the following procedure in this study:

1. *Represent skills as production rules*—Model skills in the format of production rules (i.e., condition-action pairs).
2. *Identify production rules with composite conditions and single actions*—Focus on skills where learners must evaluate multiple conditions but produce a single observable action, as these are particularly susceptible to deceptive overgeneralization.
3. *Prioritize skills where misapplication is particularly problematic*—identify rules where overgeneralization could lead to serious errors.
4. *Create tailored ‘do-not-act’ practice where learned actions should not be taken*—design problems that intentionally violate some necessary conditions to detect deceptive

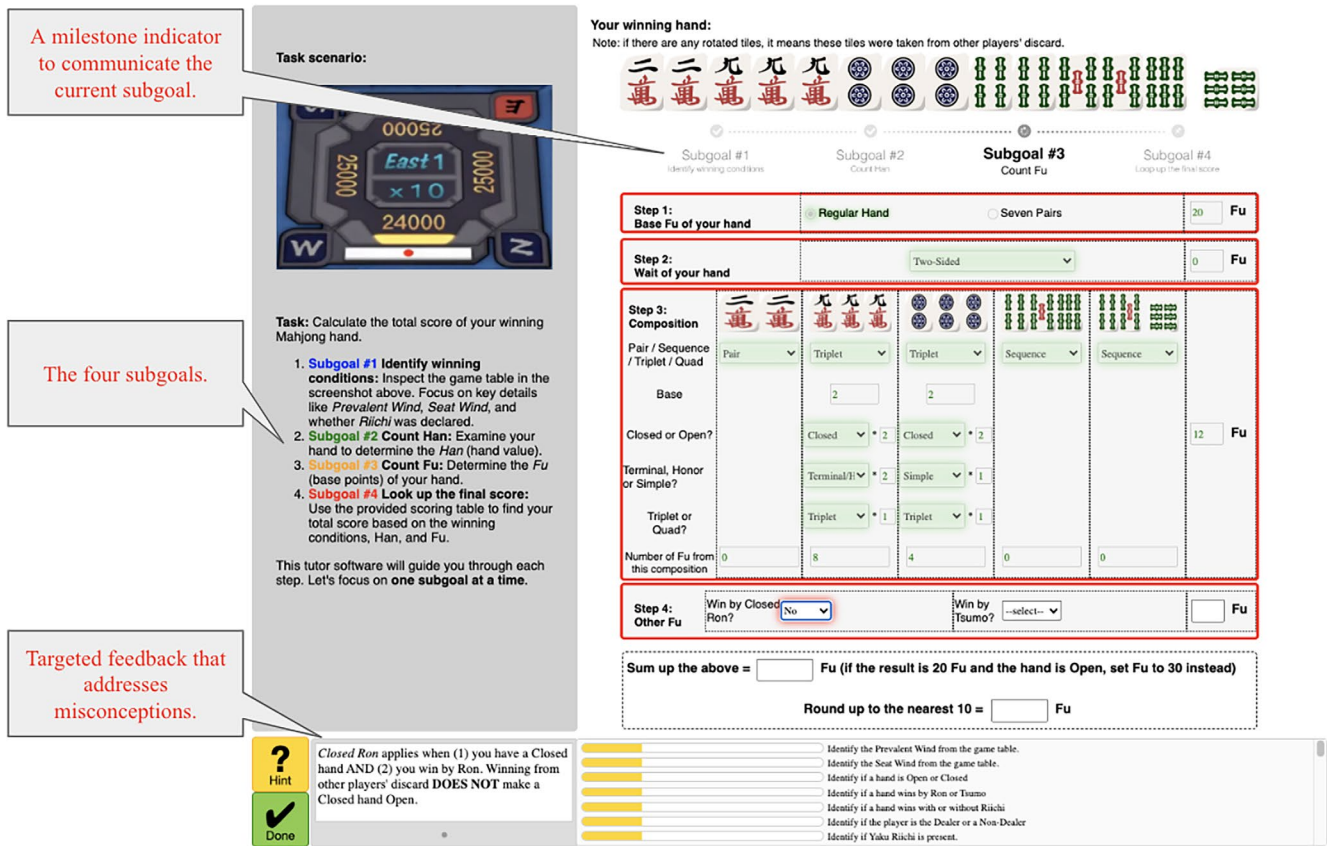


FIGURE 3 | The Riichi Mahjong score calculation ITS. The tutor guides learners through the systematic computation of a winning hand's score. Key instructional features are highlighted in the annotated interface.

overgeneralization, paired with targeted feedback that directly addresses the missing constraint to help learners refine their understanding.

From a regulative-loop perspective (VanLehn 2016), this procedure introduces an explicit task-level regulation mechanism: learner performance is evaluated against a target standard that includes both correct action execution and appropriate constraint discrimination, and targeted intervention is triggered when misapplication is detected.

2.2 | Task Domain

To investigate this phenomenon in a controlled yet ecologically valid setting, we require a task domain that inherently presents challenges conducive to deceptive overgeneralization. Specifically, the ideal domain should involve production rules with complex conditions but a single observable action, as we hypothesize that this structure is particularly susceptible to deceptive overgeneralization. We selected Riichi Mahjong because it naturally exhibits these characteristics.

Riichi Mahjong, also known as Japanese Mahjong, is an incomplete-information intellectual game that requires complex cognitive processes. Its strategic depth has attracted considerable interest from AI researchers, leading to models that outperform most human players (Li et al. 2020; Mizukami and Tsuruoka 2015). However, these AI models are optimized for

gameplay rather than learning, lacking explanatory features to support skill acquisition. This gap highlights the need for alternative approaches to help human players develop expertise. In addition, learning Riichi Mahjong presents unique challenges: optimal moves may yield unfavourable outcomes due to randomness, while incorrect moves can sometimes succeed. This inconsistency between action and outcome disrupts natural learning, as players cannot rely on in-game outcome as accurate feedback, and players may misattribute the quality of their decisions based on results rather than the decision-making process (Baron and Hershey 1988). Furthermore, since each game begins with a randomized hand and evolves through the dynamic interactions of four independent players, structured repeated practice is difficult to achieve. Given these challenges, we hypothesize that ITSs are particularly beneficial in this domain, as they can create controlled yet realistic scenarios, integrating targeted practice, immediate feedback and adaptive problem sequencing to support learning. Meanwhile, as this domain features production rules with composite conditions and single actions, it is particularly suitable for investigating deceptive overgeneralization.

2.3 | Materials

As shown in Figures 3–6, we developed a series of ITSs for Riichi Mahjong using CTAT. These ITSs incorporate key features that provide learners with favourable learning conditions, as synthesized by Koedinger et al. (2023). These features

Riichi Mahjong Intelligent Tutoring System Series: *Pinfu*

Task scenario:



Task: Follow the criteria below to determine whether the provided hand qualifies for the Yaku *Pinfu*.

- Requirement #1** The hand is closed. (An open hand does not qualify for *Pinfu*.)
- Requirement #2** The hand consists of 1 pair and 4 sequences. (Any triplets or quads give extra Fu, which disqualifies the hand for *Pinfu*.)
- Requirement #3** The pair is not composed of *Yakuhai*. (A pair of *Yakuhai* gives 2 Fu, which disqualifies the hand for *Pinfu*.)
- Requirement #4** The hand has a two-sided wait. (Inside, edge, or pair waits give 2 Fu, which disqualify the hand for *Pinfu*.)

Your hand:



Does this hand have Yaku *Pinfu*? ☐ Yes ☒ No ☐ Uncertain

?

✓

What are the prevalent wind and seat wind based on the game table provided in the scenario? Is the wind forming the pair considered *Yakuhai*?

Identify if a hand qualifies for Yaku *Pinfu*.

FIGURE 4 | The *Pinfu* ITS. Formally, identifying *Pinfu* entails systematic hand decomposition and Fu evaluation of meld composition, wait type, and pair. Experts compile this process into condition-based macro-productions that enable fast-track acceptance and rejection; however, when novices adopt these compiled rules, systematic misapplication can occur if conditions are incompletely evaluated.

Riichi Mahjong Intelligent Tutoring System Series: *Defense*

Our hand cannot compete, and it is time to play defensively. But how?

Instructions: Follow the criteria below to determine whether a numeric tile *N* is relatively safe to discard due to Kabe.

- Requirement #1** Two Kabe (*K1* and *K2*, where $K1 < K2$) must be visible. (Note: Kabe *K1* can be 0 and *K2* can be 10, representing the "virtual" Kabe at the two ends which "always exist.")
- Requirement #2** The tile *N* must be between *K1* and *K2* (i.e., $K1 < N < K2$).
- Requirement #3** The difference between *K1* and *K2* must be ≤ 3 . (If $K2 - K1$ is greater than 3, the possibility of forming sequences remains.)
- Requirement #4** The two Kabe block all possible sequences for *N*, making it impossible to use *N* in a valid sequence, while tiles outside the *K1-K2* range remain vulnerable to forming a sequence for your opponent in question.

Game Table:



Which tile is relatively safe to discard due to *Kabe*?

Type your answer using the mpsz notation format, e.g., 1m, 2p, 3s, and press the 'Enter/Return' key. If no such tile exists, enter "NA" instead (which means "Not Applicable")

?

✓

Indeed, 3s is Kabe. But does the impossibility of forming sequences increase the relative safety of 2s? The player with a ready hand has a hand of All Triplets Single Wait which can only wait for forming a pair.

FIGURE 5 | The *Kabe* ITS. The provided scenario, based on an anonymized real game, does NOT warrant applying the KC. Learners who misapply the KC due to deceptive overgeneralization receive targeted feedback designed to refine understanding.

include immediate feedback, context-specific instruction on demand, enforcement of correct responses, tailored tasks designed through cognitive task analysis (Schraagen et al. 2000), and repeated practice opportunities to ensure mastery.

An important consideration in studies that aim to observe learner mistakes is whether the observed phenomenon reflects authentic learning processes or merely results from experimental 'traps.' In our design, we made considered efforts to maintain

Riichi Mahjong Intelligent Tutoring System Series: Riichi Judgement

Game Settings

A Tenhou/MahjongSoul mid-to-high-level ranking *Hanchan* (i.e., which starts with East 1 and ends with South 4) game.
Ranking points are primarily based on final placement.
For simplicity, assume the ranking point bonus or penalty per final placement is as follows: 1st: +100, 2nd: +50, 3rd: 0, and 4th: -200.
Notably, as a significantly negative-sum game, finishing in 4th place results in a substantial ranking point deduction.

Game Scenario:



? Hint
✓ Done

Previous
Next

Step 1: Evaluate game situation

Game Round
Is this the last round of the game (All Last)? ☒ Yes ☐ No ☐ Uncertain

Current Placement: ☐ 1st Place ☒ 2nd Place ☐ 3rd Place ☐ 4th Place

Placement Goal: ☒ Aim for 1st ☐ Avoid 4th ☐ Uncertain

Step 2: Calculate hand value

How many points are needed to achieve the placement goal? That is, the score gap between you and the other player that you aim to overtake. 1000

At least how many *Han* would this hand have upon winning? Please type your answer and press the 'Enter' key. 2

The hand value with *Dama*: 2000

Is the current hand value sufficient without *Riichi*? ☒ Yes ☐ No ☐ Uncertain

Step 3: Weigh hand value against win rate

Which of *Riichi* or *Dama* has a higher hand value? ☒ Riichi ☐ Dama ☐ Uncertain

Given this scenario, which of *Riichi* or *Dama* has a higher win rate? ☐ Riichi ☒ Dama ☐ Uncertain

Based on the game situation, placement goal, and current hand value, which should be prioritized? Value or Win Rate? ☐ Value ☒ Win Rate ☐ Uncertain

Step 4: Make final decision

Riichi, or Dama?

☐ Riichi

☒ Dama

☐ I do not know

Summary of one general heuristic for Riichi judgement

IF it's All Last AND the hand has Yaku even with Dama AND winning the hand would result in achieving 1st place THEN choose Dama to maximize the win rate.

FIGURE 6 | The Riichi judgement ITS. It aids students in acquiring complex macro-productions formed via knowledge compilation. While it provides detailed steps to help learners understand how production rules are compiled into the final macro-production, the macro-production only produces a single observable action (as shown in the *Make Final Decision* section of the ITS interface, where the action is either *Riichi* or *Dama*).

ecological validity by grounding all problem scenarios in *real* gameplay. All the skills involved in these experiments—such as score calculation, Pinfu, Kabe, and Riichi judgement—are commonly taught skills featured in Mahjong textbooks (Chiba 2026). The macro-productions and their commonly overgeneralized variants were not arbitrarily invented but derived through cognitive task analysis involving both experts and novices.

2.4 | Study Overview

Table 1 reports the purpose, semester, skill, use of task-loop adaptivity, use of tailored remediation, and sample size for each experiment.

2.5 | Participants

We conducted multiple in vivo experiments across five semesters in a course on Riichi Mahjong at Carnegie Mellon University. Students enrolled voluntarily based on interest and did not receive course grades that affected their Grade Point Average (GPA). Participation in the study was voluntary. Students could

decline or withdraw without penalty, and completion or performance on the ITS activities did not affect academic standing. Participants were included in the analysis if they provided consent and completed the relevant activities.

2.6 | Procedure

2.6.1 | Common Procedure Across Experiments

Across semesters, participants began the assigned ITS problem sets during designated class sessions. If they did not finish within the session, they completed the remaining problems outside of class. The tutors provided immediate step-level feedback, including correctness verification and targeted feedback when responses were incorrect, along with hint-on-demand support. All interactions were logged by DataShop (Koedinger et al. 2010). Logged data included time-stamped student actions at each step (e.g., selected responses, submitted entries), correctness outcomes, hint requests and hint levels, feedback messages delivered by the tutor, and KC mappings associated with each step. Low-level interface behaviours such as mouse movements, cursor trajectories, or screen navigation events were not recorded.

TABLE 1 | Overview of the experiments.

Experiment	Semester	Purpose	Skill	Task-Loop adaptivity	Tailored remediation	N
1	S24–S26	Established study-internal baseline (RQ1)	Score calculation	N/A	N/A	54
2	F24	Tested deceptive overgeneralization with a non-adaptive baseline (RQ2)	Pinfu	No	No	19
3	F24	Tested deceptive overgeneralization with task-loop adaptivity (RQ2–RQ3)	Kabe	Yes	No	11
4	F24	Tested deceptive overgeneralization with task-loop adaptivity (RQ2–RQ3)	Riichi judgement	Yes	No	13
5	S25	Tested deceptive overgeneralization with task-loop adaptivity (RQ2–RQ3)	Pinfu	Yes	No	17
6	S25	Evaluated tailored remediation with task-loop adaptivity (RQ2–RQ4)	Kabe	Yes	Yes	17
7	S25	Evaluated tailored remediation with task-loop adaptivity (RQ2–RQ4)	Riichi judgement	Yes	Yes	13
8	F25	Tested deceptive overgeneralization with task-loop adaptivity (RQ2–RQ3)	Pinfu	Yes	No	13
9	F25	Evaluated tailored remediation with task-loop adaptivity (RQ2–RQ4)	Kabe	Yes	Yes	13
10	F25	Evaluated tailored remediation with task-loop adaptivity (RQ2–RQ4)	Riichi judgement	Yes	Yes	8
11	S26	Evaluated tailored remediation with task-loop adaptivity (RQ2–RQ4)	Pinfu	Yes	Yes	14

2.6.2 | Effectiveness of ITS in Learning Complex Skills (RQ1)

To evaluate the efficacy of ITSs in teaching complex skills, we employed a single-group pretest-posttest design in Experiment 1. This experiment was implemented on a rolling basis across five semesters (Spring 2024–Spring 2026) using the same design and instruments, and we report both cohort-level and pooled analyses. Two distinct but equivalent 10-question quizzes were used for the pre- and post-tests, with no feedback provided during the pre-test. This design aimed to mitigate biases, ensuring differences in learning outcomes stemmed from the ITS intervention rather than practice effects from the pre-test.

2.6.3 | Deceptive Overgeneralization and Premature Stopping Under Adaptive Learning (RQ2 and RQ3)

We begin our investigation of deceptive overgeneralization with a non-adaptive baseline (Experiment 2). In this baseline

experiment, all learners completed an identical fixed sequence of problems, guaranteeing exposure to the first detector item at the same point in the sequence, where the correct response is to *refrain* from the learned action.

As discussed in Section 1.1.3, knowledge compilation supports fluency by compressing multi-step reasoning into macro-productions. However, this same process creates the structural risk that necessary application constraints are omitted from the compiled rule. Experiment 2 therefore examines whether learners form inaccurate macro-productions during this compilation process.

The problem set was constructed to target a macro-production derived from a chain of detailed steps practiced in Experiment 1. The 15-problem sequence consists of two phases. Problems 1–9 provide repeated opportunities in which the target action is warranted, encouraging induction of a compiled rule without yet requiring discrimination among all necessary constraints. Problems 10–15 introduce detector items in which one or more constraints are violated, requiring learners to determine

whether the action remains appropriate. In this design, misapplication on the first such detector item operationalizes deceptive overgeneralization as applying the learned action when one or more required conditions are not satisfied.

To test whether learners exhibit deceptive overgeneralization (RQ2), we employ a two-stage diagnostic analysis.

First, we examine whether performance on Problem 10 deviates from the prediction of a standard knowledge-tracing model. We use BKT (Corbett and Anderson 1994) as the baseline model. BKT dynamically updates its estimate of a learner's skill mastery based on observed performance and assumes a binary latent state for each KC, such that a learner is either in an *Unlearned* or *Learned* state (Corbett and Anderson 1994, 260). If learners have accurately generalized the KC during Problems 1–9, the observed error rate on Problem 10 should align with the rate predicted by BKT under this two-state assumption. However, if the observed error rate significantly exceeds the prediction, this constitutes a violation of the two-state model. Formally, we test the following hypotheses:

Null Hypothesis (H0). *The error rate on Problem 10 matches the rate predicted by BKT based on performance in Problems 1–9.*

Alternative Hypothesis (H1). *The error rate on Problem 10 exceeds the BKT-predicted rate, indicating a violation of the two-state mastery assumption.*

Second, if such a deviation is detected, we determine whether it reflects the introduction of a new KC or systematic misapplication of an overgeneralized rule. To do so, we compare the observed error rate on Problem 10 to the expected error rate for a completely new KC under the same BKT parameterization. If the error rate matches the new-KC baseline, errors may be attributed to unfamiliarity. However, if the observed error rate significantly exceeds this baseline, it indicates systematic application of an incompletely specified macro-production rather than guessing, consistent with deceptive overgeneralization. The hypotheses for this second comparison are:

Null Hypothesis (H0). *The error rate on Problem 10 matches the expected rate for a new KC.*

Alternative Hypothesis (H1). *The error rate on Problem 10 exceeds the new-KC baseline, supporting the presence of deceptive overgeneralization.*

Experiments 3–11 were designed to (i) replicate the observation from Experiment 2 under task-loop adaptivity (RQ2) and (ii) test whether correctness-based mastery stopping rules can exacerbate deceptive overgeneralization by terminating practice before learners encounter ‘do-not-act’ cases (RQ3). Importantly, these experiments collectively provide replications across both cohorts (multiple semester offerings) and skills (multiple Riichi Mahjong KCs with composite conditions and single observable actions), allowing us to assess whether deceptive overgeneralization generalizes beyond a single cohort or KC. In these experiments, the ITSs employed BKT to dynamically update mastery estimates and determine whether

further practice was required. Once the system reached a pre-defined 95% mastery threshold—a commonly used value in CTAT-based tutors (Koedinger and Corbett 2006, 144)—it assumed mastery and stopped providing additional practice opportunities.

To evaluate whether adaptive learning prematurely ceased necessary practice, we employed a within-subjects design. Each participant completed two rounds of practice using the ITS on the same problem set:

First round (adaptive learning enabled): The ITS used the same ordered problem pool as the second round, including the ‘do-not-act’ detector items, but automatically terminated further problem assignment once it deemed the skill mastered; consequently, the system's stopping rule could end adaptive practice before those detector items were presented.

Second round (adaptive learning disabled to allow full exposure): The ITS presented all problems, ensuring participants encountered cases where previously learned actions were no longer applicable. If a skill was truly mastered when the ITS deemed so in the first round, practice in the second round should turn out to be unnecessary—learners should be able to correctly discern when an action should or should not be taken. However, we hypothesize that after adaptive learning ceases practice in the first round, learners will incorrectly apply the acquired action in the second round when exposed to scenarios where that action is inappropriate—revealing deceptive overgeneralization.

Across Experiments 3–11, we deployed ITSs targeting macro-productions hypothesized to be particularly susceptible to deceptive overgeneralization, following the schema discussed in Section 1.1.3. The corresponding skills investigated in each experiment are reported in Table 1.

Below, we present the KC from Experiment 3 as an illustrative example, as it contains relatively little Mahjong-specific jargon and partially resembles a mathematical problem with probabilistic reasoning; more importantly, it illustrates how knowledge compilation merges a chain of detailed steps into a composite-condition macro-production, increasing efficiency while also creating the structural conditions under which deceptive overgeneralization can emerge.

To provide necessary context for readers unfamiliar with Mahjong, we briefly outline the logic behind this KC before presenting its formal structure. This KC pertains to defensive play, where the goal is to determine whether discarding a tile carries minimal risk of helping an opponent complete their hand. In Mahjong, players form hands using pairs (two identical tiles), triplets (three identical tiles), and sequences (three consecutive tiles, e.g., 2-3-4). Since opponents' hands are hidden, players infer safer discards based on visible tiles. A tile is considered safer if it contributes to fewer possible patterns. For example, if all copies of 3 are visible and thus cannot be in an opponent's concealed hand, then an opponent cannot be waiting to complete a 1-2-3 or 2-3-4 sequence, making 2 safer. In contrast, a tile like 8 remains risky, as it could still complete multiple patterns (6-7-8, 7-8-9, a pair, or a triplet of 8s). Formally, the KC is structured as follows:

IF [Current Goal] The goal is to determine whether a numeric tile N is relatively safer to discard

AND [Application Constraint] The opponent in question could be waiting to form a 3-tile sequence

AND [Mathematical] All copies of tiles K_1 and K_2 ($K_1 < K_2$) are visible, meaning they cannot be in an opponent's concealed hand. (For boundary cases, K_1 can be treated as 0 and K_2 as 10, representing the conceptual 'outer limits' of the tile range)

AND [Mathematical] $K_2 - K_1 \leq 3$, preventing a 3-tile sequence from forming using tiles in between

AND [Mathematical] $K_1 < N < K_2$,

THEN [Single Action] Discard the tile N

In this case, the application constraint—the 'why'—is critical for determining whether the action is appropriate. The compiled rule is based on the probabilistic reasoning that a tile is conditionally safer if it cannot complete a sequence when the nearby tiles required to form that sequence have already run out, compared to other riskier tiles that can still complete one or more sequences. However, learners may overgeneralize this rule by focusing solely on the 'how'—discarding a tile so long as nearby tiles are visible—without fully understanding the 'why.' This could lead them to misapply the action in scenarios such as the problem presented in Figure 5. In that scenario, the opponent cannot be waiting on sequences in the first place, so invoking Kabe does not make the tile in question safer than other discard options. Therefore, the compiled composite-condition single-action rule should not be applied.

To replicate the tests conducted for RQ2, a similar comparison was applied in Experiments 3–11. Unlike Experiment 2, where BKT was used post-experiment to predict the expected error rate on Problem 10, Experiments 3–11 already employed BKT dynamically to determine when to stop providing practice. Given this setup, we calculate the expected error rate based on the mastery threshold of 95%. However, if deceptive overgeneralization is present, there will be a significant deviation from this prediction, as learners may systematically misapply an overgeneralized rule when exposed to scenarios that do not warrant the learned action in the second round.

To investigate RQ3, we examine whether adaptive learning prematurely stops necessary practice before learners encounter scenarios where the learned action is inappropriate. If learners exhibit high error rates when faced with these scenarios in the second round, yet BKT stopped practice in the first round before they had any opportunity to encounter such cases, this would suggest that adaptive learning risks amplifying the issue of deceptive overgeneralization by prematurely terminating practice.

2.7 | Effectiveness of Our Proposed Procedure in Addressing Deceptive Overgeneralization (RQ4)

To investigate RQ4, we examined the experiments that incorporated tailored remediation (Experiments 6, 7, 9, 10, and 11; Table 1). In these experiments, the tutor introduced a sequence of tailored 'do-not-act' practice problems immediately after a learner's first misapplication on the detector item (i.e., the first instance where the correct response was to refrain from the previously practiced action). These remediation problems were designed in accordance with the procedure in Section 2.1: they systematically violated certain necessary application constraints while holding other features constant, and they provided targeted feedback that explicitly named the missing constraint to support refinement of the macro-production. We operationalize remediation effectiveness as the reduction in error rates across successive remediation opportunities, with particular focus on performance on the final remediation item, where the correct response is to refrain from the previously learned action.

3 | Results

3.1 | Effectiveness of ITS in Learning Complex Skills (RQ1)

Table 2 summarizes pre–post performance by semester and overall. As shown, post-test means exceeded pre-test means in every cohort, with positive gains observed across all five semesters. Within-subject effect sizes ranged from $d_z = 0.74$ to 1.51.

To assess the statistical significance of these gains, we conducted paired analyses on the full sample ($N = 54$). Normality of the paired difference scores (post-test minus pre-test) was not rejected by a

TABLE 2 | Pre–post scores by semester.

Semester	N	Pre	Post	Gain	d_z
S24	11	4.82 (2.71)	6.82 (1.47)	2.00 (2.72)	0.74
F24	11	4.91 (3.05)	7.36 (1.86)	2.45 (2.25)	1.09
S25	14	4.00 (2.80)	6.36 (2.53)	2.36 (2.65)	0.89
F25	7	1.57 (2.44)	6.57 (2.99)	5.00 (3.32)	1.51
S26	11	2.91 (3.67)	5.73 (1.27)	2.82 (3.37)	0.84
Overall	54	3.81 (3.10)	6.56 (2.06)	2.74 (2.88)	0.95

Note: Values are mean (SD).

Shapiro–Wilk test ($W = 0.973$, $p = 0.261$). We therefore proceeded with a paired-samples t -test, which showed that post-test scores were higher than pre-test scores, $t(53) = 6.99$, $p = 4.73 \times 10^{-9}$, with a large within-subject effect size ($d_z = 0.95$). A Wilcoxon signed-rank test yielded a consistent conclusion ($W = 77.0$, $p = 1.45 \times 10^{-7}$).

To examine whether the pre–post gain pattern differed across cohorts, we used an ANCOVA framework predicting post-test score from pre-test score and semester (treated as a categorical factor). The homogeneity-of-regression-slopes assumption was tested by including a pre-test $\times$ semester interaction; the interaction terms were not statistically significant (all $p \geq 0.298$), indicating no evidence that the association between pre-test and post-test scores differed across semesters. We therefore proceeded with the additive ANCOVA model.

A Breusch–Pagan test conducted on the additive model indicated heteroscedasticity ($\chi^2(5) = 12.56$, $p = 0.028$), so we report heteroscedasticity-consistent (HC3) standard errors. In the additive model with robust standard errors, pre-test score was positively associated with post-test score ($\beta = 0.288$, robust SE = 0.100, $z = 2.88$, $p = 0.004$). After adjusting for pre-test performance, semester was not a significant predictor of post-test score (all $p \geq 0.141$), indicating no detectable semester-to-semester difference in learning gains in this sample.

Together, these results indicate substantial pre–post improvement on the targeted skills in the domain of Riichi Mahjong, with a large within-subject effect size. This pattern was observed consistently across five semester cohorts, and the ANCOVA analysis did not reveal statistically significant semester-to-semester differences in post-test outcomes after adjusting for pre-test performance.

3.2 | Observation of Deceptive Overgeneralization (RQ2)

Because several experiments involved small cohorts, we report Wilson 95% confidence intervals for key proportions. These intervals make the precision of the estimates visible, while the interpretation focuses on the repeated direction and magnitude of the pattern across skills and cohorts.

In Experiment 2, on Problem 10—the first scenario where the acquired action should not be taken—14 of 19 participants (73.7%) misapplied the action. To evaluate whether these errors stemmed from systematic overgeneralization, we simulated BKT based on performance in Problems 1–9. The following values were used for the BKT parameters: $P_{\text{learned}} = 0.25$ (initial mastery), $P_{\text{transfer}} = 0.2$ (learning rate), and $P_{\text{slip}} = 0.1$, as these are the default and commonly used values in CTAT. We set $P_{\text{guess}} = 0.5$ because the ITS provided three options: perform the action, refrain from the action, or select ‘I don’t know.’ Since selecting ‘I don’t know’ does not contribute to correctness, a learner making an intentional guess would randomly choose between performing or refraining from the action, each with an equal probability. Thus, the probability of guessing correctly is 50%. Under these conditions, the BKT model, fit using learners’

performance on Problems 1–9, predicted 2.95 collective errors on Problem 10.

However, the actual 14 errors significantly exceeded this expectation ($p < 0.001$), indicating a systematic deviation from the BKT two-state assumption. Next, we compared the observed errors to the expected error rate for a completely new KC. For a new KC under BKT assumptions: $P_{\text{error}} = P_{\text{learned}} \times P_{\text{slip}} + (1 - P_{\text{learned}}) \times (1 - P_{\text{guess}})$, with $P_{\text{learned}} = 0.25$, $P_{\text{slip}} = 0.1$, and $P_{\text{guess}} = 0.5$, the expected error rate $P_{\text{error}} = 0.4$.

With 19 participants, the expected collective errors for a new KC were 7.6. The actual errors (14) significantly exceeded this estimate ($p = 0.003$), suggesting that participants were not merely encountering a new KC but were systematically misapplying an overgeneralized macro-production. This distinction highlights that learners had acquired a systematic but flawed rule from Problems 1–9, which they applied inappropriately rather than making random guesses.

A similar pattern was observed in Experiments 3–11 (Table 3), where learners initially demonstrated seemingly good performance when performing actions, followed by starkly high error rates in overgeneralization detector problems, manifesting as systematically misapplying previously learned actions.

Under the BKT two-state assumption and the aforementioned parameterization, a 95% mastery threshold implies an expected error rate of 12% when accounting for slips (Corbett and Anderson 1994). For a completely new KC under the same parameterization, the expected error rate is 40% (as previously computed, $P_{\text{error}} = 0.4$). In contrast, the observed misapplication rates ranged from 61.5% to 100% across skills and cohorts. Notably, an astonishing error rate of 100% was consistently observed for the Kabe macro-production across three cohorts. Binomial tests indicated that misapplication significantly exceeded the 12% baseline in all experiments ($p < 0.001$). When compared against the more conservative 40% new-KC baseline, misapplication remained statistically significant in 9 of 10 experiments ($p < 0.01$). The sole exception was Pinfu in F25 (61.5%, $p = 0.156$), where the observed rate remained above the 40% baseline but did not reach statistical significance in this sample ($n = 13$).

In summary, the finding in Experiment 2 was replicated across Experiments 3–11: misapplication consistently exceeded the BKT mastery expectation and, in nearly all cases, also exceeded the new-KC baseline, providing consistent evidence of deceptive overgeneralization across cohorts and skills.

3.3 | Evidence That Adaptive Learning Risks Prematurely Stopping Practice in the Presence of Deceptive Overgeneralization (RQ3)

Given the observation of deceptive overgeneralization in RQ2, we examined whether BKT-based mastery stopping rules terminated practice before learners encountered detector items designed to reveal this phenomenon. Across Experiments 3–11,

TABLE 3 | Overgeneralization detector outcomes (RQ2). Misapplication is defined as applying the learned action on the first problem instance where the correct response is to *not* take that action. Bracketed values are Wilson 95% confidence intervals. Binomial tests compare observed rates against 12% (BKT accurate-mastery-with-slip expectation) and 40% (new-KC baseline).

Exp.	Semester	Skill	Task-loop adaptivity	Misapplication rate [95% CI]	<i>p</i> vs. 12%	<i>p</i> vs. 40%
2	F24	Pinfu	No	14/19 (73.7%) [51.2%, 88.2%]	< 0.001	0.004
5	S25	Pinfu	Yes	13/17 (76.5%) [52.7%, 90.4%]	< 0.001	0.005
8	F25	Pinfu	Yes	8/13 (61.5%) [35.5%, 82.3%]	< 0.001	0.156
11	S26	Pinfu	Yes	11/14 (78.6%) [52.4%, 92.4%]	< 0.001	0.003
3	F24	Kabe	Yes	11/11 (100.0%) [74.1%, 100.0%]	< 0.001	< 0.001
6	S25	Kabe	Yes	17/17 (100.0%) [81.6%, 100.0%]	< 0.001	< 0.001
9	F25	Kabe	Yes	13/13 (100.0%) [77.2%, 100.0%]	< 0.001	< 0.001
4	F24	Riichi judgement	Yes	11/13 (84.6%) [57.8%, 95.7%]	< 0.001	0.001
7	S25	Riichi judgement	Yes	11/13 (84.6%) [57.8%, 95.7%]	< 0.001	0.001
10	F25	Riichi judgement	Yes	7/8 (87.5%) [52.9%, 97.8%]	< 0.001	0.009

among participants who later exhibited systematic misapplication in the full-exposure round, *none* encountered a detector item during the first round of practice with adaptive learning enabled because BKT reached the mastery threshold and stopped assigning further problems. Consequently, without the second (full-exposure) round, these learners would not have been exposed to any ‘do-not-act’ cases capable of revealing deceptive overgeneralization.

This consistent pattern across cohorts and skills indicates that correctness-based mastery inference can terminate practice prematurely before learners are exposed to conditions that require discriminating when the learned action should *not* be applied. Consequently, adaptive learning risks leaving deceptive overgeneralization unaddressed by halting practice prior to refinement of critical application constraints.

3.4 | Effectiveness of Our Proposed Procedure in Addressing Deceptive Overgeneralization (RQ4)

Across Experiments 6, 7, 9, 10, and 11, where tailored remediation was deployed, error rates decreased across successive remediation opportunities in all five cohorts, despite minor fluctuations (Figure 7). Such a smoothly decreasing error pattern is indicative of effective student learning (Stamper and Koedinger 2011). In the present context, this pattern reflects

learners refining when the previously learned action should *not* be applied.

The qualitative pattern is further reflected in the magnitude of change between the error rate on the initial detector item and that on the final remediation item. As summarized in Table 4, initial misapplication rates on the first detector item ranged from 78.6% (11/14) to 100.0% (17/17; 13/13), indicating systematic overgeneralization prior to remediation. Following several instances of targeted practice designed via our proposed procedure, post-remediation error rates on the final detector item decreased substantially, ranging from 0.0% (0/14) to 23.1% (3/13). Corresponding effect sizes were uniformly large, with Cohen's *h* values between 1.696 and 2.441.

For contextual calibration, we compared post-remediation error rates against the 12% error rate implied by the BKT accurate-mastery-with-slip expectation. In three of the five cohorts, post-remediation rates fell at or below this benchmark. In the remaining two cohorts, observed rates were modestly higher (12.5% and 23.1%) but were not statistically distinguishable from the 12% benchmark under one-sided exact binomial tests (all $p > 0.05$; for the highest rate, 23.1%, $p = 0.198$).

Together, these findings indicate that tailored remediation substantially reduced systematic misapplication and restored performance to a level consistent with mastery expectations across skills and cohorts.

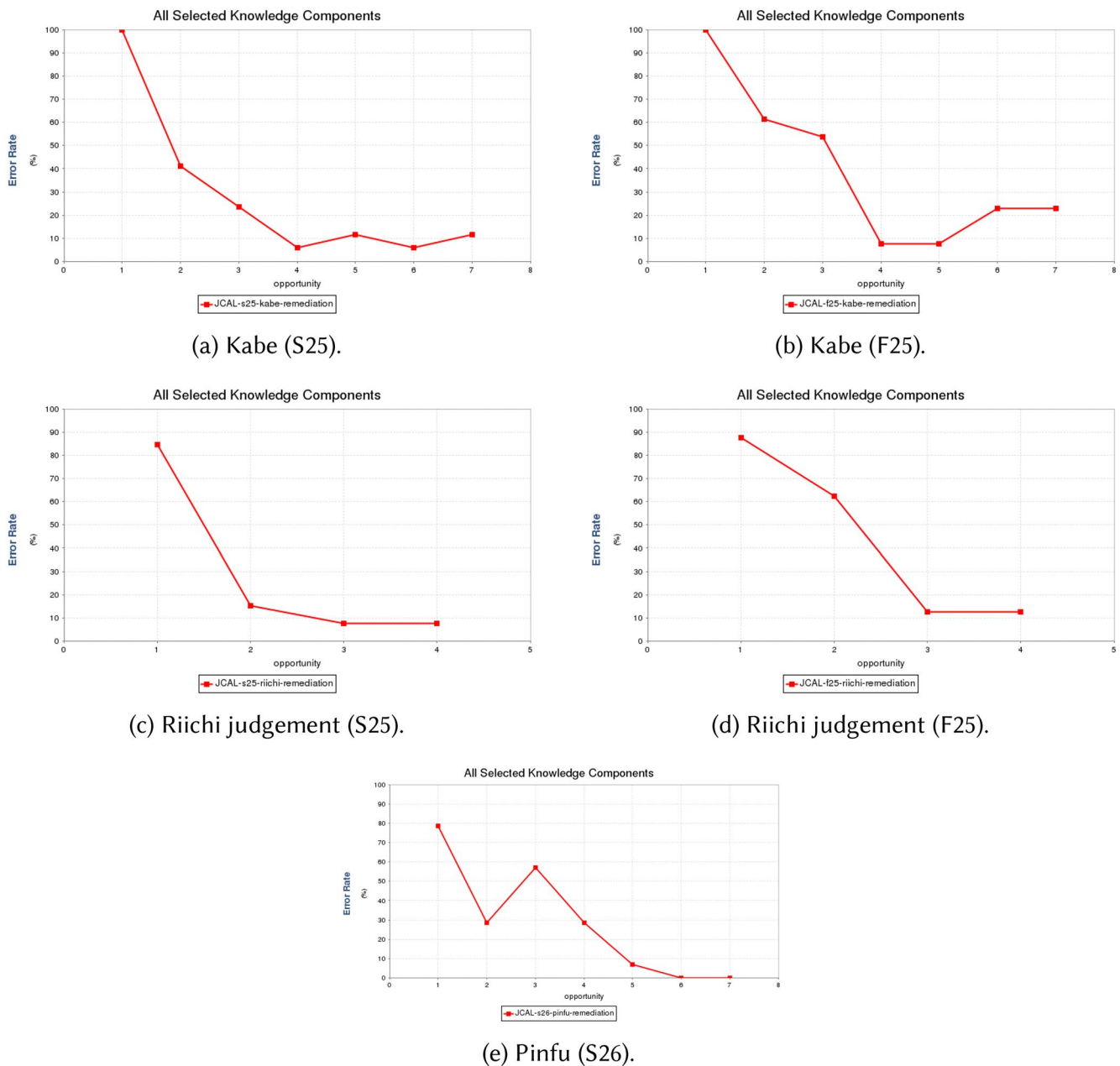


FIGURE 7 | Learning curves for tailored remediation across skills and cohorts (RQ4). Within each skill, error rates decrease over successive remediation opportunities.

4 | Discussion

Our findings provide consistent evidence that deceptive overgeneralization differs from encountering new skills. In Experiments 2–11, learners systematically misapplied acquired actions under inappropriate conditions, with error rates exceeding the new-KC baseline, indicating that learners were not merely guessing but systematically applying an overgeneralized rule. Supported by the results of RQ4, our proposed procedure enables instructional designers to systematically detect and remediate deceptive overgeneralization, ensuring that learners acquire not only the ability to execute an action but also a precise understanding of its application constraints.

4.1 | Applying the Framework Beyond the Study Domain: A Case Study of Whole-Number Bias in K–12 Decimal Learning

To illustrate the cross-domain applicability of the deceptive overgeneralization framework, we examine one representative misconception in K–12 mathematics: the whole-number bias (WNB) (Ni and Zhou 2005; Van Hoof et al. 2015). In decimal comparison, this bias can manifest as a ‘longer-is-larger’ judgement when comparing decimals (e.g., 0.34 vs. 0.213), in which a student treats the decimal with more digits as larger because the digits after the decimal point are interpreted like a whole-number string (e.g., 34 vs. 213). Prior syntheses and empirical

TABLE 4 | Effectiveness of tailored remediation for deceptive overgeneralization (RQ4).

Exp.	Semester	Skill	Initial misapplication [95% CI]	Post-remediation error [95% CI]	Cohen's <i>h</i>
6	S25	Kabe	17/17 (100.0%) [81.6, 100.0]	2/17 (11.8%) [3.3, 34.3]	2.441
9	F25	Kabe	13/13 (100.0%) [77.2, 100.0]	3/13 (23.1%) [8.2, 50.3]	2.139
7	S25	Riichi judgement	11/13 (84.6%) [57.8, 95.7]	1/13 (7.7%) [1.4, 33.3]	1.773
10	F25	Riichi judgement	7/8 (87.5%) [52.9, 97.8]	1/8 (12.5%) [2.2, 47.1]	1.696
11	S26	Pinfu	11/14 (78.6%) [52.4, 92.4]	0/14 (0.0%) [0.0, 21.5]	2.179

Note: Misapplication refers to applying the previously learned action on the first detector problem where the correct response is to *not* take that action. Bracketed values are Wilson 95% confidence intervals. The post-remediation column reports the error rate on the *final* problem in the tailored remediation practice where the correct response is to refrain from the previously misapplied action.

work characterize this bias as a well-documented and persistent obstacle in decimal learning (Ni and Zhou 2005; Van Hoof et al. 2015).

4.1.1 | Theoretical Modelling of Whole-Number Bias

From the perspective of deceptive overgeneralization, this misconception can be formalized as omitting the critical application constraint of the following composite-condition, single-action production rule:

IF [Current Goal] The goal is to compare two numbers

AND [Application Constraint] The numbers are whole numbers

AND [Mathematical] The numbers have different lengths

AND [Mathematical] The numbers are positive

THEN [Single Action] Treat the longer number as larger

Through the lens of deceptive overgeneralization, the misconception can be understood as a systematic consequence of instructional sequencing. Learners typically begin by learning whole numbers before they learn decimals. Note that the overgeneralized rule can consistently produce seemingly satisfactory performance when comparing positive whole numbers. Within the initial learning context of positive whole-number comparison, there are no counterexamples to the rule that numbers with more digits are larger; and the missing boundary condition has no observable consequence until decimals are introduced. According to ACT-R, through repeated application that produced seemingly satisfactory outcomes, this overgeneralized production rule can become highly strengthened (Anderson 1983, 250).

Importantly, this deceptive-overgeneralization account does not claim that students struggle because instruction failed to

introduce decimal-comparison procedures. Rather, according to ACT-R, newly introduced productions do not automatically replace older ones; they coexist, and *conflict resolution* determines which production is selected (Anderson 1983, 237, 132). Because production strength reflects prior successful use (Anderson 1983, 250), a newly introduced decimal-comparison production may initially be weaker than an older whole-number production with a longer history of success. Without tailored instruction that directly contrasts the old and new rules and helps learners encode the missing application constraint, the old overgeneralized whole-number production may remain eligible during conflict resolution and sometimes be selected instead of the newly taught decimal-comparison production.

4.1.2 | Exploratory Secondary Analysis of a Large-Scale Decimal Learning Dataset

We conducted an exploratory secondary analysis using data from *Decimal Point*, a K–12 mathematics learning game with an underlying ITS (McLaren 2024). We use this analysis not as a confirmatory replication of the Riichi Mahjong experiments, but as a worked example demonstrating how the deceptive overgeneralization framework can provide a mechanistic account of, and offer a targeted remediation strategy for, a well-known mathematics misconception.

The *Decimal Point* data¹ are hosted in DataShop, which allows access upon request for secondary analysis. After requesting and obtaining access, we analysed assessment data from 13 *Decimal Point* datasets spanning Fall 2015 through Fall 2025. The assessments used three counterbalanced versions (A, B, and C), administered across the six possible pretest–posttest–delayed-test sequences to control for practice effects and test difficulty.

Prior work on *Decimal Point* has provided substantial evidence that the game can support learning gains in decimal reasoning (McLaren 2024). Our question here was therefore not whether *Decimal Point* is effective overall. Instead, we asked whether whole-number bias remains a challenging residual

misconception even in a learning environment with demonstrated overall efficacy, and whether the deceptive overgeneralization framework helps explain why.

Among these assessment items, we identified the item family shown in Figure 8 as a close analogue to the framework's 'do-not-act' scenario. In version A, the correct response is that no number can be larger than 4.8 and smaller than 4.12. However, if students overgeneralize the 'longer-is-larger' rule, they may incorrectly interpret 4.12 as larger than 4.8, leading them to enter one of three values: 4.9, 4.10, or 4.11. The same 'do-not-act' structure appears in the corresponding items in assessment versions B and C, with only the numerical values differing across forms. For the phase comparison reported below, we restricted the analysis to students with a final answer to this item family at all three assessment phases (3299 students; 9897 final answers).

As Table 5 shows, the 'do-not-act' item continued to elicit the WNB after the *Decimal Point* intervention. Among students included in the phase comparison, the overall error rate decreased from 53.11% at pretest to 48.50% at immediate posttest and 47.47% at the one-week delayed posttest. However, these reductions were modest in magnitude (4.61 and 5.64 percentage points; Cohen's $h = 0.092$ and $h = 0.113$, respectively). The WNB-aligned error rate changed even less, decreasing from 44.71% to 41.77% at both later phases, a reduction of 2.94 percentage points ($h = 0.059$). All these standardized differences were below Cohen's conventional small-effect benchmark of ($h = 0.20$). Notably, 84.19%–87.99% of incorrect responses were WNB-aligned across phases, which indicate that students' errors overwhelmingly took the form predicted by the 'longer-is-larger' overgeneralization, rather than being distributed across arbitrary decimal responses. The high proportion of errors

aligned with WNB, rather than being scattered, suggests that these students are not guessing or lacking fundamental decimal knowledge, but are systematically misapplying an overgeneralized production.

This demonstrates how a formerly acquired production reinforced in earlier learning contexts can persist as a 'deceptive overgeneralization' when its application conditions are not sufficiently constrained. The implication for instructional design is clear: simply teaching the correct decimal procedure may be insufficient; the older, competing production must be explicitly refined with a missing application constraint.

4.1.3 | Instructional Implications of the Whole-Number Bias Analysis

This secondary analysis should be interpreted cautiously, as it does not test the same mechanism with task-loop adaptivity that led to premature automated practice termination. Its value is analytic and supplementary: it shows how the deceptive overgeneralization framework can provide a concrete mechanism-level account of a representative misconception in traditional academic domains.

This decimal-learning case study also clarifies that task-loop adaptivity is neither a precondition for deceptive overgeneralization nor the source of the underlying cognitive mechanism; our findings should not be interpreted as suggesting that adaptive learning is inherently flawed. Rather, in our main experiments, correctness-based task-loop adaptivity served as a controlled diagnostic context. Because observed performance was used to infer mastery and trigger explicit stopping rules, the experiments provided a clear, quantifiable boundary between continued practice and its termination. This precision made an otherwise diffuse phenomenon empirically observable and analytically tractable: learners could appear to have mastered a production before encountering the boundary-violating cases needed to refine its application conditions.

Critically, the same underlying cognitive mechanism can arise in traditional instructional sequences, independent of adaptive learning. In many domains, instruction first introduces a production that works well within an early, restricted—and often simplified—problem space, where boundary-violating cases are not yet relevant. When the curriculum later expands that problem space, the earlier production may need to be refined with its missing application constraints rather than merely supplemented

Decimal Test

Write down any number that is larger than 4.8, but smaller than 4.12 or, if it's not possible, check "That's impossible."

☐ That's impossible

FIGURE 8 | Example decimal-comparison assessment item from *Decimal Point*. The item asks whether any number can be larger than 4.8 and smaller than 4.12; the correct response is 'That's impossible.' Students with whole number bias may instead enter 4.9, 4.10, or 4.11.

TABLE 5 | Error rates and whole-number-bias-aligned errors on the 'do-not-act' decimal-comparison assessment item among students with final answers at all three assessment phases.

Assessment phase	Final answers	Incorrect answers	WNB-aligned errors	WNB among incorrect
Pretest	3299	1752 (53.11%)	1475 (44.71%)	84.19%
Immediate posttest	3299	1600 (48.50%)	1378 (41.77%)	86.12%
One-week delayed posttest	3299	1566 (47.47%)	1378 (41.77%)	87.99%

Note: WNB-aligned errors are responses consistent with misapplying the 'longer-is-larger' rule. Abbreviation: WNB = whole-number bias.

by a newly introduced production that must compete with the older, more practiced rule during conflict resolution.

Importantly, the framework also suggests a specific, testable, and potentially domain-general instructional implication. Instructional designers should audit instructional sequences for earlier productions that may have been learned with underspecified application constraints and later come into competition with newly introduced productions as the problem scope expands. Rather than only teaching the correct new production, instruction could include contrastive cases in which the earlier production is tempting but inappropriate, paired with feedback that makes the missing application constraint explicit. Such ‘do-not-act’ practice may help students recognize when the earlier production should be withheld, thereby reducing competition between the older and newer productions in contexts where the earlier rule no longer applies.

4.2 | Limitations and Future Work

Several limitations of our work should be noted. First, our empirical studies were conducted in Riichi Mahjong. This domain is suitable for studying composite-condition skills with single observable actions, but the data do not establish the prevalence of deceptive overgeneralization in other domains. We complement these studies with a secondary analysis using a large-scale decimal learning dataset, which illustrates how the deceptive overgeneralization framework can provide a mechanistic account of whole-number bias in decimal comparison and shows that whole-number-bias-aligned errors were prevalent and persistent in that dataset. However, this analysis should be interpreted as a theory-grounded exploratory application, not as a direct empirical test of deceptive overgeneralization in K–12 mathematics. Future work should test whether incorporating ‘do-not-act’ practice that refines the application constraints of earlier productions—rather than only teaching the new production—improves learners’ ability to apply newly introduced productions appropriately.

Second, our participant pool may reflect self-selection. The course was voluntary and interest-driven, and study participation was optional; therefore, learners who enrolled and/or consented may differ from the broader population in prior knowledge, motivation, or available time. This limits the generalizability of prevalence estimates (e.g., misapplication rates) to settings with different recruitment and participation constraints.

Third, although the phenomenon replicated across cohorts and skills, individual experiments had relatively small samples, limiting precision for estimating effect magnitudes and for testing moderators. These constraints also shaped two design choices and their limitations discussed next.

Fourth, our design structured practice sequences to position ‘do-not-act’ cases after ample non-diagnostic cases. This design choice helped make premature termination before such cases observable with relatively limited samples, but it may differ from learners’ exposure under more organic instructional sequencing. For example, in an adaptive practice sequence that shuffles rather than fixes item order, some learners may

encounter detector-like cases early enough to refine an overgeneralized rule, whereas others may, by chance, receive a sequence effectively equivalent to the present design and have the adaptive system terminate practice upon deemed mastery after seeing only non-diagnostic cases in which the overgeneralized rule consistently appears successful. Larger-sample deployments in authentic curricula are therefore needed to estimate prevalence under natural instructional sequencing.

Fifth, our within-subjects design (adaptive round followed by full-exposure round) was also motivated by these sample constraints, because it allowed each learner to contribute evidence under both exposure conditions. However, this fixed order may introduce order and carryover effects (e.g., familiarity, fatigue, or learning from feedback in Round 1) that could influence Round 2 performance. Future larger-sample studies could supplement the present evidence by counterbalancing order or assigning exposure conditions between subjects to test whether the pattern persists when these order effects are removed.

Sixth, our quantitative baselines rely on a two-state BKT reference model and a standard CTAT-style parameterization; the exact expected error rates depend on assumptions about slip/guess and mastery thresholds. Future work should test robustness under plausible parameter ranges and compare against richer tracing models that explicitly represent partial or condition-incomplete knowledge states.

Seventh, our remediation results demonstrate an immediate reduction in misapplication within the ITSs, but we did not evaluate longer-term retention or transfer beyond the remediation sequence. Future work should test whether these reductions persist after a delay and transfer to novel detector items or authentic tasks without immediate remediation support.

5 | Conclusion

This work examined a theoretically domain-general risk in knowledge compilation that we coin ‘deceptive overgeneralization’: learners may omit critical application constraints yet repeatedly perform a seemingly correct action during practice, thereby reinforcing an overgeneralized macro-production that they later misapply when a critical application constraint is violated. Correctness-based task-loop adaptivity is not the source of the mechanism, but it makes the risk especially salient and analytically tractable because adaptive learning systems using such stopping rules can terminate practice before learners encounter cases where the learned action should not be applied.

Across multiple skills and cohorts, we found that learners frequently misapplied acquired actions on the first ‘do-not-act’ detector item, with misapplication rates far exceeding both the BKT accurate-mastery-with-slip expectation and, in nearly all cases, the error rate expected if the detector item simply required a yet-to-be-learned KC. We also found that BKT-based mastery stopping rules can terminate practice before learners ever encounter such detector items, leaving deceptive overgeneralization undetected. Finally, when we introduced short sequences of tailored ‘do-not-act’ practice problems paired with feedback that addresses the missing constraints, learners made

progressively fewer errors across remediation opportunities, supporting the interpretation that targeted practice helped them refine the missing application constraints.

Together, these findings suggest that observed correctness alone is not sufficient evidence of conditional understanding for composite-condition, single-action skills. For CAL, and especially for adaptive and AI-supported learning systems, the implication is therefore not only to evaluate whether a learner can execute an action but also to assess when that action should be withheld. A practical design implication is to incorporate targeted detector items in which the correct response is to *refrain* from applying the skill being practiced and to place these items before mastery stopping rules trigger. Meanwhile, the framework is not limited to CAL: the whole-number bias case study suggests that traditional instructional sequences can create similar risks when an earlier production learned in a restricted problem space may be overgeneralized and later compete with newer productions after the problem space expands. More broadly, deceptive overgeneralization offers a theoretically grounded, empirically supported, and instructionally actionable framework for treating mastery as discrimination of application constraints as well as action execution.

Future work should test the domain-general value of this framework by examining whether similar patterns arise in other domains that rely on compiled macro-productions, including both CAL and non-CAL settings, and should evaluate how to integrate ‘do-not-act’ detectors into knowledge tracing models and adaptive policies without substantially increasing learner time-on-task. Larger-sample studies could also estimate how frequently deceptive overgeneralization occurs under natural instructional sequences and identify which instructional designs most reliably support refinement of missing application constraints.

Author Contributions

Marshall An: conceptualization, methodology, software, data curation, investigation, validation, formal analysis, visualization, writing – original draft, project administration, resources, writing – review and editing. **Bruce M. McLaren:** conceptualization, writing – review and editing, supervision, methodology, validation. **John Stamper:** conceptualization, writing – review and editing, supervision, methodology, validation.

Acknowledgements

This article is an expanded and substantially revised version of a paper presented at the Twentieth European Conference on Technology-Enhanced Learning (EC-TEL 2025): An, M., & Stamper, J. (2025, September). *Detect and Remedy Deceptive Overgeneralization in Adaptive Learning*. In *European Conference on Technology Enhanced Learning* (pp. 3–17). Cham: Springer Nature Switzerland. The authors used ChatGPT and Codex (GPT-5.2 and GPT-5.5, OpenAI) for language editing (spelling, grammar and clarity) only. All scientific content, analyses, interpretations, and conclusions are solely those of the authors, who take full responsibility for the manuscript.

Funding

The authors have nothing to report.

Ethics Statement

This study was approved by the Carnegie Mellon University Institutional Review Board (IRB Protocol #STUDY2023_00000510). All participants provided informed consent prior to participation.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Endnotes

¹ <https://pslcdatashop.web.cmu.edu/Project?id=567>.

References

- Aircraft Accident Investigation Bureau. 2002. “Final Report of the Aircraft Accident Investigation Bureau on the Accident to the Saab 340B Aircraft, Registration HB-AKK of Crossair Flight CRX 498 on 10 January 2000 Near Nassenwil/ZH.”
- Aleven, V., B. M. McLaren, J. Sewall, and K. R. Koedinger. 2006. “The Cognitive Tutor Authoring Tools (CTAT): Preliminary Evaluation of Efficiency Gains.” In *Intelligent Tutoring Systems: 8th International Conference, ITS 2006*, 61–70. Springer.
- Aleven, V., B. M. McLaren, J. Sewall, et al. 2016. “Example-Tracing Tutors: Intelligent Tutor Development for Non-Programmers.” *International Journal of Artificial Intelligence in Education* 26: 224–269.
- Aleven, V., E. A. McLaughlin, R. A. Glenn, and K. R. Koedinger. 2016. “Instruction Based on Adaptive Learning Technologies.” *Handbook of Research on Learning and Instruction* 2: 522–560.
- Ambrose, S. A., M. W. Bridges, M. DiPietro, M. C. Lovett, and M. K. Norman. 2010. *How Learning Works: Seven Research-Based Principles for Smart Teaching*. John Wiley & Sons.
- Anderson, J. R. 1982. “Acquisition of Cognitive Skill.” *Psychological Review* 89, no. 4: 369.
- Anderson, J. R. 1983. *The Architecture of Cognition*. Harvard University Press.
- Anderson, J. R. 1992. “Automaticity and the ACT* Theory.” *American Journal of Psychology* 105, no. 2: 165–180.
- Anderson, J. R. 1996. “ACT: A Simple Theory of Complex Cognition.” *American Psychologist* 51, no. 4: 355.
- Anderson, J. R., and C. D. Schunn. 2013. “Implications of the ACT-R Learning Theory: No Magic Bullets.” In *Advances in Instructional Psychology, Volume 5*, 1–33. Routledge.
- Baron, J., and J. C. Hershey. 1988. “Outcome Bias in Decision Evaluation.” *Journal of Personality and Social Psychology* 54, no. 4: 569.
- Bjork, R. A., J. Dunlosky, and N. Kornell. 2013. “Self-Regulated Learning: Beliefs, Techniques, and Illusions.” *Annual Review of Psychology* 64, no. 1: 417–444.
- Bloom, B. S. 1968. “Learning for Mastery.” *Evaluation Comment* 1, no. 2: 1–12.
- Chang, N. M. 2006. *Learning to Discriminate and Generalize Through Problem Comparisons*. Ph. D. Dissertation. Carnegie Mellon University.
- Chi, M. T. H., P. J. Feltovich, and R. Glaser. 1981. “Categorization and Representation of Physics Problems by Experts and Novices.” *Cognitive Science* 5, no. 2: 121–152.

- Chiba, D. 2026. *Riichi Book I: A Mahjong Strategy Primer for European Players* Version 14. <https://dainachiba.github.io/RiichiBooks/>.
- Collins, A., J. S. Brown, and A. Holum. 1991. "Cognitive Apprenticeship: Making Thinking Visible." *American Educator* 15, no. 3: 6–11.
- Corbett, A. T., and J. R. Anderson. 1994. "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge." *User Modeling and User-Adapted Interaction* 4: 253–278.
- Kalyuga, S. 2009. "The Expertise Reversal Effect." In *Managing Cognitive Load in Adaptive Multimedia Learning*, 58–80. IGI Global.
- Koedinger, K. R., and V. Aleven. 2007. "Exploring the Assistance Dilemma in Experiments With Cognitive Tutors." *Educational Psychology Review* 19: 239–264.
- Koedinger, K. R., J. R. Anderson, W. H. Hadley, and M. A. Mark. 1997. "Intelligent Tutoring Goes to School in the Big City." *International Journal of Artificial Intelligence in Education* 8: 30–43.
- Koedinger, K. R., R. S. J. D. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. 2010. "A Data Repository for the EDM Community: The PSLC DataShop." In *Handbook of Educational Data Mining*, 43–56. CRC Press.
- Koedinger, K. R., P. F. Carvalho, R. Liu, and E. A. McLaughlin. 2023. "An Astonishing Regularity in Student Learning Rate." *Proceedings of the National Academy of Sciences of the United States of America* 120, no. 13: e2221311120.
- Koedinger, K. R., and A. T. Corbett. 2006. "Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom." In *The Cambridge Handbook of the Learning Sciences*, edited by R. K. Sawyer, 61–78. Cambridge University Press. <https://doi.org/10.1017/CBO9780511816833.006>.
- Koedinger, K. R., A. T. Corbett, and C. Perfetti. 2012. "The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning." *Cognitive Science* 36, no. 5: 757–798.
- LearnLab. 2011. "Feature Validity." Accessed May 29, 2024. https://learnlab.org/wiki/index.php?title=Feature_validity.
- Li, J., S. Koyamada, Q. Ye, et al. 2020. "Suphx: Mastering Mahjong With Deep Reinforcement Learning." arXiv Preprint arXiv:2003.13590.
- Luchins, A. S. 1942. "Mechanization in Problem Solving: The Effect of Einstellung." *Psychological Monographs* 54, no. 6: i-95.
- McLaren, B. M. 2024. "Decimal Point: A Decade of Learning Science Findings With a Digital Learning Game." In *Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*, edited by P. Ilic, I. Casebourne, and R. Wegerif, 145–203. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71232-6_9.
- Mizukami, N., and Y. Tsuruoka. 2015. "Building a Computer Mahjong Player Based on Monte Carlo Simulation and Opponent Models." In *2015 IEEE Conference on Computational Intelligence and Games (CIG)*, 275–283. IEEE.
- National Research Council. 2000. *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. National Academies Press. <https://doi.org/10.17226/9853>.
- Newell, A. 1994. *Unified Theories of Cognition*. Harvard University Press.
- Ni, Y., and Y.-D. Zhou. 2005. "Teaching and Learning Fraction and Rational Numbers: The Origins and Implications of Whole Number Bias." *Educational Psychologist* 40, no. 1: 27–52.
- Oxford Reference. n.d. "Computer Assisted Learning." Accessed July 11, 2026. <https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095629962>.
- Perkins, D. N., and G. Salomon. 1992. "Transfer of Learning." In *International Encyclopedia of Education*, 2nd ed. Pergamon Press.
- Ross, B. H. 1984. "Reminders and Their Effects in Learning a Cognitive Skill." *Cognitive Psychology* 16, no. 3: 371–416.
- Schraagen, J. M., S. F. Chipman, and V. L. Shalin. 2000. *Cognitive Task Analysis*. Psychology Press.
- Shute, V. J. 2008. "Focus on Formative Feedback." *Review of Educational Research* 78, no. 1: 153–189.
- Stamper, J. C., and K. R. Koedinger. 2011. "Human-Machine Student Model Discovery and Improvement Using DataShop." In *International Conference on Artificial Intelligence in Education*, 353–360. Springer.
- Van Hoof, J., L. Verschaffel, and W. Van Dooren. 2015. "Inappropriately Applying Natural Number Properties in Rational Number Tasks: Characterizing the Development of the Natural Number Bias Through Primary and Secondary Education." *Educational Studies in Mathematics* 90, no. 1: 39–56.
- VanLehn, K. 2006. "The Behavior of Tutoring Systems." *International Journal of Artificial Intelligence in Education* 16, no. 3: 227–265.
- VanLehn, K. 2016. "Regulative Loops, Step Loops and Task Loops." *International Journal of Artificial Intelligence in Education* 26, no. 1: 107–112.
- Vygotsky, L. S. 1978. *Mind in Society: The Development of Higher Psychological Processes*, edited by M. Cole, V. John-Steiner, S. Scribner, and E. Souberman. Harvard University Press.
- Woolf, B. P. 2010. *Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing e-Learning*. Morgan Kaufmann.