A Principled Intelligent Occupational Training of Psychomotor Skills in Virtual Reality

Edward Kim
Zachary Pardos
Sanjit A. Seshia
Björn Hartmann

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-17
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-17.html

February 7, 2023

A Principled Intelligent Occupational Training of Psychomotor Skills in Virtual Reality

Edward Kim
EECS Department
UC Berkeley

Zachary Pardos
Education Department
UC Berkeley

Sanjit Seshia
EECS Department
UC Berkeley

Bjoern Hartmann
EECS Department
UC Berkeley

***Abstract*—The use of virtual reality (VR) for occupational training of psychomotor skills has been investigated for decades. Previous literature show that training in VR increases engagement, expedites learning, enhance safety, and improves skills in reality. Grounding on these progress, industry has begun to adopt VR to train their workers. However, we observe a disconnect between learning sciences and the current practice in VR training. Numerous literature across domains rely on self-assessment of the learners to predict whether they have reached a sufficient level, or mastery, of skill. However, it is well-established that self-assessment is inaccurate. Yet, there is no alternative for self-assessment to predict mastery of psychomotor skills. We propose to use bayesian knowledge tracing (BKT), a de facto standard in education to predict students' mastery of cognitive skills, for psychomotor skill mastery prediction. Using BKT, we design an intelligent occupational training system in VR. We conduct a between subjects study with 18 participants, with the control group relying on self-assessment to adjust curriculum and progression speed, while BKT is used instead for the experimental condition. Our results demonstrate the negative impact of self-assessment on psychomotor skill learning in VR, and shows the benefits of BKT as an alternative to self-assessment.**

***Index Terms*—learning sciences, human-computer interaction, formal methods, virtual reality, psychomotor skill, bayesian knowledge tracing**

I. Introduction

Over the past decade, there has been an increasing body of literature investigating the effectiveness of occupational training in immersive virtual reality (VR) to train high precision psychomotor skills (i.e. physical skills with cognitive planning) [51]. The literature spans various occupations such as healthcare [19], defense [2], manufacture [41], and sports [35]. These literature show that the VR training can increase engagement, expedite learning, enhance safety in training, and improves skills in reality. Grounded on these progress, a growing number of businesses are adopting VR to train their workers [34].

In this context, personalization, or individualized adaptation, of training contents displayed in VR to maximize learning and user experience becomes a practical design problem, which can positively impact across various workplaces. Then, a natural design question is: *who* should control the contents? The learners themselves? Or the training system? Numerous literature on VR-based occupational training systems are designed to let *learners* modify the training scenario contents by controlling the curriculum (i.e. the order of skills to train) and/or the progression speed (i.e. the rate at which you transition through skills in the curriculum). These design choices implicitly assumes that *the learners have accurate self-assessment of their skills during training*.

However, in learning science, it has been well-established that learners are inaccurate in the self-assessment of their skill mastery, or proficiency [11], [14]. Yet, currently there is a lack of principled mechanism grounded in learning sciences to replace self-assessment for an accurate skill mastery prediction. We propose to use bayesian knowledge tracing (BKT) [52], a de facto standard in education research to predict mastery of cognitive skills in domains like algebra, to predict psychomotor skill mastery. BKT is traditionally developed for purely cognitive skills that do not involve motor skills. Hence, one of key components of our study is to evaluate the suitability of BKT in its canonical form for psychomotor skill mastery prediction, and to identify which aspects of BKT to extend to improve its accuracy. Using BKT, we design and implement an intelligent occupational training system for psychomotor skills. The proposed system adapts the curriculum and the progression speed to individual needs.

We conduct a between subjects study with 18 participants who trained for ten different psychomotor skills. The control condition employs self-assessment and the experimental condition relies on BKT to adjust the curriculum and the progression speed. The control group represents the current common practices in VR-based occupational training literature. Our results show that the control group does improve on average in its accuracy as demonstrated in numerous literature. However, its fluctuations, measured by standard deviation, of mastery prediction error nearly doubled than the BKT's. Consequently, some participants fall behind with as low as 0% learning gain. In terms of accuracy, BKT's skill mastery prediction accuracy is not statistically different on average compared to self-assessment. However, its more consistent prediction leads to more uniform average improvement in learning gains, which is pedagogically useful for instructors to bring their students up to a similar level of skill to then tailor instruction to as a group. As a result, the experimental group had 32.3% higher learning gains with respect to the control group. We identify weaknesses of BKT for skill mastery prediction and suggests ways to improvement.

This work presents three primary contributions in VR-based occupational training: (1) re-evaluation of the current common practice of using learners' self-assessment to personalize train-

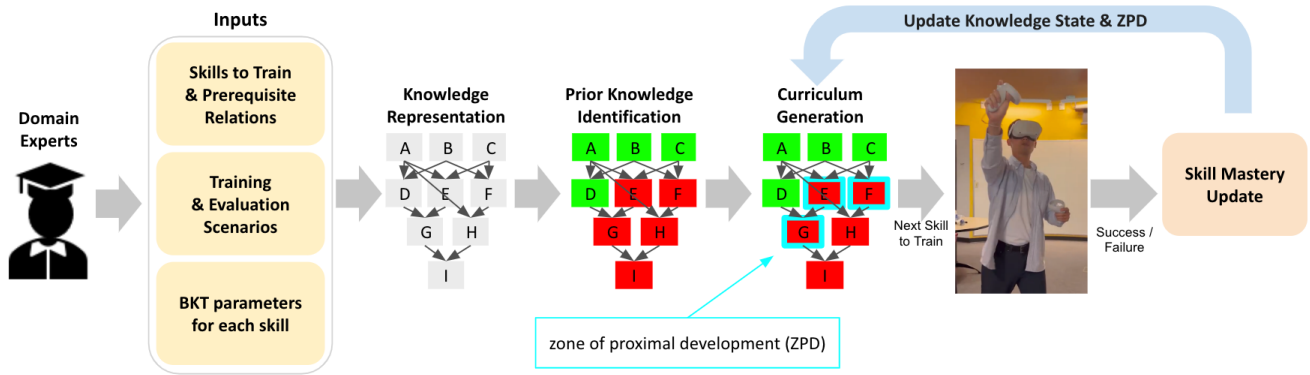

Fig. 1: An overview of our proposed intelligent tutoring system architecture for training psychomotor skills in virtual reality

ing contents, (2) a intelligent occupational training system in VR for psychomotor skills with a novel use of BKT for skill mastery prediction, and (3) user evaluation which demonstrates the negative impact of self-assessment on skill learning and the benefits of BKT as an alternative, principled mechanism for skill mastery prediction.

II. Related Work

A large body of literature have investigated different aspects of psychomotor skill training. Some focused on the construction of training simulators [20], [39]. Another investigated diverse forms of feedbacks to correct the learners by employing existing visual, tactile, and auditory haptic feedback [5], [43], [50], developing new media like augmented mirror [1], or accounting for social interactions [47] and cognitive science [21]. Others implemented physical devices (e.g. airRacket [45], tacTower [30]) to enhance sensory realism and engagement in training.

In comparison, our work assumes training in VR that a VR simulator and any feedback mechanisms are already provided for an occupational training. We focus on how to personalize, or individually adapt, the sequential contents of the training scenarios that each learner will experience in VR to optimize learning and user experience. In this vain, previous literature explored the *gamification* of training to increase engagement and motivation in VR [15], [46]. Others also investigated the personalization of a curriculum to increase the effectiveness of training [33], [40]. However, there is a lack of investigation in understanding the design implications of *which* entity should control the learning contents for occupational training in VR. In this absence, we observe a growing number of literature having learners control the contents despite the well-established results in the inaccuracy of self-assessment of their skills [11], [14].

For training systems to accurately assess learners' skill, previous literature focused on various task analysis techniques to formally specify evaluation metrics to measure the *correctness* of performance in solving tasks [7], [22], [32] designed to train or evaluate for particular skills. Yet, how many repetitions of practices each learner should have per skill to reach mastery has not been sufficiently investigated, even though it has been brought to attention more than a decade ago [3]. In this paper, we propose to use BKT [52] to address this problem of personalized repetition.

The design of our training system is built on top of curriculum personalization framework proposed in [48]. This work provided a generic framework to assess skills and personalize curriculum without relying on any dataset a priori. In general, across occupations, it is difficult to access a big data of dataset in relation to training. Hence, we adapted the curriculum generation aspect of this work to incorporate BKT to estimate skill mastery.

III. Background

A. Scenic: *Probabilistic Programming Language for Scenario Modeling*

To robustly train or evaluate a learner's mastery of psychomotor skills against realistic variations in the environment, we need to generate various scenarios in a VR headset for training or evaluation. For example, suppose we train a learner to throw a ball to a moving teammate. The teammate can be moving in various directions with different speeds. How can we efficiently model and generate such variations in a VR headset to train psychomotor skills? To address this issue, we used Scenic [13], a scenario modeling language whose syntax and semantics are designed to intuitively *model* and *generate* dynamic and interactive scenarios involving multiple agents and objects. More technically, Scenic is a probabilistic programming language that allows users to easily specify distributions over environment parameters (e.g. teammate's speed and moving direction). A Scenic program, therefore, models a distribution of concrete scenarios. A *concrete* scenario is defined as a tuple, (I_{v1}, B_{v2}), where I_v is an initial state consisting of different objects, their positions, orientations, etc., with concrete values, v1. B_v defines a behavior assigned to each object in the scenario, where each behavior is parametrized with concrete value, v2. For training or evaluation, we iteratively sample a concrete scenario and generate it in a VR headset.

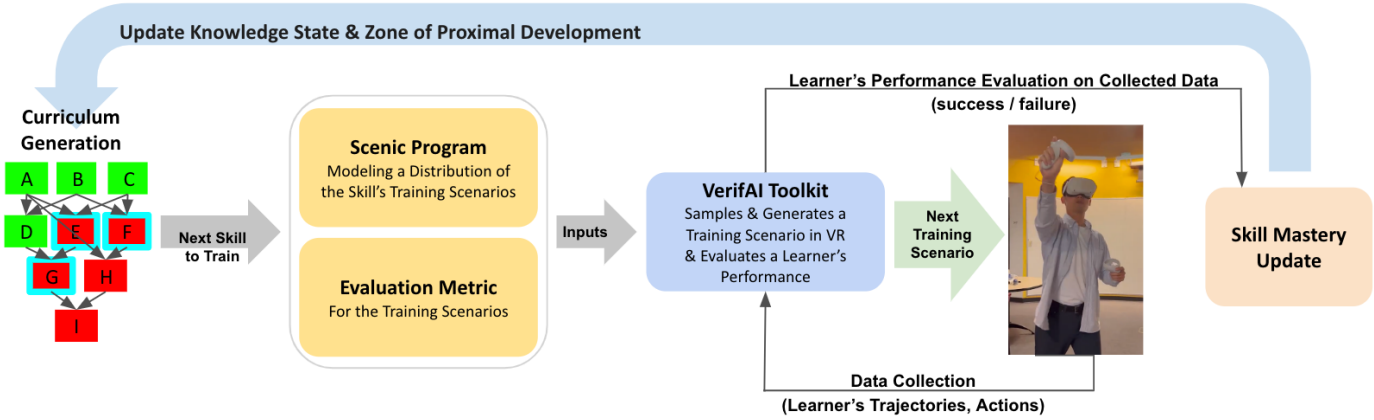

Fig. 2: This figure expands the curriculum generation aspect visualized in Fig. 1. Once we select which next skill to train for, we input a SCENIC program modeling a distribution of training scenarios for the skill and its evaluation metric to VERIFAI toolkit [10]. Then, VERIFAI samples a training scenario from the program, generates it in the learner's VR headset, and assesses the learner's performance using the given metric. This assessment is used to update the skill's BKT model.

B. VERIFAI *Toolkit*

For both training and evaluation, we need to assess the performance of a learner in VR using an evaluation metric. However, SCENIC cannot specify an evaluation metric. Hence, we use VERIFAI toolkit [10] which takes a SCENIC program and an evaluation metric as inputs. Then, as shown in Fig. 4, it samples a concrete scenario from the program, generates it in a VR headset, collects the learner's telemetry data (e.g., trajectory and actions), and evaluates the learner's performance by computing the given evaluation metric on the collected data.

C. *Bayesian Knowledge Tracing*

Bayesian Knowledge Tracing (BKT) [52] has become the standard in education research for modeling a student's mastery of a skill. BKT assumes a binary knowledge state, meaning that the student is either in the learned (i.e. mastered) or unlearned state with respect to a skill. It also assumes a binary-graded response from a student's attempt to solve a tutorial/training exercise (i.e., the student either correctly or incorrectly solved the exercise). The underlying statistical architecture of BKT is a hidden Markov model with observable nodes representing students' known binary response sequences obs_t to training exercises and hidden nodes representing students' latent knowledge state at a particular time step t. A canonical BKT model has four parameters: initial probability of knowing the skill a priori (prior), probability of student's knowledge of a skill transitioning from not known to known state after an opportunity to apply it (learn), probability to make a mistake when applying a known skill (slip), and probability of correctly applying a not-known skill (guess). For more detail, please refer to [52]. The mathematical definitions of these parameters and the Bayesian update rule is formulated below.

$$\text{prior} = P(L_0)$$

$$\text{learn} = P(T) = P(L_{t+1} = 1 | L_t = 0)$$

$$\text{guess} = P(G) = P(obs_t = 1 | L_t = 0)$$

$$\text{slip} = P(S) = P(obs_t = 0 | L_t = 1)$$

Note that while $P(L_0)$ denotes the prior parameter, we also define $P(L_t)$ as the probability that the student has mastered the skill at time step t. Bayesian Knowledge Tracing updates $P(L_t)$ given an observed correct or incorrect response to calculate the posterior with:

$$P(L_t | obs_t = 1) = \frac{P(L_t)(1 - P(S))}{P(L_t)(1 - P(S)) + (1 - P(L_t))P(G)}$$

$$P(L_t | obs_t = 0) = \frac{P(L_t)P(S)}{P(L_t)P(S) + (1 - P(L_t))(1 - P(G))}$$

The updated prior for the following time step, which incorporates the probability of learning from immediate feedback and any other instructional support, is defined by:

$$P(L_{t+1}) = P(L_t | obs_t) + (1 - P(L_t | obs_t))P(T)$$

IV. METHODOLOGY

We present the methodology for designing our intelligent tutoring system (ITS) for psychomotor skill training in virtual reality. To maximize the number of skills mastered, or learned, within a bounded training time, two adaptive mechanisms are deployed in our ITS.

First, we allocate the training time efficiently, by primarily focusing on the skills that the learners have not yet learned. At the beginning of the training, we characterize the prior knowledge (i.e. skills the learner mastered versus not mastered), which varies across learners. Then, we adaptively determine the order of skills to train from the skill not yet learned.

Second, we use Bayesian knowledge tracing (BKT) as the adaptive mechanism that estimates a learner's mastery of a particular skill. For additional details, please refer to the background (Sec. III-C). An accurate skill mastery estimation is crucial to determine *when* is appropriate to transition a

```
from scenic.simulators.vr.actions import *
from scenic.simulators.vr.behaviors import *
model scenic.simulators.vr.model

behavior teammateBehavior(endPoint, disc, trainee, reaction_distance, catch_radius):
    try:
        do Idle()
    interrupt when (distance from ego to disc) < reaction_distance:
        take MoveToAction(endPoint, TEAMMATE_SPEED)
    interrupt when (distance from disc to ego) < catch_radius
        take GrabDiscAction(True, CATCH_RADIUS)

# Define Regions
egoRegion    = MeshVolumeRegion(dimensions = (4, 4, 4), position = (0, 0, 0))
discRegion   = MeshVolumeRegion(dimensions = (2, 4, 4), position = (13, 0, 1))
tmRegion1    = MeshVolumeRegion(dimensions = (2, 2, 3), position = (9.88, -7.34, 1.2))
tmRegion2    = MeshVolumeRegion(dimensions = (2, 2, 3), position = (10.58, 7.34, 1.2))
region_list = [teammateRegion1, teammateRegion2]
random.shuffle(region_list)

# Define Objects using Regions
ego         = HumanPlayer in egoRegion, facing toward goal
disc        = Disc in discInitialRegion
destination = Point in reg_list[0]
teammate    = Teammate in region_list[1], facing toward ego,
                with behavior teammateBehavior(dest, disc, ego, 1.5, 1.5)
```

Fig. 3: An example of a SCENIC program (left) modeling a distribution of training scenarios for throwing a frisbee disc to a moving teammate and a snapshot (right) of a generated scenario in a VR headset where the teammate is moving as the learner is throwing a disc towards it

learner to the next skill for training. An underestimation and an overestimation both result in low number of mastered skills. The former results in an incomplete training, where the learner does not have a chance to train for certain skills at all. The latter results in covering all skills during training but none of the skills are mastered.

A. Domain Expert Informed Curriculum

We first recruit experts to gather domain knowledge. We ask domain experts to help build our curriculum by asking them to provide the following information: (i) a set of skills to train, (ii) prerequisite relations among the skills, and (iii) corresponding sets of training and evaluation scenarios for the skills, and (iv) parameters to tune a distinct knowledge tracing model for each skill. The details of our interaction with experts to gather these domain knowledge are explained in Appendix A.

1) Modeling and Generating Training and Evaluation Scenarios in VR: To model and generate training/evaluation scenarios with realistic environment variations and to evaluate a learner's performance, we used open-sourced SCENIC (Sec. III-A) and VERIFAI (Sec. III-B). For each skill, we encoded two SCENIC programs, each encoding a distribution of training/evaluation scenarios, respectively. Experts crafted these scenarios with specific tasks designed to train/evaluate particular skills. As shown in Fig. 2, for each skill, we input the corresponding SCENIC program and the task evaluation metric to VERIFAI which iteratively samples a scenario from the program, generates it in a VR headset to train/evaluate a learner, and measure the learner's performance according to the given metric.

2) Fine-tuning Knowledge Tracing Models: A BKT model estimates student mastery of a single skill as a probability. Hence, a separate BKT model is used for each skill. We used 0.99 as the probability threshold such that if the knowledge tracing model's estimate of mastery is greater than 0.99, then we determined that a learner mastered the skill. Typically, in traditional ITS, a threshold of 0.95 or 0.98 is used. We use 0.99 because of the high slip rate which increases BKT's skill mastery more steeply than for typical cognitive skills. Hence, we use 0.99 as a more conservative measure to ensure mastery. A BKT model requires four parameters to be tuned (refer to Sec. III-C). To tune these parameters for each skill, we asked the domain experts to respond to the following statements regarding the training scenario they provided for each skill in Likert 5-Point scale [28], assuming that the trainee already mastered the prerequisite skills.

1) There is a high chance a novice trainee will learn the skill after a single training exercise. (learn)
2) A trainee is likely to solve the task in a training scenario without having mastered the necessary skill. (guess)
3) Considering the complexity of the maneuvers that a novice trainee has to make to solve for the training scenario, a trainee is likely to make a mistake and fail to solve a task in this scenario even if they had already mastered the necessary skills. (slip)

Additionally, to determine the mapping from the Likert 5-Point scale to probability $\in [0, 1]$, we also asked the expert how many times *in a row* should a learner solve the training exercises to master each skill. We found a mapping such that if the learner answers the *first* three training exercises consecutively correct, then the KT model should output > 0.99. Regarding the "prior" parameter, we conservatively uniformly set it to 0.05 across all skills since we do not have data a priori for estimation. The practice of enlisting experts to help hand set BKT parameters based on expected skill learning trajectories, is not unique to our work. In the first few years of operation, this was the practice established by the Cognitive Tutor [37] for setting their skill parameter values, although data-driven refinements were proposed after substantial student response data had been collected [24], [38].

B. Knowledge Graph Generation

Knowledge is represented as a *knowledge graph*, i.e. a directed, pre-order graph as shown in Fig. 5. Given a set of skills to train and their prerequisite relations by the experts, we construct a knowledge graph. Each node represents a pscyhomotor skill to train and is associated with its distinct knowledge tracing model and SCENIC programs encoding abstract training and evaluation scenarios. The directed edges encode the prerequisite relation among skills such that the parent nodes pointing to other nodes are prerequisite skills to the children nodes being pointed at.

C. Prior Knowledge Identification

A knowledge state, as visualized in Fig. 5, is defined as a colored knowledge graph, where a binary color, red or green, represents mastered or not mastered skill, respectively. Learners have diverse prior knowledge where some may already have learned certain skills. We aim to identify the knowledge state of a learner to allocate more training time to skills not mastered yet.

There is an exploration/exploitation trade-off to consider where more time could be spent increasing our confidence in a student's prior knowledge of a particular skill by giving more assessments of that skill, but this would be at the expense of assessing prior knowledge in additional skills under a tight time budget. Hence, we *approximate* the prior knowledge using the provided prerequisite relations. If the trainee has already mastered a skill, then we estimate that there is a higher chance that the learner has mastered the prerequisite skills. In this case, we update the "prior" parameter of the prerequisite nodes' KT models to a higher probability in consultation with the experts. On the other hand, if the learner has not mastered a skill, then the *post-requisite* skills, i.e. the skills that have the unmastered skill as a prerequisite, are also likely not mastered. To reflect this, we keep the "prior" parameter of the post-requisite skills to the default value of 0.05. After updating the prior for all prerequisites or post-requisites, if the prior > 0.99, we color the node green, otherwise red.

Using these prerequisite relations, we can efficiently approximate the prior knowledge state if we carefully sample which node to assess. To account for time-efficiency, for each node that is not colored yet, we approximate how much time is saved if the learner already mastered a skill, s, by adding the time constraint for the skill as well as its prerequisites skills'. We denote this time as t_s^+. Similarly, we approximate time saved if the learner did not master the skill by adding the time constraint for the skill and its post-requisites'. We denote this time as t_s^-. Hence, for each uncolored node, s, the worst saved time is $\min(t_s^+, t_s^-)$. We sample for the uncolored node, which maximizes the worst saved time, for a time-efficient prior knowledge identification. We mathematically formulate the algorithm for sampling skill for prior knowledge identification in Equation (1).

$$s^* = \underset{\text{s is uncolored}}{\arg\max}\ \min(t_s^+, t_s^-) \tag{1}$$

D. Personalized Curriculum Generation

Zone of proximal development (ZPD) is a concept from psychology, which we adopt to generate a personalized curriculum. ZPD defines the "boundary zone" of human knowledge, which defines the zone that is not learned yet but has close relation with those already learned. Previous literature shows that, with activities selected from ZPD, students can learn on their own with little guidance from instructors [27], [29], and feel more engaged in learning [6].

As highlighted in light blue in Fig. 5, we define the ZPD to be a *set of red color nodes* that are either one edge away from the green nodes or red nodes with no prerequisite skill. An example of a knowledge state with a ZPD highlighted in light blue is shown in Fig. 5. From the ZPD set, we select for the *next* skill to train, which has the minimum number of prerequisites. If there are more than one such node, we choose the one with a shorter time constraint for its training scenario. We use these heuristics to expedite the training.

As shown in Fig. 2, once a node is selected, we generate a variable number of training exercises until the learner has mastered the skill according to its BKT model. We sample a training exercise from its corresponding SCENIC program and generate it in the VR headset. After each training exercise, we compute a Boolean to represent whether the learner solved the task or not, and update the BKT model with the Boolean outcome. Once the BKT model outputs a probability > 0.99, we update the color of the node from red to green, indicating mastery. Then, we update the ZPD set and select the next skill to train and generate a variable training exercises again until mastery. We repeat this process until either the training time expires or the ZPD set is empty.

V. EXPERIMENT

We conduct a study to understand the design implications of having *the learners versus the training system* control the training contents in VR. We specifically investigate their impact on learning gains and user experience.

A. Example Application Domain: Echo Arena VR Esports

As an example application domain, we select an esports called Echo Arena [26] which is a zero gravity, frisbee VR esports game owned by Meta [31]. Our rationale for the choice is the following. First, the characteristics of psychomotor skills that Echo Arena requires are general enough to represent those of many occupations. This esports demands gross (e.g. arms, legs, head) and/or fine (e.g. wrists, fingers) motor movements often under varying time limits. Hence, our findings in the design implications of training content control may have broader implications across workplaces. Second, there is no considerable VR-to-reality gap to consider since the application domain itself is in VR. Although skill transfer from VR to reality is crucial, we aim to first investigate whether the two contrasting designs have implications in VR training setting itself. If there is a significant difference in VR, then such finding would motivate future work on its impact on the skill transfer. Third, recruiting experts/instructors in Echo

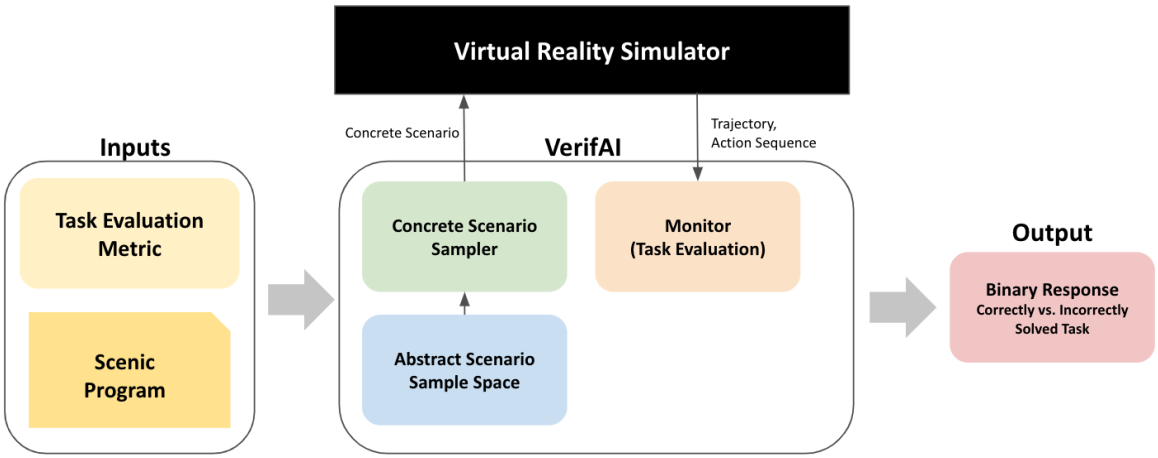

Fig. 4: VerifAI is an open-source tool that we use to generate training/evaluation scenarios in VR and assess a learner's performance in those scenarios. Its architecture is visualized in this figure. VerifAI takes as inputs a SCENIC program and an evaluation metric. It samples iteratively a concrete scenario from the program, generates it in VR, and outputs the learner's performance in the training scenario.

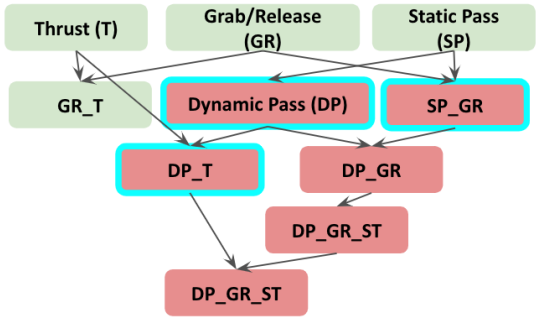

Fig. 5: We represent a knowledge state as a colored, acyclic, directed, pre-order graph as visualized in this figure. Each node represents a skill. The directed edges encode prerequisite relations. The color represents mastery (green: mastered, red: not mastered). The zone of proximal development (ZPD) highlighted in light blue is a set of not mastered skills that are *in proximity* to mastered ones.

Arena is relatively easier and *does not warrant any legal debate between organizations*. For these reasons, we select Echo Arena as our example application domain to conduct our study. We reconstruct Echo Arena in Unity [16] and interfaced SCENIC (refer Sec. III-A) to model and generate the desired training and evaluation scenarios in VR.

B. Experts/Instructors Recruitment

We recruit four professional Echo Arena esports players via direct messaging in Discord [9], who provide us with necessary inputs (refer Sec. IV-A) to our training system through 2 hours of joint video call. Each professional is paid $50 for their time and inputs. These professionals have achieved the top 10 in ranking over the last few years in the VR Master League [26], which hosts the largest annual Echo Arena tournament. For context, in the most recent tournament in 2022, nearly 8,000 people around the world joined the competition [25]. These four experts also had experience in coaching novice or amateur Echo Arena players.

C. Participants

We recruit participants through university online forums and mailing lists from a community of VR users. We receive 25 responses of people with prerequisite dynamic VR game experience. Out of the 25 respondents, we exclude 7 people according to three *pre-determined* exclusion criteria: 1) exhibiting motion sickness, 2) too much skill expertise (no opportunity for learning) and 3) extreme lack of hand-eye coordination (unlikely to master any skill during our short training session). Eligibility criteria are listed in supplemental material (Appendix B). The accepted 18 participants' ages range from 19 - 25 years old, with 4 females and 14 males. Each participant is financially compensated with $40 gift card for their 2 hours of participation. For the participants who are excluded according to our pre-determined criteria, they are compensated for the time they participate at $20 per hour rate.

D. Procedure

We conduct a between subjects experiment to avoid learning and fatigue effects. We randomly divide the accepted 18 participants into two disjoint groups, i.e. the control and the experimental groups, with 9 participants in each condition. The study is conducted individually, not in groups. The study consists of the following procedures: tutorial (5 min), pre-test (15 min), training (25 min), post-test (15 min), and exit questionnaire (5 min), with 10 min breaks in between parts including the half way through the training session. The details of these procedures are explained in Appendix C. Per skill, the pre/post tests examine each skill 3 times by randomly sampling 3 concrete scenarios from the corresponding SCENIC

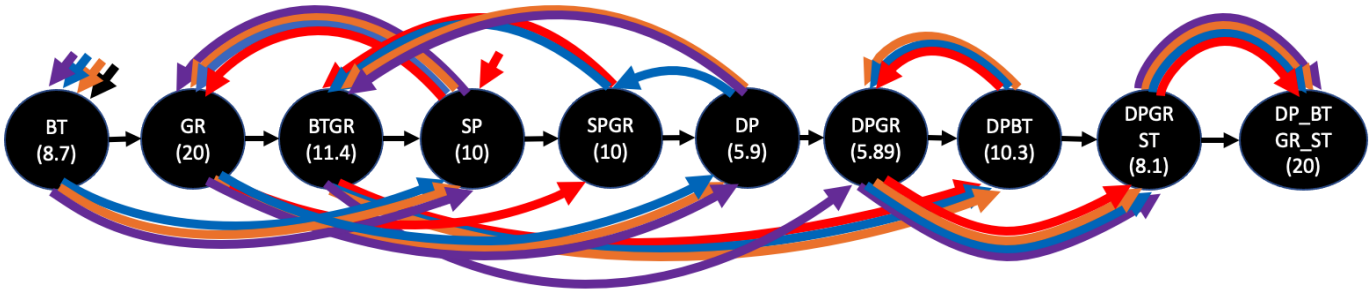

Fig. 6: A comparison in curricula between the control and the experimental groups. The black circles represent the skills. The control group's single curriculum is visualized as a sequence of skills traced by the black arrows. The experimental group's diverse, personalized curricula are shown with colored arrows. The short, straight line segments pointing at skills, BT and SP, mark the start of different curricula.

programs modeling distributions of evaluations scenarios (refer Sec. III-A). Three is the maximum number of tests we could afford within our time bound study to examine the mastery of each skill.

Both conditions follow these same procedures and train for the same set of 10 skills (a video of a post-test visualizing these ten skills can be found in this drive). The only difference is *the entity* that controls the training contents, namely the curriculum (i.e. the order of skills to train) and the progression speed (i.e. the rate at which you transition through the skills in the curriculum). In the control group, *the learners* control the training contents, hence self-guided. Additionally, the control group is provided with the experts' suggested curriculum to train on (as is the case in many occupational VR training), but can alter the curriculum as they see fit. In contrast, in the experimental group, *the training system* controls the training contents. After each training scenario, we ask the participants in both groups to report their binary self-assessment (i.e. mastered vs. not mastered) for the skill they are training at the moment. During training, we collect learners' binary performance score (i.e. correct or incorrect) for each training / evaluation scenario, binary self-assessment, BKT's probability output of [0,1], and the time stamp of when these data are collected.

E. Measurements

1) Relative Quantitative Comparison: Because access to our limited sample data is expensive, we report the experimental group's quantity *with respect to* the control group's quantity to better appreciate the differences. For example, suppose the learning gains of the control group is 50% and the experimental group, 60%. Then, the experimental group improves $25\% = (60 - 50)/50$ = (experimental's quantity - control's quantity) / (control's quantity) x 100 ($\%$) with respect to the control group.

2) Learning Gains: A learning gain for a participant is computed by one's score improvement (i.e. post test - pre test scores), where the pre and post test scores are computed in the following way: $\sum_{i=1}^{10}$(# of successes for skill i in the test)$/3$. Recall that each participant is evaluated on 10 different skills where each skill is evaluated 3 times in the pre/post tests (refer Sec. V-D).

3) Statistical Significance Test: We use Mann-Whitney's U test using Python Scipy's stats package [23] for all the statistical significance tests reported in the Results Section. We choose this test because, although both the control and the experimental groups are sampled from the same population, the sample size is too limited to expect normal distributions to hold for unpaired t-test. In case when statistically non-significant results are found, we conduct bayes factor analysis [8] to further check if null hypothesis is accepted. The bayes factor quantitatively represents the degree of *closeness* to accepting the null hypothesis when having a statistically non-significant result. We use the online calculator [44] to compute the bayes factors.

4) User Experience: We equally ask participants in both groups to fill out NASA task load index (TLX) [17] and a verbal interview. The NASA TLX is answered before and after their training. The verbal interview is conducted after the post tests.

During our verbal interview, we asked own questionnaire below to inquire specific attributes of curricula and progression speed through the curricula.

1) The training session was engaging.
2) The training session was incrementally challenging.
3) The training has helped me learn physical skills in virtual reality.

A table listing out the 5 point scale and their meanings, i.e. strongly disagree, disagree, neutral, agree, strongly agree) was provided underneath each statement. In addition, we further ask participants for any negative experiences during training.

5) Skill Mastery Estimation Error: This error is computed using the difference between the expected and actual post test score for the mastered skills for a participant, i.e. $N - \sum_{i=1}^{N}$(# of successes for skill i in the test)$/3$ where $N \leq 10$ is both the number of mastered skills and expected score for each participant.

F. Results

1) Effectiveness in Learning Gains: Prior to comparing the learning gains between the two conditions, we check whether

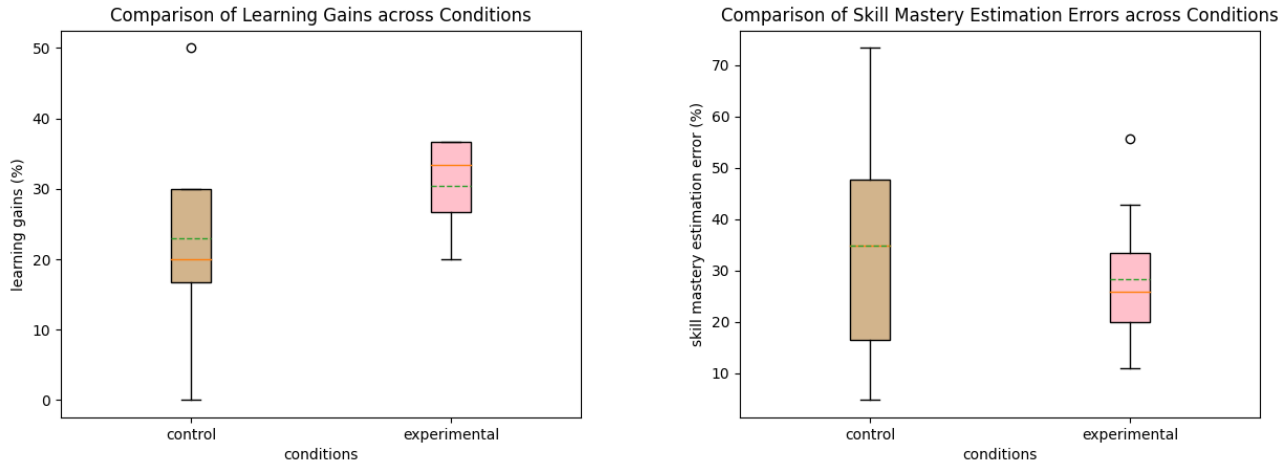

Fig. 7: The left box plot shows that the experimental group had higher learning gains in comparison to the self-guided control group. The right box plot shows that the bayesian knowledge tracing models have lower error in estimating skill mastery when compared to learners' self-assessment in the control group. The green dotted line in the box plot represents the average and the orange line, the median.

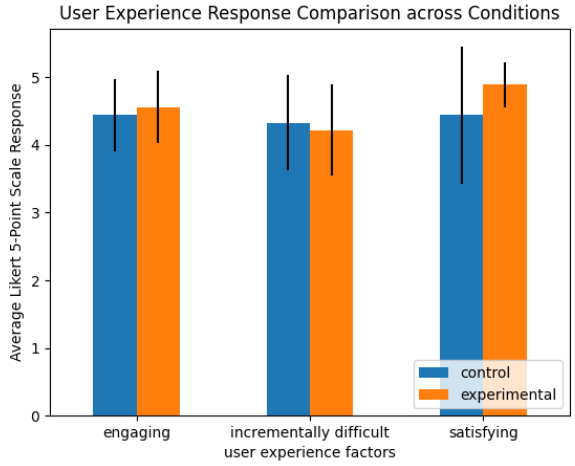

Fig. 8: The bar plot compares the learners' Likert 5-point scale responses on their learning experience. The experimental group's responses are on par with the control's, where both reported positively on their learning experience.

there is any imbalance in the prior skills between the two conditions. The difference in the distributions of the pre-test scores is not statistically significant (p-value 0.26). Regarding learning gains, the experimental group outperform the control group on average with statistical significance (p-value of 0.04) as shown in Fig. 7 with an effect size of 0.41. On average, the control group improve $22.96 \pm 12.90\%$, whereas the experimental group improve $30.37 \pm 5.97\%$. The experimental group experience $32.3\% = (30.37 - 26.67)/26.67$ higher learning gain with respect to the control group. Furthermore, the standard deviation of the experimental group's learning gains is 46% less with respect to the control group's.

2) User Experience: After the post tests, we equally ask participants in both conditions to share their user experience using NASA task load index (TLX) and and a verbal interview.

Regarding verbal interview, we ask three questions as listed in V-E4. Both conditions positively rate their training experience as summarized in Fig. 8. Mann-Whitney U test show that the differences in distributions across conditions for engagement, incremental difficulty, and helpfulness are not statistically significant, reporting p-values of 0.86, 0.43, and 0.34, respectively, and bayes factors were 0.33, 0.30, and 0.96, respectively. Hence, our statistical analysis shows that the user experience is *on par* regarding engagement and incremental difficulty attributes, but inconclusive for helpfulness aspect.

Despite the experimental group's higher average learning gains, the average NASA TLX score improvement after training for the experimental group is nearly a third of the control group's. The experimental group improved 6.56 ± 16.00, while the control group improved 17.56 ± 11.77. During the verbal interview, the control group (participants denoted as C1-C9) and the experimental group (E1-E9) share feedback that could explain the gap between learning gains and NASA TLX scores. We inquire participants in both groups for negative experience about the training. The experimental group share conflicting feedback related to the *progression speed* of the training. Some participants share frustration from too many assigned practices for a specific skill:*"I got frustrated towards the end because I was stuck in a task"* (E3) and *"getting stuck in a task was a bit frustrating in the beginning, but frustration went down as I saw myself improving"* (E5). On the contrary, some report unexpected early transitions: *"sometimes, the training algorithm transitioned you a bit earlier than you expected"* (E6) and *"during the training, I thought I still needed some more practice, but during evaluation I actually performed*

better than I expected" (E1). Hence, either the frustration or the early transitions could have likely affected the participants perception of the skills. The rest of the participants (E2,E4,E7-E9) report no negative feedback. The control group did not report any negative feedback.

3) Skill Mastery Estimation: We compute the skill overestimation error for each condition for the skills that participants report to have mastered *during training*. For the control group, the mastered skills are identified using the binary self-assessment that the learners report during training after each training scenario. For the experimental group, they are the skills which the corresponding bayesian knowledge tracing (BKT) models output greater than 0.99 (refer to Sec. III-C).

Comparing the average overestimation errors, BKT shows higher accuracy in skill mastery estimation on average than the learners' self-assessment as shown in Fig. 7. BKT overestimates participants' skill mastery by $28.21 \pm 13.06\%$, whereas participants in the control group overestimate their own skill mastery by $34.81 \pm 23.67\%$. Note that the standard deviation for BKT's error is 55% decreased with respect to the control's. However, Mann-Whitney's U test shows that this result is not statistically significant (p-value 0.46). We further investigate the effect of this considerable reduction in the standard deviation. We correlate (using Pearson correlation [12]) the participants' actual scores to the corresponding expected scores for the control and the experimental groups, respectively. The BKT models shows much higher correlation coefficient of 0.96 (p-value < 0.01) than the self assessment's 0.59 (p-value 0.09). In short, BKT is $62.71\% = (0.96 - 0.59)/0.59$ more highly correlated to the consecutive successful demonstration of skill mastery with respect to self-assessment.

4) Curriculum Generation: We observed that our system personalized curricula better optimized training time efficiency compared to the control group. The comparison of the emerged curricula between the two conditions is visualized in Fig. 6. We observed that the control group all *uniformly* adhered to the expert's fixed curriculum (highlighted in black arrows and the black nodes represent skills) despite the freedom to change it. In contrast, our training system generated various personalized curricula, automatically *skipping* over skills (e.g. BT, DP, SPGR, DPBT) participants already mastered to focus the training on the skills yet mastered.

The personalized curriculum reduced the time wasted for redundantly training on already mastered skills. In the control group, the participants spent on average 11.67 ± 10.21 trials on skills they already mastered (i.e. scored 3 out of 3 in the pre-test) out of 66.33 ± 14.53 total average number of trials they experienced during the training session, using close to 16.67% of the total trials on skills they already mastered. On the other hand, the experimental group spent 6.57%, using 4.16 ± 4.53 trials on average on mastered skills out of the total average number of 63.33 ± 8.77 trials during the training session. We report time in terms of trials, not minutes, for fairness because each skill's training scenarios required different training time on average, ranging from 6 to 20 seconds. Depending on which skills a participant mastered, 1 trial for a skill may worth 3 trials for another skill in regards to time.

VI. Discussion

In this paper, we question the current design choices in the personalization of VR-based occupational training to let *learners* control training contents. We base our suspicion on the well-established literature in learning sciences demonstrating the inaccuracy of learners' self-assessment [11], [14]. Therefore, we hypothesize that learning gains would be maximized if *the training system*, not learners, control the training contents, namely the curriculum and the progression speed. To test this hypothesis, we design a training system that controls the training contents and does not yield or share control with the learners. The key component of this system is skill mastery estimation specifically related to the problem of repetition: *how many times a learner should correctly solve training tasks, designed for a particular skill, with realistic variations in order to achieve mastery? And, without prefixing this number, can we adaptively and systematically adjust the number of repetitions per skill for each individual based on one's performance during training time?* We conduct a between subjects study to investigate the impact of learners versus system-controlled training on the learning gains and user experience.

Learning Gains: Our study shows higher average learning gains when the *training system*, not learners, assumes full control of training contents. Our results show that the experimental group achieves 32.3% higher learning gains (p-value 0.04) with respect to the self-guided baseline with the effect size of 0.41. Furthermore, the system's control over contents considerably *reduces* the fluctuations in learning, as evidenced by 46% reduction in the standard deviation of learning gains in the experimental group. In Fig. 7 (left plot), we observe much higher fluctuations in the control group's learning gains than those of the experimental group. For occupational training, this is pedagogically useful for instructors to bring their students up to a similar level of skill to then tailor instruction to as a group. Our findings are consistent with Corbett and Anderson's literature concerning the control problem [4], where learners have the highest learning gains when the training system has the most control over its content sequencing and interventions in an academic setting with no motor skills involved. This prevents the inaccuracy of self-assessment negatively affecting the learning process. Our study also reveals a consistent message but in psychomotor skill training setting.

Skill Mastery Estimation: The control group's underperformance in learning gains is attributed to inaccuracy and inconsistency of learners' self-assessment of skill mastery. Note that, for both conditions, accurate skill mastery estimation is crucial for adequate control of the curriculum and the progression speed. We observe that BKT and self-assessment are both inaccurate with nearly 30% average estimation errors with no statistically significant difference. However, the control group's self-assessments have much higher variations in its estimation errors. In Fig. 7 (right plot), we observe the control group's much wider range of skill estimation

error. The BKT considerably reduces the standard deviation of the skill mastery estimation error by 55% with respect to the self-assessment. Consequently, BKT is 62.71% more highly correlated to skill mastery than self-assessment. Hence, the inaccuracy and the wide individual variations in their inaccuracy contributed to the control group's lower learning gains.

User Experience: We measured user experience via (1) NASA-TLX, (2) verbal interview, and (3) custom questionnaire after the post-test. In both conditions, participants report positive overall training experience to our custom questionnaire, with approximately 4.5 out of 5 point likert scale on average across the three aspects of training as shown in Fig. 8. Despite the system taking control over the training contents, the experimental group's average ratings on the engaging and the incrementally difficult aspects of the training are *on par* with the control group's.

However, despite the experimental group's higher average learning gains, they showed *noticeably lower* average improvement than the control group in their subjective perception of skills after training. The self-guided group improved in the NASA TLX score by $17.56 \pm 11.77\%$ on average, whereas the experimental group only improved by 6.56 ± 16.00. The contrasting verbal interview results between the two conditions reveal the limitations of withholding control from learners. While the control group reported no negative experience with training, nearly half of the experimental group participants experienced frustration or undesired transitions. As we report in the Results section, the participants (E3,E5) repeatedly use the word "stuck" to share their frustration from not being able to stop excessive training one a particular skill. Also, the "early transitions" (E1,E6) to new skills, contrary to the learners' expectations, left them feeling unprepared. This discrepancy in objective and subjective (or perceived) learning has been observed in academic learning setting [42], where the condition with the highest objective learning gains has the lowest perceived learning gains.

This noticeable discrepancy between the learning gains and the perception of skills opens up a new design optimization problem for a time-efficient, effective, and user-friendly occupational training in VR. There needs to be further design explorations to carefully determine the learners' degree of control over the learning process, in order to optimize for both higher learning gains and better perception of skills. In the remaining discussion, we suggest challenges to consider for the design exploration.

A. Challenges of Skill Mastery Estimation & Our Suggestions

We provide our suggestions and insights from our study to help improve personalization of VR-based occupational training. From our study, we observe two issues (i.e. frustration and early transitions) when controlling the training contents, stemming from the system's exclusion of learners' feedback. We describe the potential challenges to consider when addressing these two issues.

To lower the learners' chance of frustration, there are important factors to consider. First, the pre-requisite relations imposes the order in which skills should be *mastered*. Consequently, if we design the training system to violate this order to avoid frustration, then this may result incurring more frustration as time progresses. For example, suppose after a learner fails to solve training tasks multiple times, the system transitions the learner to train for the next skill whose pre-requisite is the current skill. The learner is now *overloaded*, having to learn both the current and the next skill, simultaneously. This may likely result in accumulation of frustration, contrary to the motive. Second, the violation of the pre-requisite relation also poses an issue in skill mastery estimation. This violation may invalidate the assumptions of the system designers and the instructors/experts. For example, in our system, we tune the parameters of each BKT model per skill with the experts *under the assumption* that each skill's pre-requisite skill(s) are already mastered. Hence, violating the pre-requisite relations to avoid frustration may degrade the BKT's accuracy. To circumvent these issues, scaffolding [18] could help the learner fully master the skill before transitioning to the next skill. However, scaffolding each training scenarios to easier ones could be labor intensive, especially as the number of skills scales.

To prevent "early transitions" (E1,E6), it may be appropriate to accept learner's self-assessment. Once BKT determines a skill is mastered, then the system could ask the learners whether they are ready to move on. If they are not, then provide further practices for the skill until the learner is ready. However, this comes at the risk of, in worst case, consistent underestimation of skill mastery, resulting in redundant training due to the learner's inaccurate self-assessment. For this reason, it may be reasonable to explore *sharing* the BKT's estimate of skill mastery on VR display during training. This way, the learners do not have to solely rely on their self-assessment.

Lastly, although BKT was more accurate on average than self-assessment, the difference in distributions was not statistically significant. Hence, there is a room for improvement in BKT's accuracy. Recall that we only used crude *binary* (i.e. correct/incorrect) performance feedback to update BKT models (refer to Sec. III-C). However, using telemetry data collected in VR, we could compute how "close" (i.e. partial credit) the learner is to correctly solving the training task, which provides a richer signal than binary to update the BKT models. To enable this, the current traditional BKT model will need to be adapted to accept non-binary partial credit feedback. Deep knowledge tracing (DKT) [36] may provide a suitable framework for this effort. Yet, due to DKT's reliance on neural networks, this will require data.

B. On Interacting with Experts/Instructors & Suggestions

We recommend tuning BKT parameters for skills with more than one expert to reduce subjective bias. This is also a recommended practice in academic education to improve instructors' reliability [49]. During our joint discussion with the experts, each expert's suggested BKT parameters disagreed

by more than 2 or 3 likert scale points on a number of skills. At each occasion, the experts corrected each other's bias through discussions and converged to closer parameters.

In hindsight, we suggest to have the experts not directly insert the BKT parameters on a shared document to further reduce bias. For each skill, experts took different lengths of deliberation time to propose BKT parameters. Consequently, some wrote down the parameters earlier than others on a shared document, potentially influencing or even biasing others who wrote down parameters afterwards. Hence, we suggest that the experts share the parameters separately to the researcher, and the researcher should notify the experts to jointly discuss and correct any bias only when there is a notable discrepancy in the proposed parameters.

VII. Limitations & Future Work

There are a number of limitations in our study. Although we do not experience this issue in our study, it can be challenging to extract *tacit* domain knowledge from the experts to specify accurate evaluation metrics. We do not fully explore a methodological approach to cope with this difficult problem. Furthermore, the transfer of skills from VR to reality is a critical component that is yet investigated in our study. However, the results of our work motivates the need to explore further impact of self-assessment in skill transfer.

Another major limitation of VR-based occupational training is motion sickness. Depending on the dynamic nature of the vocation, the degree of motion sickness may vary. In our study, we train learners in extremely dynamic esports setting, which results in two participants dropping out due to motion sickness. To be *inclusive* of all participants in occupational training, it would be necessary to explore effective methods to (gradually) adapt participants to VR in order to reduce motion sickness. Another feasible direction to lower motion sickness is to use mixed reality (MR), where virtual images or avatars were overlaid on top of or interact with real physical objects. However, this realistic physical interactions between the virtual and the real objects impose constraints on the types of training scenarios one can generate in MR. Hence, for our future work, we plan to extend our work to MR for more inclusive, VR-based occupational training across workplaces.

VIII. Conclusion

This work re-evaluates the current common practice of using learners' self-assessment to personalize training contents in VR-based occupational training. It identifies the disconnect between the practice and the well-established finding in learning sciences that self-assessment is inaccurate. To address this discrepancy, we propose to use BKT as an alternative for self-assessment in predicting skill mastery for psychomotor skills. Using BKT, we design an intelligent occupational training system in VR. Our study shows that BKT, although it is traditionally designed to predict purely cognitive skills, it is still better suited for mastery prediction than self-assessment. In the future, we plan to extend BKT with physical factors such as fatigue to enhance its accuracy and, ultimately, provide a generic means for skill mastery prediction in VR.

References

[1] Anderson, F., Grossman, T., Matejka, J., and Fitzmaurice, G. Youmove: enhancing movement training with an augmented reality mirror. In *Computer Human Interaction (CHI) Conference on Human Factors in Computing Systems* (2013).

[2] Bhagat, K. K., Liou, W.-K., and Chang, C.-Y. A cost-effective interactive 3d virtual reality system applied to military live firing training. In *Virtual Reality* (2016), vol. 20, pp. 127–140.

[3] Brunner, W. C., Korndorffer, J. R., Sierra, R., Massarweh, N. N., Dunne, J., Yau, C., and Scott, D. J. Laparoscopic virtual reality training: Are 30 repetitions enough? In *Journal of Surgical Research* (2004), vol. 122, pp. 150–156.

[4] Corbett, A. T., and Anderson, J. R. Locus of feedback control in computer-based tutoring: impact on learning rate, achievement and attitudes. In *Conference on Human Factors in Computing Systems (CHI)* (2001), pp. 245–252.

[5] Correia, N. N., Masu, R., Primett, W., Jurgens, S., Feitsch, J., and da Silva, H. P. Designing interactive visuals for dance from body maps: Machine learning and composite animation approach. In *Designing Interactive Systems (DIS)* (2022).

[6] Csikszentmihalyi, M., and Csikszentmihalyi, I. S. *Optimal experience: Psychological studies of flow in consciousness*. Cambridge University Press, 1988.

[7] Demirel, D., Butler, K. L., Halic, T., Sankaranarayanan, G., Spindler, D., Cao, C., Petrusa, E., Molina, M., Jones, D. B., De, S., and deMoya, M. A. A hierarchical task analysis of cricothyroidotomy procedure for a virtual airway skills trainer simulator. In *The American Journal of Surgery* (2016), vol. 212, pp. 475–484.

[8] Dienes, Z. Using bayes to get the most out of non-significant results. In *Frontiers in Psychology* (2014), vol. 5.

[9] Discord Inc. Discord.

[10] Dreossi, T., Fremont, D. J., Ghosh, S., Kim, E., Ravanbakhsh, H., Vazquez-Chanlatte, M., and Seshia, S. A. Verifai: A toolkit for the formal design and analysis of artificial intelligence-based systems. In *International Conference on Computer Aided Verification (CAV)* (2019).

[11] Falchikov, N., and Boud, D. Student self-assessment in higher education: A meta-analysis. In *Review of Educational Research* (1989), vol. 59, pp. 395–430.

[12] Freeman, D., Pisani, R., and Purves, R. Statistics. *WW Norton & Company (4th Edition)* (2007).

[13] Fremont, D. J., Kim, E., Dreossi, T., Ghosh, S., Yue, X., Sangiovanni-Vincentelli, A. L., and Seshia, S. A. Scenic: A language for scenario specification and data generation. *Machine Learning Journal* (2022).

[14] Gordon, M. J. A review of the validity and accuracy of self-assessments in health professions training. In *Academic Medicine* (1991), vol. 66, pp. 762–769.

[15] Graafland, M., Schraagen, J. M., and Schijven, M. P. Systematic review of serious games for medical education and surgical skills training. In *British Journal of Surgery* (2016), vol. 99, pp. 1322–1330.

[16] Haas, J. K. A history of the unity game engine.

[17] Hart, S. G., and Staveland, L. E. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds., vol. 52 of *Advances in Psychology*. North-Holland, 1988, pp. 139–183.

[18] Hogan, K., and Pressley, M. Scaffolding student learning: Instructional approaches and issues. In *Brookline Books* (1997).

[19] Hogg, M. E., Tam, V., Zenati, M., Novak, S., Miller, J., Zureikat, A. H., and Zeh, H. J. Mastery-based virtual reality robotic simulation curriculum: The first step toward operative robotic proficiency. In *Journal of Surgical Education* (2017), vol. 74.

[20] Ipsita, A., Erickson, L., Dong, Y., Huang, J., Bushinski, A. K., Saradhi, S., Villanueva, A. M., Peppler, K. A., Redick, T. S., and Ramani, K. Towards modeling of virtual reality welding simulators to promote accessible and scalable training. In *Computer Human Interaction (CHI) Conference on Human Factors in Computing Systems* (2022).

[21] JENSEN, M. M., RASMUSSEN, M. K., AND GRONBAEK, K. Design sensitivities for interactive sport-training games. In *Designing Interactive Systems (DIS)* (2014).
[22] JOHNSON, S. J., GUEDIRI, S. M., KILKENNY, C., AND CLOUGH, P. J. Development and validation of a virtual reality simulator: Human factors input to interventional radiology training. In *Human Factors and Ergonomics Society* (2011), vol. 53, pp. 612–625.
[23] JONES, E., OLIPHANT, T., PETERSON, P., AND ET AL. SciPy: Open source scientific tools for Python, 2001.
[24] KOEDINGER, K. R., MCLAUGHLIN, E. A., AND STAMPER, J. C. Automated student model improvement. In *International Conference on Educational Data Mining (EDM)* (2012).
[25] LEAGUE, E. A. V. M. Statistics, 2022.
[26] LEAGUE, V. R. M. Echoarena vr master league, 2022.
[27] LEE, C. D. *An Introduction to Vygotsky*. Routledge, London, 2005.
[28] LIKERT, R. A technique for the measurement of attitudes. *Archives of Psychology 22*, 140 (1932).
[29] LUCKIN, R. Designing children's software to ensure productive interactivity through collaboration in the zone of proximal development (zpd). *Information Technology in Childhood Education Annual 2001*, 1 (August 2001), 57–85.
[30] LUDVIGSEN, M., FOGTMANN, M. H., AND GRONBAEK, K. Tactowers: An interactive training equipment for elite athletes. In *Designing Interactive Systems (DIS)* (2010).
[31] META. Meta: Social metaverse company, 2022.
[32] NATHANAEL, D., MOSIALOS, S., VOSNIAKOS, G.-C., AND TSAGKAS, V. Development and evaluation of a virtual reality training system based on cognitive task analysis: The case of cnc tool length offsetting. In *Human Factors and Ergonomics in Manufacturing Service Industries* (2016), vol. 26, pp. 52–67.
[33] NEAGU, L.-M., RIGAUD, E., GUARNIERI, V., DASCALU, M., AND TRAVADEL, S. Selfit v2 – challenges encountered in building a psychomotor intelligent tutoring system. In *International Conference on Intelligent Tutoring Systems (ITS)* (2022).
[34] NIJLAND, T. The 22 best examples of how companies use virtual reality for training, 2023.
[35] OAGAZ, H., SCHOUN, B., AND CHOI, M.-H. Performance improvement and skill transfer in table tennis through training in virtual reality. In *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS* (2022), vol. 28.
[36] PIECH, C., BASSEN, J., ND SURYA GANGULI, J. H., SAHAMI, M., GUIBAS, L. J., AND SOHL-DICKSTEIN, J. Deep knowledge tracing. In *In Advances in neural information processing systems (NeuriPS)* (2015).
[37] RITTER, S., ANDERSON, J. R., KOEDINGER, K. R., AND CORBETT, A. Cognitive tutor: Applied research in mathematics education. In *Psychonomic Bulletin and Review* (2007), vol. 14, pp. 249–255.
[38] RITTER, S., HARRIS, T. K., NIXON, T., DICKISON, D., MURRAY, R. C., AND TOWLE, B. Reducing the knowledge tracing space. In *International Conference on Educational Data Mining (EDM)* (2009).
[39] SCHIJVEN, M., AND JAKIMOWICZ, J. Virtual reality surgical laparoscopic simulators. In *Surgical Endoscopy And Other Interventional Techniques* (2003).
[40] SIU, K.-C., BEST, B. J., KIM, J. W., OLEYNIKOV, D., AND RITTER, F. E. Adaptive virtual reality training to optimize military medical skills acquisition and retention. In *Military Medicine* (2016), vol. 181, pp. 214–220.
[41] STONE, R. T., WATTS, K. P., ZHONG, P., AND WEI, C.-S. Physical and cognitive effects of virtual reality integrated training. In *Human Factors: The Journal of the Human Factors and Ergonomics Society* (2011), vol. 5.
[42] SZAFIR, D., AND MUTLU, B. Artful: Adaptive review technology for flipped learning. In *Conference on Human Factors in Computing Systems (CHI)* (2013).
[43] TANG, R., YANG, X.-D., BATEMAN, S., JORGE, J., AND TANG, A. Physio@home: Exploring visual guidance and feedback techniques for physiotherapy exercises. In *Computer Human Interaction (CHI) Conference on Human Factors in Computing Systems* (2015).
[44] TATTAN-BIRCH, H. Bayes factor.info, 2022.
[45] TSAI, C.-Y., TSAI, I.-L., LAI, C.-J., CHOW, D., WEI, L., CHENG, L.-P., AND CHEN, M. Y. Airracket: Perceptual design of ungrounded, directional force feedback to improve virtual racket sports experiences. In *Computer Human Interaction (CHI) Conference on Human Factors in Computing Systems* (2022).
[46] VANDERMASESEN, M., WEYER, T. D., FEYS, P., LUYTEN, K., AND CONINX, K. Integrating serious games and tangible objects for functional handgrip training: A user study of handly in persons with multiple sclerosis. In *Designing Interactive Systems (DIS)* (2016).
[47] VIDAL, L. T., SEGURA, E. M., AND WAERN, A. Movement correction in instructed fitness training: Design recommendations and opportunities. In *Designing Interactive Systems (DIS)* (2018).
[48] WANG, S., HE, F., AND ANDERSEN, E. A unified framework for knowledge assessment and progression analysis and design. In *Computer Human Interactions Conference on Human Factors in Computing Systems (CHI)* (2017).
[49] WARD, M., GRUPPEN, L., AND REGEHR, G. Measuring self-assessment: Current state of the art. In *Advances in Health Sciences Education* (2002), vol. 2, pp. 63–80.
[50] WOŹNIAK, M. P., DOMINIAK, J., PIEPRZOWSKI, M., AND ET AL. Subtletee: Augmenting posture awareness for beginner golfers. In *Computer Human Interaction (CHI) Conference on Human Factors in Computing Systems* (2020).
[51] XIE, B., LIU, H., ALGHOFAILI, R., ZHANG, Y., JIANG, Y., LOBO, F. D., LI, C., LI, W., HUANG, H., AKDERE, M., MOUSAS, C., AND YU, L.-F. A review on virtual reality skill training applications. In *Frontiers in Virtual Reality* (2021), vol. 2.
[52] YUDELSON, M. V., KOEDINGER, K. R., AND GORDON, G. J. Individualized bayesian knowledge tracing models. In *International Conference on Artificial Intelligence in Education (AIED)* (2013).

APPENDIX

A. Interaction with Experts

We had a 2-hour joint video call with the four experts, each of whom is compensated $50. To facilitate the discussion we shared a Figma document on which the experts can collectively draw or jot down ideas. First, we asked the experts to provide us with the ten most fundamental skills to play Echo Arena. To facilitate brainstorming, we initially asked them to individually jot down relevant skills on the shared document for five minutes. They proposed more than ten skills so we asked them to reach consensus to ten through discussion without our intervention. Next, we inquired the prerequisite relations among the ten skills. To facilitate the discussion, we created a knowledge graph of ten nodes (as shown in Fig. 5) without any directed edges or coloring on the shared document. We first asked which skills are the most fundamental among the ten skills. We place them at the top of the knowledge graph, becoming the *root nodes* with no prerequisite skill. Then, we asked them to place the those skills, or nodes, beneath the root nodes and draw directed edges to indicate prerequisite relations. We iterated this process until all nodes are referenced, forming the knowledge graph as in Fig. 5 without colors.

Next, for each skill, we asked the experts to describe training and evaluation scenarios by drawing top-down views of these scenarios in the shared document. Then, we asked for BKT parameters for each skill considering its training scenario as explained in Sec. IV-A2, starting with the root nodes and traversing down the knowledge graph by depth. Finally, we asked them for a curriculum, i.e. a carefully ordered sequence of skills, to train the control group. We provide the visual explanations of the training/evaluation scenarios for these ten skills in the Supplement.

B. Eligibility Criteria for the Experiment

During recruitment phase, we only participants should regularly play dynamic and interactive VR games like Echo Arena, for at least an hour per month without symptoms of motion sickness. This meant that their physical bodies are already accustomed to VR.

During study, we excluded participants after pre-test. These exclusion criteria set the lower and the upper bound for our study's eligibility criteria. As to the lower bound, participants who cannot score at least 50% in the evaluation tasks during pre-test, regarding the three fundamental skills, i.e thrust, grab/release, and static pass as shown in Fig. 5. As to the upper bound, participants who scores over 50% in the overall pre-test are also excluded because they already know majority of the skills, and would limit the potential range of learning gains.

C. Details of Our Experiment Design

1) Tutorial Session: Because all the participants never played Echo Arena, we asked all participants to watch a short tutorial video that we prepared covering basic controls (e.g. thrusts for navigation, grabbing an object, brake) and then asked them to wear Oculus Quest 2 VR headset and familiarize the controls in a few simple scenarios.

2) Pre / Post Test Sessions: For each skill in the curriculum enumerated by the experts, we gave participants three evaluation exercises representing the skill. To account for potentially high slip rate, our experts suggested that we evaluate each skill repeatedly with variations to accurately measure learning. Three was the maximum number of evaluation exercises we could afford to complete the study within 2 hours. We sampled the three evaluation exercises, i.e. concrete scenarios, for each skill from the corresponding evaluation SCENIC program and generated them in VR headset in sequence (refer to Sec. IV-A1).

3) Training Session: At the beginning of the training session, for both groups, we asked participants to watch another tutorial video we prepared, which provided game intuition on how to solve tasks more easily. Then, both groups were instructed to "master" the 10 skills in 25 minutes of training time with a 10 min break in the middle. We did not provide the definition of mastery, and left it to the learners. This instruction posed a *trade-off* between committing much time per skill for mastery versus training for all 10 skills within the limited time. The self-guided control group needed to *manually* control the learning speed and the curriculum to balance this trade-off. On the other hand, these two are automatically controlled by our algorithm for the experimental group.

By default, the control group trained with a non-adaptive, fixed curriculum that experts provided (Appendix A). However, they could change the curriculum by skipping and returning to skills of their choice. And, they could adjust the learning speed by choosing to repeatedly train for a particular skill.

During training, the training scenarios were generated in sequence in the VR headset. After each scenario, for both groups, we displayed a question in text whether they mastered the skill to solve the training scenario. We asked all the participants to laser tag either 'Yes' or 'No' button on the screen to indicate their self-assessment of skill mastery. Answering this question was mandatory. Otherwise, participants were not allowed to proceed to the next training scenario. This Yes/No answer was only logged and was not used as feedback to the tutoring system in either groups. In addition, only for the control group, we provided an extra 'Skip' button which when pressed, allowed the participant to skip to the next skill in the expert's fixed curriculum. Pressing this skip button was optional. If not pressed, a variant training scenario for the current skill was generated. The control group was also allowed to return to the skill they skipped by verbally requesting to the experiment conductor.

4) Exit Questionaire: We asked the participants in both groups to rate their user experience in Likert 5-Point scale.