Altered Records and Available Actions: Counterfactual Search in Knowledge Tracing and the Limits of Educational Recourse

Flyxion
Independent Researcher

November 2026

Abstract

Knowledge tracing models predict whether a learner will answer a future question correctly from a record of past answers. A recent proposal, RA4KT, searches for a sparse set of changes to that record, restricted to turning incorrect answers into correct ones, under which a trained predictor would forecast success on a target concept, and then orders the associated concepts into a study path using a curriculum graph. This essay examines what such a procedure establishes. Its defensible result is a statement about the sensitivity of a fixed predictor to edited inputs, together with a graph-guided ordering of the concepts involved. A recommendation to a learner is a different kind of object, because it concerns an action taken now whose effect on later performance requires evidence that ordinary predictive training does not, by itself, provide. The essay states this distinction formally, shows how the evaluation design keeps the two objects from being separated empirically, and identifies several internal inconsistencies in the reported statistics, the temporal constraint, and the claimed ordering guarantees. The paper's properties are sorted into those the method enforces, those it only encourages, those supported by agreement with a supplied curriculum structure, and those that remain unvalidated, and the essay closes by specifying the evidence needed to support the stronger educational claims.

1 Introduction

Knowledge tracing, in the tradition of Corbett and Anderson (1994) and its neural descendants (Piech et al., 2015), treats a learner's history of answers as evidence about latent competence and uses that evidence to predict performance on questions not yet attempted. The models are evaluated by predictive accuracy, and they are good at it. Predictive accuracy alone does not establish which teaching action should follow. A teacher or a learner who sees a low predicted probability of success on a concept wants to know what to do about it, and the literature on explanation in machine learning supplies an obvious computational starting point: counterfactual explanation, the search for changes to an input that would alter a prediction (Wachter et al., 2018; Mothilal et al., 2020). Algorithmic recourse adds the further requirement that the proposed changes correspond to feasible actions (Karimi et al., 2021).

RA4KT, introduced under the name reverse attribution for knowledge tracing (Wang et al., 2026), adapts this idea to answer sequences. Given a history in which the model predicts failure on a target concept, it searches for a small set of earlier incorrect answers that, if they had been correct, would lead the same model to predict success. It penalizes the number of changes, favors concepts that are close to the target in a curriculum graph, forbids turning correct answers into incorrect ones, and finally arranges the concepts attached to the changed answers into an ordered path ending at the target. The result is presented as an actionable teaching plan.

The computational idea is reasonable and the engineering is carefully described in many places. The difficulty lies in the interpretive step. Between the sentence "this edited history makes the predictor optimistic" and the sentence "this learner should study these concepts in this order" there is a chain of assumptions that the method does not discharge and that its evaluation cannot test. The purpose of this essay is to make that chain explicit, to separate what the paper itself concedes from what its framing implies, and to record several places where the reported details do not cohere. The treatment is deliberately narrow. It concerns the logic of the claims and the internal consistency of the reported methods and statistics, and it does not attempt to reproduce the experiments.

2 What the Procedure Computes

Let $H_T = ((q_1, r_1), \ldots, (q_T, r_T))$ denote a learner's record of questions and binary responses, and let $f(H_T, k) \in [0, 1]$ denote the predictor's probability of success on concept k. RA4KT searches for a modified record H'_T that differs from H_T only at positions where the original response was incorrect, where the modification sets the response to correct, and that is intended to satisfy

$$f(H'_T, k) \; > \; \tau_{\text{flip}}$$

where the paper's flip condition sets $\tau_{\text{flip}} = 0.5$ (Eq. 5, p. 9). Three thresholds must be kept apart: this flip threshold, the stricter evaluation threshold $\tau_{\text{eval}} = 0.8$ used to define validity (Eq. 14, p. 14), and the total-loss stopping threshold τ_{stop} in Algorithm 1 (p. 11), which is a condition on the optimization loss and not on the predictor's output. Success of the search is an empirical outcome and is not guaranteed by construction. The search is carried out over a continuous relaxation of the response vector, with a penalty on the squared distance from the original responses, a penalty that weights modified positions by their graph distance from the target concept, and a final discretization. The positions that changed identify concepts, and the concepts are ordered by a greedy traversal of an undirected graph anchored at the target and then reversed so that the target appears last.

Stated this way, the output is a triple: a sparse edit set, a set of concepts attached to that edit set, and a linear order on those concepts. Each component has a clear meaning inside the model. The edit set identifies one combination of historical response changes that raises the predictor's output under the imposed penalties and restrictions, and it reflects initialization and optimization success as well as sensitivity. The concept set inherits that sensitivity through the item-to-concept mapping. The order reflects the geometry of a graph that the authors

supply. The recorded answers concern an actual learner and the graph encodes educational content, but these constructions do not, by themselves, identify the effect of a future teaching action on the learner.

3 From Altered Records to Available Actions

A recommendation has a different logical form. It names an action a available to the learner or teacher at the present time T, and it asserts something about a later outcome $Y_{T+\Delta}$, defined as performance on a specified target assessment administered after a fixed interval under a common assessment protocol, so that an intervention cannot change the measured outcome merely by changing which items are attempted. The quantity such a claim concerns is closer to

$$\mathbb{E}[\, Y_{T+\Delta} \mid \mathrm{do}(A_T = a),\, H_T\,],$$

compared across candidate actions a and b, each of which must specify its activity content, duration, feedback, and progression rules, as in the contrast $\mathbb{E}[Y^a_{T+\Delta} - Y^b_{T+\Delta} \mid H_T]$, where the superscripts index the assigned action. An instruction such as "study fractions" is not a sufficiently specified action in this sense.[1] Finding H'_T with $f(H'_T, k) > \tau_{\text{flip}}$ does not identify this expectation, for at least three separable reasons.

The first is that the altered record and the intervention occur at different times. A record altered at position $t < T$ changes the evidence the predictor sees about the past. An action taken at time T changes the learner's future, and the evidence that accumulates afterwards will include the learner's behavior on the action itself, on whatever intervening material is encountered, and on the target. The predictor was trained to forecast the next response given a fixed past. Its response-prediction objective does not, by itself, identify the effect of inserting a new teaching event, and no property of that objective ties its response to edited history with the learner's response to instruction.

The second is that an earlier success is evidence of a state that the action may not reproduce. A history containing a correct answer on a fractions item will lead a reasonable predictor to infer greater fractions competence. The inference holds under the assumption that the answer was produced by competence. Reviewing fractions now may generate the same competence, a partial version of it, or none, depending on the quality of the review, the interval since last exposure, and the learner's other circumstances. The edit tells us which inference the predictor would draw from a successful outcome. It does not tell us the probability that the action yields that outcome.

The third is that the directional constraint addresses only a narrow part of feasibility. The authors restrict edits to incorrect answers so that the hypothetical record depicts improvement and never regression, and they describe this as the Reality property. The restriction is a sensible way to avoid an obviously perverse explanation. It does not establish that the improvement

[1] The preceding sentence specifies the contrast, so identification and estimation remain the open problems. Identification asks whether that effect is determined by the available data distribution together with causal assumptions. Estimation then concerns its numerical recovery from finite observations. Section 8 returns to the evidence these steps require.

is within the learner's reach, that suitable materials exist, that the time cost is acceptable, or that the proposed concepts are the right ones to study. Reality in the paper's sense is a monotonicity condition on the edit, and actionability is a claim about the world.

Taken together, these points mean that the step from H'_T to a study plan requires evidence connecting the proposed teaching actions to subsequent learner outcomes, whether through a validated causal transition model or a credible evaluation of the recommendation policy. In the language of causal inference (Pearl, 2009), the edit changes the values supplied to a fixed prediction function, whereas the plan intervenes on the learner's instructional experience, and a causal interpretation of the former requires an additional model of how the historical responses were generated. The literature on recourse has long recognized the same difficulty in other settings: Karimi, Schölkopf and Valera argue that counterfactual explanations become recommendations only when they are interpreted as interventions within a causal model of the system, and that otherwise they may recommend actions that do not produce the stated change (Karimi et al., 2021). Barocas, Selbst and Raghavan make a related point about the hidden assumptions behind counterfactual explanations (Barocas et al., 2020). Knowledge tracing adds the further complication that the object being intervened on is a person whose competence is a latent variable.

4 What the Paper Concedes

It is important to credit the authors where they draw the relevant line themselves. The paper states that its use of the word causal refers to the relationship between modified inputs and the model's own predictions, and not to a causal relationship with real learning outcomes. Its intervention evaluation is described as a simulation in which a fixed pretrained model stands in for the student, with no observed learning outcomes, and the results are labeled as evidence of the model's counterfactual sensitivity to intervention ordering rather than demonstrated learning gains. The discussion also acknowledges that a counterfactual which flips a prediction does not guarantee that a real student would master the target concept.

These qualifications are accurate, and a fair summary of the paper must preserve them. The difficulty is one of consistency. The title, the abstract, the stated contributions, and the property definitions use the vocabulary of actionable recourse, teaching guidance, and reduced learning burden, and the sparsity metric is glossed as a measure of lighter learning burden for students. If the evidence only supports statements about a predictor's response to edited inputs, then the headline vocabulary is stronger than the evidence, and the concessions sit at some distance from the claims they qualify. A paper can be honest in its caveats and still leave a reader with a stronger impression than its results warrant.

5 Evaluation as Optimization Against the Predictor

The evaluation design reinforces the interpretive gap in two ways. Validity is measured by whether the edited sequence raises the predictor's output above a threshold. Since the search is an optimization of exactly that output, a high validity rate shows that the optimizer succeeds

against the predictor. It cannot show that the success corresponds to anything outside the predictor. The same holds for the offline educational simulation, in which mastery is simulated, the answer sequence is updated, and the predictor is queried again. The procedure for translating simulated mastery into updated answers determines the reported gains, and the predictor that scores the result is the same predictor that guided the search.

Pedagogical agreement is evaluated in a similarly coupled way. The method is penalized for selecting concepts far from the target in a curriculum graph, and it is then credited with a high pedagogical relevance score for selecting concepts close to the target in that graph. This demonstrates that the penalty has the intended effect on the output. It is not independent evidence that proximity in the graph corresponds to educational usefulness. The comparison against expert gold-standard paths has a related dependence, because the expert prerequisite hierarchy that supplies the gold paths is also the source of the graph used by the method, so that agreement largely measures conformity to supplied structure. A prerequisite satisfaction rate of 97.2 percent is useful as a statement about how well the algorithm follows its prior. It is a weak test of whether following that prior helps anyone learn.

The baseline comparison has a further limitation. A sympathetic reading holds that post hoc projection is a standard repair baseline, and that the experiment is meant to show that constraint-aware search dominates constraint repair. The experiment is consistent with that narrower comparison, namely an advantage for this implementation over these post hoc projected baselines on particular metrics, subject to the tuning and query budgets given to the baselines. In the constrained comparison, competing methods generate counterfactuals without the restriction, and the restriction is then applied afterwards by projection. Removing forbidden edits after a search has finished predictably damages validity, since the baseline was allowed to rely on those edits while searching. RA4KT, by contrast, searches inside the feasible set from the start. The comparison therefore demonstrates the cost of imposing a constraint post hoc and it does not answer the question the paper's framing invites, which is whether the proposed relaxation and initialization machinery outperforms competing methods when all are given the same feasible domain. A fair test would give the baselines the same persistent restrictions during optimization and the same tuning and query budgets, and would include simple constrained searches such as greedy single-flip selection or bounded beam search as reference points. Without such comparisons it is not possible to say how much of the reported advantage derives from the definition of the search space and how much from the optimizer.

6 Inconsistencies in the Reported Details

Several details in the paper do not fit together as stated. They are separable from the conceptual argument above, and some may have ordinary implementation explanations, but the explanations are absent from the text.

6.1 Reported statistics

Validity is defined (Eq. 14, p. 14) as the mean over N cases of the indicator that the predicted success probability reaches $\tau_{\text{eval}} = 0.8$, and the captions of Tables 2 to 5 (pp. 17 to 18) state

that all values are computed across 200 test students. For a single run per student, the estimate must then be a multiple of $1/200 = 0.005$, and its student-level standard deviation must be approximately $\sqrt{\hat{v}(1-\hat{v})}$. For $\hat{v} = 0.941$ this is about 0.236. Table 2 (p. 17) reports 0.941 ± 0.017 for RA4KT-gs on XES3G5M.

A partial explanation appears elsewhere in the paper: the caption of Fig. 5 (p. 21) states that each sample's value is averaged over five repeated optimization runs with independent seeds. If the same averaging underlies the main tables, the validity values would be multiples of 0.001, which is consistent with entries such as 0.941 and 0.922, whereas a single run per student would force multiples of 0.005, which 0.941 is not. The table captions do not say this, so it remains an inference. It does not, however, rescue the reported standard deviation. If each of the five runs yields a binary validity indicator, as the definition of validity and values such as 0.941 are consistent with, then each student's validity takes values in $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The bound that follows requires no assumption about independence or exchangeability of the runs. It assumes only that each student's value is the mean of five binary outcomes, that 0.941 is the mean of those student-level values, and that the reported deviation is meant to describe their spread across students. If X takes values in $\{0, 0.2, \ldots, 1\}$ with mean $\mu \in [0.8, 1]$, then $(X - 0.8)(X - 1) \geq 0$ on that support, and taking expectations gives

$$\mathrm{Var}(X) \;\geq\; (\mu - 0.8)(1 - \mu).$$

At $\mu = 0.941$ this yields

$$\mathrm{SD}(X) \;\geq\; \sqrt{0.141 \times 0.059} \;\approx\; 0.091,$$

which is more than five times the reported figure. The reported 0.017 therefore cannot be the student-level standard deviation. It requires a different aggregation scheme, such as a spread of aggregate rates across seeds or folds, or a correction.

The confidence intervals show a related pattern. For sparsity, Table 5 (p. 18) reports 49.64 ± 5.58 with a 95% interval of $[44.26, 55.02]$. If 5.58 is the student-level standard deviation over 200 students, the conventional interval for the mean is

$$49.64 \pm 1.96\,\frac{5.58}{\sqrt{200}} \;\approx\; [48.87,\ 50.41].$$

The published half-width of 5.38 is nearly one full standard deviation. The validity interval $[0.925, 0.957]$ in the same table around 0.941 has the same property, with a half-width of 0.016 against a standard deviation of 0.017. The published intervals are not explained by the stated student-level standard deviations and sample size under the conventional calculation above, and their construction, aggregation level, and intended interpretation require clarification. The discrepancies undermine confidence in the reported uncertainty and in the significance claims, and the aggregation ambiguity leaves the weighting of cases and the unit underlying the uncertainty estimates insufficiently specified. The Wilcoxon objection is a reproducibility objection and does not show that the tests were computed incorrectly. It matters because the paper rests its claims of significance on paired Wilcoxon signed-rank tests applied to per-sample differences (p. 16). If the per-sample quantities are five-run averages, their paired differences can take values in increments of 0.2, rather than only -1, 0, and 1 as differences between

single binary outcomes do. The aggregation level therefore affects the distribution of paired differences, including ties and zero differences, and the paper does not specify which quantities entered the validity test. The significance claims cannot be assessed independently until the unit of observation is stated and the paired outcomes are released.

6.2 The temporal constraint

The paper defines a temporal sub-constraint (p. 7) under which only time steps strictly before the target step T may be modified, and reports a temporal consistency metric for it. The mask in the method description (Eq. 11, p. 11, repeated in line 1 of Algorithm 1) is

$$m_t = \mathbf{1}\Big(r_t^{\text{orig}} = 0\Big), \qquad t \in \{1, \ldots, T\},$$

which runs through T itself. Because the target answer is assumed to be incorrect, $m_T = 1$ and the target position is modifiable under the mask. The paper explains on page 7 that the temporal restriction is enforced at initialization by restricting candidate positions to $t < T$, and this is stated openly. An initialization restriction, however, does not constrain subsequent updates. A coordinate set to its original value at the start is free to move under later gradient steps unless the projection also fixes it. The guarantee would hold if the mask were defined as $m_t = \mathbf{1}(r_t^{\text{orig}} = 0)\,\mathbf{1}(t < T)$, or if the optimizer acted only on eligible coordinates. As written, the equations do not enforce what the text claims, and the temporal consistency metric is then an empirical measurement of whether the property happened to hold, not a verification of a guarantee. Modification of the target position would contaminate the explanation only if that response can influence the prediction being evaluated, which depends on the input-output alignment of the underlying implementation, and a correctly shifted next-response implementation might prevent it. The initialization argument applies when later updates can change every coordinate the displayed mask allows, and if the implementation parameterizes only eligible coordinates, initialization may be accompanied by a persistent restriction that the displayed equations omit. The unqualified defect is that the displayed mask does not enforce the stated restriction.

6.3 Ordering guarantees

The ordering procedure builds a greedy Hamiltonian path on an undirected graph from the target outward and reverses it. On page 12 the paper states plainly that this construction does not structurally guarantee the correct relative order of two non-target concepts that are prerequisite-related to each other but equally distant from the target, and it presents the procedure as a practical approximation supported by empirical agreement and not as a formal guarantee. On page 22 the description of the experimental condition says that the RA4KT path "strictly follows the prerequisite dependency relationship." The two statements cannot both describe a guaranteed property of the same procedure. A small example shows how the weaker reading arises. Let the true dependencies be $A \to B \to C$ with C the target, and let the undirected graph also connect A directly to C, as a topical association or a redundant prerequisite relation, so that A and B are both at distance one from C. A permitted tie-break

in a greedy traversal from C may visit A first, giving C, A, B, and reversal then yields B, A, C, which places B before its own prerequisite A. Ending at the target does not entail satisfying prerequisites, and the information needed to recover direction is exactly what an undirected distance graph discards. The stronger argument does not depend on the greedy search or on any tie-break. Two curricula with opposite prerequisite orientations on a single edge share the same undirected graph, so no traversal of that graph, Hamiltonian or otherwise, can in general recover a topological order of the underlying dependencies.

6.4 Metric definitions and experimental design

The pedagogical relevance metric is described in prose as the reciprocal of the average shortest-path distance between modified concepts and the target, whereas Eq. 16 (p. 15) averages the reciprocals of the distances. These are different functionals, as the distances 1 and 5 show: the mean of reciprocals is 0.6, while the reciprocal of the mean is 1/3. The equation also requires a convention for the case in which a modified concept coincides with the target, since the distance is then zero.

The sequence-length experiment (p. 19) truncates each student's history to its first L steps for $L \in \{50, 100, 150, 200\}$ while holding the target time step and target concept fixed. Elsewhere the target is the final step of the input. Truncating to the first L steps ordinarily moves the final step, so holding the target fixed requires a construction the paper does not describe, such as growing a window that ends just before a common target. The interpretation of any trend in validity with L depends on which construction was used.

Finally, the continuous relaxation introduces gaps between the formal problem and the algorithm, and the paper discloses most of them. It states (Eq. 7 and Eq. 10, pp. 10 to 11) that the Hamming-distance term is not optimized directly and is replaced during the gradient loop by a squared distance from the original responses, which agrees with the count of flipped answers at binary endpoints but behaves differently in between: one coordinate moved by 0.6 costs 0.36, while ten coordinates moved by 0.1 cost a total of 0.10. The surrogate can therefore prefer many small shifts to few large ones, so minimality is encouraged and not enforced, even though the constraint loss is described as the core guarantee of the minimalism property. The coherence surrogate is likewise acknowledged to be a non-deduplicated relaxation of the concept-level definition, so changing several occurrences of the same concept incurs repeated position-level penalties, whereas the deduplicated concept-level definition counts that concept once, a frequency effect the paper does not examine. Thresholding at the end may also undo a flip achieved in the relaxed space. The paper states that all reported validity checks use the final discretized sequence, which is the correct choice and means that failures are counted as failures. Algorithm 1 (p. 11) terminates when the total loss falls below a threshold or the iteration limit is reached, then maps the relaxed sequence to binary values and returns it. It does not display a final discrete-validity check or an explicit failure-return rule, although the accompanying text states that reported validity is evaluated after discretization. Validity is a measured outcome of a heuristic search and should be reported as a success rate and not described as guaranteed by the optimization.

7 Histories, Continuations, and Reachability

This section is conceptual framing, and the objections of the preceding sections do not depend on it. The gap described above can be restated in terms of three operations that are easily run together: changing a representation, extending a history, and making a state reachable. The paper does not instantiate any formal system for them, and they are offered because RA4KT is a concrete case in which the three operations appear interchangeable even though each requires different evidence.

An earlier incorrect answer remains part of a learner's history after the learner has understood the material. Subsequent learning adds an event to the record and does not make the original mistake disappear. Counterfactual reasoning can legitimately explore a branch in which that answer was correct, but the branch must remain distinguished from a continuation of the actual history. The educational interpretation of RA4KT needs a bridge between two different maps,

$$H_T \longmapsto \widetilde{H}_T \qquad \text{and} \qquad H_T \longmapsto H_T \parallel e_{T+1} \parallel \cdots \parallel e_{T+m}.$$

Here $\parallel$ denotes appending an event to the record. The first map substitutes a hypothetical past, and the second records a possible future learning process. The missing obligation is to show that an admissible extension of the actual history can produce the relevant competence, and the discovery of an alternative history that yields a favorable prediction does not discharge it. Retrospective counterfactuals are not illegitimate, but they must not be silently promoted into executable continuations. The same distinction can arise in tabular recourse when a snapshot conceals the process required to change an attribute, and educational sequences make the temporal structure explicit. Educational history is cumulative and path dependent, so substituting an earlier answer need not reproduce the state that later instruction would have generated, and the recorded error remains part of the actual history even where its educational consequences have faded.

The same issue appears from the standpoint of reachability. A desirable endpoint in response-vector space need not be reachable through the activities available to the learner. Changing a zero to a one is a small representational operation, whereas acquiring the competence needed to answer correctly may require several prerequisites, substantial practice, or resources the learner lacks. Hamming distance therefore measures the size of an edit to the record and does not establish the cost of reaching the corresponding educational condition. A search that aims at plans needs an account of admissible transitions and their costs, and a nearby representation can describe a state that is distant or inaccessible.

The treatment of the curriculum graph is a concrete instance of a distinction lost under projection. The method merges prerequisite relations and topical correlations into one undirected graph, although these relations support different inferences. The statement that two concepts are related does not preserve the statement that one must precede the other, and once direction has been discarded, a short path cannot in general reconstruct it. Two curricula can share the same undirected graph while permitting different learning orders. The later ordering procedure cannot recover the missing information merely by ending its path at the target, which is the content of the three-concept example given earlier.

The evidential boundary follows the same line. RA4KT can verify that a proposed response

history raises a fixed predictor's score and that its output agrees with a supplied curriculum structure, and these are properties of the computational construction. The authors acknowledge the boundary between verifying the generated object and verifying its claimed consequence in their offline-simulation caveats, and their stronger language sometimes crosses it.

Finally, treating an error primarily as a binary coordinate eligible for hypothetical replacement leaves its diagnostic content unmodeled. Counterfactual editing need not overwrite the actual record, and the point concerns what the intervention analysis leaves out. An incorrect answer can reveal something that a correct answer would conceal, such as a specific misconception, a failed procedure, or sensitivity to the wording of a question. A repair process could preserve that failure, record the attempted explanation or exercise, and test what changes afterward. The informative object would then be the trajectory from error through intervention to subsequent performance. Two learners who ultimately answer correctly may still require different continuations, because one guessed, one memorized a procedure, and one demonstrated transferable understanding. The educational task is accordingly to discover and verify a feasible continuation from the learner's actual history. Counterfactual answer editing can help generate hypotheses about that continuation, while historical fidelity, transition admissibility, and observed learning effects remain separate obligations.

8 The Defensible Claim and What Would Support More

Several parts of the design deserve credit. The combination of continuous relaxation with a single discretization step, hard projection enforcing the correctness-direction restriction, and a gradient-free treatment of the predictor is a legitimate and non-trivial engineering contribution, and the authors disclose the surrogate losses and the final-discretization evaluation openly. A curriculum graph is also a reasonable inductive bias for a diagnostic, since it steers the search toward concepts that are plausibly related to the target. None of the objections above depends on denying these points.

The defensible contribution is narrower than the motivation. RA4KT is a method for finding sparse, directionally constrained counterfactual answer histories for a trained knowledge-tracing model, and for converting the concepts attached to those histories into graph-guided suggestions about which topics bear on a target. As a diagnostic of what evidence a predictor responds to, this is informative, and the graph prior is a reasonable way to bias such a diagnostic toward curriculum structure. As a source of study plans, it is a hypothesis generator. The curriculum-agreement results support structural plausibility, but the paper does not establish that following the plans improves learning, and the structure of the evaluation prevents that question from arising.

Closing the gap requires evidence of a different kind, and it can be sought by either of two routes that may also be combined. One is a validated causal transition model that maps teaching actions to changes in learner state and is checked against observed outcomes, so that the quantity $\mathbb{E}[Y_{T+\Delta} \mid \mathrm{do}(A_T = a), H_T]$ is estimated and not assumed. The other is direct evaluation of the recommendation policy, for which the cleanest design is a randomized comparison in which learners are assigned to follow the generated path, a path of equal length chosen by a different

method, or ordinary practice, with outcomes measured on held-out items after the intervention. A policy evaluation of this kind can establish the benefit of a recommendation without identifying latent competence or fitting a transition model. Observational data can target the same causal effect, but they require stronger identification assumptions, and the confounding between who studies what and who is already likely to succeed is severe in educational logs. At the evaluation level, independent judgments of recommendation quality by teachers, who are not given the graph used to produce the plans, would address a different and more relevant question than agreement with the same graph. And at the baseline level, the constrained comparisons described above would show how much of the reported advantage belongs to the optimizer.

None of this makes the original idea unworthy. It makes the claim hinge on a measurement the paper has not yet made. A method that reports "the predictor can be pushed this way with this many edits" and a method that reports "learners will improve if they do this" are separated by a bridge that requires empirical justification beyond counterfactual optimization, and a paper that crosses it in its title and abstract owes the reader the evidence for the crossing.

9 A Design for Evaluating the Recommendation Policy

The preceding section named direct policy evaluation as one route to the missing evidence. A design for it can be stated concretely, because the questions it must answer follow from the separate claims the paper makes. The unit under evaluation is the policy that maps a learner's record and a target concept to an ordered set of study activities, and the comparison of interest is between that policy and alternatives that receive the same amount of instructional time. Evaluating a transition model would be a larger undertaking that depends on longitudinal data with recorded interventions, which typical educational datasets lack, and policy evaluation is the more tractable design for that reason.

Before randomization, each selected concept must be mapped to a specified activity module under a common allocation rule governing instructional time, feedback, and progression, since otherwise the trial evaluates an unspecified combination of recommendation and teaching-material design, the same unspecified link for which the preceding sections fault the original paper. Eligibility can mirror the population the paper itself selects, namely learners with a high error rate who are predicted to fail on a specified target concept. Randomization should be at the level at which contamination occurs. If learners within a class discuss or share materials, assignment by class or teacher is preferable to assignment by individual, and the analysis must then account for within-cluster correlation.

Four parallel arms address four distinct questions. A within-learner cross-over, in which different targets or periods are assigned to different arms, would reduce between-learner variance but raises carry-over concerns, since earlier study changes the later starting state, and the design is parallel for that reason. The first arm follows the generated path in its generated order. The second follows a path of equal length and equal time budget chosen by a different method, such as the expert prerequisite hierarchy applied directly to the learner's failed concepts. This contrast compares two complete recommendation procedures, and it isolates the contribution of counterfactual concept selection only if ordering and activity construction are otherwise

held constant. The third arm receives the same concepts as the first in a different order, which isolates ordering from concept selection and replaces the reversed-order control used in the paper's simulation with a comparison that can be run on learners. That contrast holds the selected concepts constant and tests the effect of ordering under a common instructional protocol. The activities, allocated time, feedback rules, and opportunities for repetition should be held constant across the two arms, and equal concept counts alone do not ensure comparable instructional exposure. A common protocol is preferable to identical realized exposure, because ordering may itself affect completion, requests for feedback, and repetition, and those differences can be mechanisms of the policy effect. Because the essay argues that prerequisite order may matter, the alternative orders must be restricted to sequences that preserve prerequisites or must pass educational review before use.

That restriction changes the question the contrast answers, and it also changes who can enter it. The ordering contrast is therefore a prespecified comparison within the eligible subgroup of learners whose selected concept sets admit at least two acceptable orders, while the other contrasts are estimated in the full randomized population. It tests whether the generated order adds value among educationally admissible orders, and it cannot test whether prerequisite compliance itself matters. The ordering comparison should therefore include only selected concept sets that admit at least two acceptable orders, with a predefined rule for cases in which the generated order fails review. If those cases are repaired, the evaluated policy is the method plus the repair procedure, and if they are excluded, the resulting claim applies to the eligible subset. The fourth arm continues ordinary practice with matched time on task, so that every learner receives reasonable instruction and no one is assigned to a condition expected to harm them.

The primary outcome should be performance on the target concept measured after a fixed interval under a common assessment protocol, using items withheld from the intervention activities and from any trial-driven tuning of the method. Independent learner outcomes break the dependence of the paper's simulation on the predictor as its own evaluator, and they do not require items the system has never seen, since a previously calibrated item can serve as an outcome measure if it has not been differentially practiced across arms. A separate item pool, with matched forms of comparable difficulty and concept coverage constructed in advance, is a stronger option for testing transfer and potential exposure effects. Secondary outcomes should include delayed retention, transfer to related concepts that were not named in the path, and time to reach a predefined mastery criterion. The analysis should follow the assigned arm regardless of adherence, adjust for a pretest on the target concept, report effect sizes with interval estimates in preference to significance tests alone, and examine heterogeneity by initial mastery, since the paper reports that its simulated gains differ across such groups. The protocol, outcomes, and analysis plan should be registered before data collection. Sample size should be calculated from the target effect or precision, the number and size of clusters, and the intracluster correlation, since the loss of precision under cluster assignment depends on those quantities. Educational interventions often produce modest effects, so the calculation may demand large samples, and the paper's simulated gains are correspondingly an expectation that a real trial would have to confirm.

The same trial affords a test of the predictor's role, but not through individual pre-post

improvement, which is not an individual treatment effect: it can reflect ordinary practice, maturation, test familiarity, or other changes that would have occurred under any comparator, and a parallel trial does not reveal both potential outcomes for any learner. The appropriate comparison is between the effects the simulator predicts and the effects the trial estimates, namely the between-arm contrasts, reported overall and within prespecified baseline strata with their uncertainty. The simulator's predicted contrast must concern the same comparator, the same outcome scale, and the same assessment interval as the trial, and predicted mastery probabilities must be mapped onto the observed assessment scale before the two can be compared. Close correspondence of the contrasts would be direct evidence that the simulation can serve as a proxy for the tested population, policies, assessment, and interval. Agreement on a single contrast can result from compensating errors and does not validate the simulator across other interventions, so agreement should be prespecified across several contrasts or baseline strata, although a small trial may estimate those comparisons too imprecisely to be decisive. A discrepancy large enough to be distinguished from sampling and simulation uncertainty would count against using the simulated gains as estimates of instructional effects in the tested setting, while leaving the method's diagnostic use intact.

10 Conclusion

A summary of the paper's standing requires four categories of claims that the RA4KT paper does not consistently distinguish. Some properties are enforced, such as the correctness-direction restriction through projection. Some are only encouraged, such as minimality and graph coherence through surrogate penalties. Some are supported by empirical agreement with supplied structure, such as the prerequisite satisfaction rate. And some, namely reduced learning burden, actionable recourse, and improved learning, remain unvalidated. The computational question is legitimate, and the objections here concern how far the published algorithm and its evaluation carry the stated claims.

Counterfactual search over answer histories answers a question about a model. Educational recourse answers a question about a learner. The two coincide only under assumptions about how instruction changes competence and how competence generates answers, and predictive training on logged responses does not, on its own, identify or validate those assumptions. RA4KT is open about this in several passages, and the more useful reading of the paper is the one those passages invite: a constrained, graph-informed procedure for interrogating a predictor, accompanied by a hypothesis about how its output might guide study. The surrounding framing, the metric glosses, and several reported details tend to blur that reading, and the statistical and algorithmic inconsistencies identified above should be resolved before the reported margins over baselines are relied upon. The decisive question requires evidence about intervention effects on learners, obtained through a credible experimental design, such as the one sketched in the preceding section, or an explicitly justified causal identification strategy.

References

Barocas, S., Selbst, A. D., and Raghavan, M. The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 80–89, 2020. https://doi.org/10.1145/3351095.3372830

Corbett, A. T., and Anderson, J. R. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1994. https://doi.org/10.1007/BF01099821

Karimi, A.-H., Schölkopf, B., and Valera, I. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–362, 2021.

Mothilal, R. K., Sharma, A., and Tan, C. Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617, 2020.

Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press, 2009.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., and Sohl-Dickstein, J. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems 28*, 2015.

Wang, L., Fu, M., Jian, Z., Tan, J., and Zhou, X. RA4KT: A Causality-Informed Explainable Reverse Attribution Framework for Knowledge Tracing. *Scientific Reports*, 16:36215, 2026. https://doi.org/10.1038/s41598-026-72278-4

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2):841–887, 2018.