Computers & Education: X Reality 2 (2023) 100012

Contents lists available at ScienceDirect

Computers & Education: X Reality

journal homepage: www.journals.elsevier.com/computers-and-education-x-reality

Higher-order thinking skills assessment in 3D virtual learning environments using motifs and expert data

Nuket Nowlan [a], Ali Arya [a,*], Hossain Samar Qorbani [a], Maryam Abdinejad [b]

[a] Carleton University. 1125 Colonel By Drive, Ottawa, Ontario, K1S5B6, Canada
[b] Delft University of Technology, Mekelweg 5, 2628 CD, Delft, Netherlands

ARTICLE INFO

Keywords:
Higher-order thinking skills
Virtual learning environments
Motif
Assessment
Process metric

ABSTRACT

The research reported in this paper addresses the problem of assessing higher-order thinking skills, such as reflective and creative thinking, within the context of virtual learning environments. Assessment of these skills requires process-based observations and evaluation, as the output-based methods have been found to be insufficient. Virtual learning environments offer a wealth of data on the process, which makes them good candidates for process-based evaluation, but the existing assessment methods in these environments have shortcomings, such as reliance on large data sets, inability to offer specific feedback on actions, and the lack of consideration for how actions are integrated into bigger tasks. Demonstrating and confirming the ability of three-dimensional virtual learning environments to work with process metrics for assessment, we propose and evaluate the use of motifs as an assessment tool. Motifs are short and meaningful combination of metrics. Combining time-ordered motifs with a similarity analysis between expert and learner data, our proposed approach can potentially offer feedback on specific actions that the learner takes, as opposed to single output-based feedback. It can do so without the use of large training datasets due to reliance on expert data and similarity analysis. Through a user study, we found out that such a motif-based approach can be effective in the assessment of higher-order thinking skills while addressing the identified shortcomings of previous work. We also address the limited research on similarity-based analysis methods, compare their effectiveness, and show that utilizing different similarity measures for different tasks may be a more effective approach. Our proposed method facilitates and encourages the involvement of instructors and course designers through the definition of motifs and expert problem-solving paths.

1. Introduction

Numerous educational organizations and think tanks have put forth the concept of 21st-*century skills* to refer to the mental skills required in the increasingly competitive 21st-century environment (Abdullah, 1998; Almond et al., 2010). The demands of the 21st century (such as critical thinking, problem-solving, and drawing conclusions) have reinforced the importance of Higher-Order Thinking Skills (HOTS), including critical, logical, reflective, metacognitive, and creative thinking, which are particularly important when dealing with new and uncertain situations (King et al., 1998).

Assessment is essential to any learning process, as it helps identify if the learner is on the path to mastery and what areas need more attention and development (Hill, 2013). However, current output-based assessment approaches are not suitable for assessing HOTS (Fullan & Langworthy, 2013), as they focus on the result (output or outcome) of the learning process. Such focus can be effective for specific and straightforward tasks, but for the assessment of HOTS, observing the learning process itself is required for deeper insight into learners' abilities. Observing the process of learning means noticing the type, order, quantity, and quality of interactions between learners and the learning environment throughout a learning task. The metrics that represent such observation are generally referred to as *process metrics* (Bennett, 2003). Assessments based on process metrics are considered richer than those that rely on output-based data (*output metrics*) (Griffin & Care, 2014). Using process metrics provides information on the learning methods and strategies that learners use throughout tasks and thus better captures HOTS development (Greiff et al., 2012).

Unfortunately, observing each individual learner performing a task is extremely time-consuming and requires significant resources. To address

* Corresponding author.
E-mail addresses: NuketNowlan@cmail.carleton.ca (N. Nowlan), arya@carleton.ca (A. Arya), HossainSamarQorbani@cmail.carleton.ca (H.S. Qorbani), M.Abdinejad@tudelft.nl (M. Abdinejad).

https://doi.org/10.1016/j.cexr.2023.100012
Received 3 November 2022; Received in revised form 22 February 2023; Accepted 22 February 2023

this problem, researchers have turned to *computer-based assessment* as a way of capturing rich information about the learning process (Burelson et al., 2014). Computer-based platforms designed to support learning are commonly called Computer-Based Learning Environments (CBLEs) or Virtual Learning Environments (VLEs) (Duncan, Miller, & Jiang, 2012; Scavarelli, Arya, & Teather, 2021). The most common VLEs are Learning Management Systems (LMS), frequently used in educational institutions for communication and assessment (Kasim & Khalid, 2016). They offer many functionalities, including presenting learning content, discussions, and evaluation. These VLEs are generally text-based with multimedia elements. They usually have a limited ability to collect process metrics and rely on end (submitted) results. Still, they can track some process metrics, such as files opened or content elements read by the learner. Assessment tools in these environments can include various graded items, rubrics, and automated formulas (Hussain et al., 2018; Kasim & Khalid, 2016; Młynarska et al., 2016).

A particular type of VLE with the potential to facilitate HOTS development and assessment is the 3DVLE, i.e., three-dimensional graphical multi-user virtual environments specifically designed for educational purposes (Dede, 2007; Scavarelli et al., 2021). 3DVLEs allow avatar-based interaction and navigation of virtual spaces that can simulate real ones or visualize novel spaces (Arya et al., 2012). They have been increasingly attracting the attention of scholars interested in skills development due to their ability to facilitate simulated experiential learning (Schmidt & Stewart, 2009; Alqahtani et al., 2017; Scavarelli et al., 2021; Elme et al., 2022). They also have the potential to observe learners throughout the process and track and assess a significantly wider range of user activities (Loh & Sheng, 2015a). In this paper, we use the term 3D Virtual Environment as synonymous with Virtual Reality (VR), which can be on a desktop, mobile device, or Head-Mounted Display (HMD). Some literature uses the term VR only for HMD-based experiences.

As described in the next section, HOTS assessment in VLEs is done through score-based and series-based methods, which assess the learner by assigning scores to elements of performance or to the sequence of those elements, respectively. As we show in our literature review in the next section, score-based methods lack attention to the importance of when an activity happens and the order of the activities (Snow et al., 2015). Series-based methods address that shortcoming but are usually looking at the full set of activities and do not have a localized assessment (Reilly & Dede, 2019). As a result, providing meaningful feedback on the elements of learning activities is not easily possible in the existing assessment approaches. Also, there is limited research on what type of analysis on performance metrics should be done. On the other hand, assessment methods may require large amounts of data to establish patterns, and so they frequently use machine learning algorithms which again negatively affect the possibility of offering meaningful feedback on specific elements of the learner's activities (Conati et al., 2018; Loh & Sheng, 2015a).

The concept of motif as a small yet meaningful set of tasks within the learning process has been suggested by some researchers (Gibson & de Freitas, 2016). Motifs have the potential to offer more localized assessment and feedback, but their effectiveness has not been properly investigated. We propose that motifs, together with the use of similarity analysis with expert data (Floryan et al., 2015), can address the existing shortcomings in score-based and series-based assessments, i.e., the lack of localized feedback, the use of large amounts of data, and limited studies on analysis methods. This paper reports on our investigation of the effectiveness of this proposed approach and similarity analysis measures that are appropriate for it. We defined motifs for various HOTS within an educational context (Chemistry lab) and compared the motifs based on students' performance in a custom-designed 3DVLE to that of experts. Four different similarity measures were used for comparison and resulted in an "auto" assessment for students. We then did a correlation analysis between these assessments to see which similarity measures aligned with the instructor's observation-based assessment. A high correlation would suggest that the automated assessment was based on an appropriate method.

Our findings show that motif-based similarity analysis can be successful in providing meaningful assessment and feedback with a limited need for data. Our comparison of similarity analysis methods shows that using a unique method may not be the best approach, and a combination of similarity measures may be an optimal choice. We were motivated by the need to make HOTS assessment in 3DVLEs more suitable for instructors through methods that correspond to the instructors' knowledge of what processes need to be followed for learning tasks, and offer insight into specific learner actions. We contribute to the field of 3DVLE and HOTS assessment by providing evidence in support of 3DVLE-based assessment, particularly.

- Proposing and evaluating the use of motifs and expert data to allow specific feedback with a limited data set. We showed that the combination of these two has the potential to assess HOTS.
- Investigating the suitability of different similarity measures for expert/learner comparison. We found preliminary evidence that using different measures for different tasks may be more appropriate.

It is not possible to investigate a large set of HOTS in one study. As such, our study is focused on problem-solving, drawing conclusions, and crisis management as part of lab safety training. We use this example to illustrate the more general notion of HOTS assessment. This particular case was chosen by our partner instructor as a typical situation that helps strengthen and evaluate the students' HOTS. Since the skills required are typical of many HOTS-related scenarios (Robinson & Schraw, 2011; Shute et al., 2015), we expect the findings to show potential and offer initial insights into the suitability of our proposed method. Further investigations are admittedly required to expand these initial insights and see how they can be generalized to other cases.

2. Related works

2.1. Higher order thinking skills

Robinson and Schraw (2011) defined Higher-Order Thinking Skills (HOTS) as those skills that "enhance the construction of deeper, conceptually-driven understanding" (p. 23), including the skills of reasoning, argumentation, problem-solving and critical thinking, and metacognition. Traditionally, assessment in education has focused on knowledge (Leighton, 2011). Therefore, it might be expected that traditional assessment approaches are not adequate to assess HOTS, which is more process-oriented. The question-and-answer approach, mainly developed for knowledge assessment, cannot assess processes that involve using complex competencies (Shute et al., 2015). To assess and understand the areas for improvement for each student, more sophisticated assessment tools are needed to follow students' decision-making, thinking, and investigation processes (Code & Zap, 2013; Shute & Kim, 2014).

Robinson and Schraw (2011) argue that HOTS can be assessed by observing the thinker while they are engaged in an activity (process), such as inquiring or identifying questions, assumptions, or issues to investigate. However, observing a learner when they perform activities can be difficult due to location and time constraints, or the possibility of influencing the process. Computer-based assessment, with its capacity to capture rich information about students' learning process, may help educators with this task. Emerging technologies such as VLE/3DVLE enable such practices by offering simulated space and digital tracking capability (Borgman et al., 2008; Dede, 2009; Warburton, 2009).

2.2. HOTS in virtual learning environments

The history of VLEs dates back as early as the 1960s when computer-based courses were being developed, yet it was only computer advances

in the 1980s and 1990s that allowed the creation of learning systems that are recognizable today as widespread Internet-based educational media (Duncan et al., 2012). VLEs provide educational content, allow communication, and can facilitate skills development. They are available in many different formats: single or multi-user, gamified or not gamified, and 3D immersive or not immersive.

Closely related to VLEs, there is the concept of Smart Learning Environment (SLE). According to Koper (2014), SLEs are "physical environments that are enriched with digital, context-aware, and adaptive devices to promote better and faster learning." SLEs are the result of the advances in information technologies, particularly online services, mobile devices, and the Internet of things, that allow the increasing use of smart devices in all aspects of the learning process. SLEs allow the integration of VLEs into the physical environments and can increase flexibility and personalization (Chen et al., 2021). Research on SLEs can be categorized into technology-related, domain-related, and learning process-related, where feedback and assessment form a potential area of future research due to limited existing work (Muller et al., 2019; Chen et al., 2021). In this paper, we focus on 3DVLEs due to their simulation and tracking abilities, as mentioned earlier.

Kelman (1989) identified 3DVLEs as a potential environment to foster HOTS development. In a study by Hopson et al. (2001), students' self-reports indicated that the 3DVLE resulted in increased motivation, creative tendencies, inclination towards exploring the unknown, perseverance, and taking individual initiatives. Dede and his colleagues developed River City (Dede & Ketelhut, 2003; Ketelhut et al., 2010) with the claim that HOTS are best developed when learners construct knowledge rather than passively ingest information, and information-gathering tools and evaluation systems are used to measure complex higher-order skills, rather than the simple recollection of facts.

As noted in a highly cited article by,Roschelle et al. (2000) "Although active constructivist learning can be integrated into the classroom with or without computers, the characteristics of computer-based technologies make them a particularly useful tool for this type of learning" (p. 79). Dede (2007) argued that 3DVLEs are learning environments well-suited for the promotion and assessment of learning with the following strategies: active, experiential, and situational learning. However, research on how VR affordances can be used to implement learning and educational strategies is still limited (Dede, 2007; Scavarelli et al., 2021).

2.3. HOTS assessment methods in virtual learning environments

As Queiroz et al. (2019) identified, the existing assessment methods in VLEs mostly focus on more tangible skills such as biology, computer science, and medicine. Limited work has been done on the assessment of higher-order thinking skills, analysis of learning patterns, and investigating methods that allow the use of AI. At the same time, researchers (Spector & Ma, 2019) have warned about over-emphasis on the use of AI and the need to rely on human intelligence, and the development and assessment of higher-order thinking skills in learners through simulations and games (Ketelhut et al., 2010; Van Voorhis & Paris, 2019).

While some researchers (Kuang et al., 2021) have tried to assess HOTS through output metrics (such as the difference between pre-intervention and post-intervention test scores), it is generally accepted that HOTS require a more process-oriented assessment approach (Griffin & Care, 2014; Van Voorhis & Paris, 2019). *Stealth Assessment* (SA; Shute, 2011) is the generic term for the methods used on computer-based learning platforms, including VLEs, to assess learners' progress through the collection of interaction logs and analytics (process metrics)—without interrupting learners' flow. SA approaches in existing literature can be organized into the following two categories: Score-based and series-based (Shute, 2011).

2.3.1. Score-based stealth assessment

Score-based stealth assessment methods are used when learners' interactions with the VLE are scored based on a logic compatible with the learning activity, usually with the help of an assessment script. Scoring can be performed by giving equal or different weights to each interaction, or it can be done via scripting based on a defined algorithm.

Veenman et al. (2014) proposed that log files of students' actions in computer-based learning environments can reliably track students' learning process while they engage in learning tasks and support them when needed, thus helping them to improve their meta-learning skills. Correlation results showed a stronger relationship between metacognitive skills traced through user action captured in log files and Groninger Intelligence Test results (Veenman et al., 2014).

Arroyo et al. (2014) conducted a similar study with Wayang Outpost, a computer-based tutoring system that provides pedagogical meta-learning skills assessment and support for students' mathematical problem-solving skills. Arroyo et al.'s study showed that not only can metacognitive skills be traced through user interactions, but also meaningful tutoring support can be given to foster metacognitive skills.

Azarnoush et al. (2015) investigated simulation metrics to identify expert and resident surgeons in a virtual reality simulator, NeuroTouch, that simulates neurosurgical procedures, including brain tumor resection.

Although scoring different types of interactions—e.g., simply adding up each categorical interaction and performing a manual assessment in relation to each interaction category and frequency with the learner's success—offers encouraging insights on students' progress, general scoring for each interaction can be misleading depending on the position where this interaction occurred. On the other hand, the increase in data has motivated researchers to investigate the use of Machine Learning (ML; Jordan & Mitchell, 2015) methods to study students' learning task interactions (Sabourin et al., 2013; Shute & Kim, 2014). However, they require a model to be prepared in advance, a lengthy process in which all actions' probabilities and meanings must be identified. As such, it is not easy to change and customize the flow of action. Also, it is a challenge to assign a practical, real-life meaning to many machine learning (especially deep learning) results and interpret their probabilities as actionable insights that can improve performance (Conati et al., 2018; Loh & Sheng, 2015a).

To make the assessment results more explainable, researchers have considered data patterns instead of single metrics (Baker & Clarke-Midura, 2013; Gibson & de Freitas, 2016). Baker and Clarke-Midura (2013) collected log files which were further distilled for analysis, producing a set of 48 semantically meaningful features. Gibson and de Freitas (2016) used the same log files showing that clustering was ineffective until the study's subject domain experts identified a two or three-element chain of actions, which the researchers called a *motif*. For example, a data element named 'opened door' by itself was relatively meaningless compared to knowing that it was a particular door, opened after another significant event, such as talking to a scientist. Thus, patterns of action were transformed into motifs, which then became the transformed units of analysis. The method was still score-based as it did not take into account the order of elements.

To deal with the large amount of data required to train the machine learning algorithms, the use of expert knowledge has been suggested. Floryan et al. (2015) used (1) basic features and (2) expert knowledge-based features to train the machine learning algorithm in a biology course. Their work suggests that the use of expert knowledge can be helpful for HOTS assessment, as well, to reduce the large amount of data needed for machine learning models.

To summarize, while score-based methods can offer overall assessments, they are limited in providing localized feedback and evaluation, as the order and relation of individual activities within a task are not considered properly. They may also require large amounts of data to train machine learning models. The use of motifs and expert data have been suggested by researchers, but the existing literature has not investigated the application of these concepts within a series of actions where the order and relation of activities matter. Series-based methods have been suggested to address some of these shortcomings.

2.3.2. Series-based stealth assessment

The alternative approach to score-based stealth assessment is flow or series-based assessment. This method generally uses learners' full activity series for all types of interactions or selected interactions as input rather than a single interaction or metrics derived from interactions. Unlike the motif approach above, where a series of interactions are divided into small meaningful, self-contained task components, this approach tries to make sense of full series without considering different categorical tasks within the learning session.

A notable study using this approach was conducted by Snow et al. (2015), who aimed to identify stability in the action series for students. They collected digital activity traces of college students performing learning tasks in the Interactive Strategy Training for Active Reading and Thinking (iSTART), a 2D game-based intelligent tutoring system designed to improve students' reading comprehension. Snow et al. (2015) concluded that learners' interaction log provides valuable information on their progress and future performance. This method is noteworthy, as it does not require previous data collection to train a model or definition of a rubric to assess learners' performance. However, it has the shortcoming of offering actionable development points to learners as an overall series-based assessment, so it is not clear where the actual failing points are that learners should focus on to improve their performance.

Another notable study was reported by Loh and Sheng (2014). Unlike Snow et al., who analyzed the stability on learners' interaction series, Loh and Sheng compared learners' series with an expert series to identify the difference by employing string similarity index analysis. The term *similarity* covers a wide range of scores and measures that assess differences among various kinds of data. Similarity metrics were originally used to statistically define (dis)similarities between two strings in a database (Winkler, 1999). Loh and Sheng (2014, 2015a, 2015b) suggested using multiple similarity indices for different experts and using the maximum similarity index (MSI) to identify specific players' expertise levels. This approach offers a single metric and easy-to-measure practicality that is not offered by score-based approaches, but it only provides an overall performance assessment rather than a performance component. Sawyer et al. (2018) also suggested that comparing students' problem-solving paths to an expert's problem-solving path might provide strong insight into students' problem-solving skills.

Building on Sawyer et al. (2018), Reilly and Dede (2019) conducted a similar analysis on ecoMUVE, an inquiry-based 3DVLE curriculum. In this study, students' trajectories were clustered by time series instead of comparing them with experts. Students within the golden path cluster (belonging to an expert) were noted to be those with the highest knowledge gains. Reilly and Dede suggested further studies on performing clustering on the slopes and the distance to see if the patterns emerge in play styles that meaningfully correlate with learning gains or effective dimensions.

More studies are needed to help instructors in offering more specific and targeted assessments identifying weaknesses and strengths in students. This might be achieved by performing series-based assessments on smaller tasks or elements of activity (for example, a motif). Expert actions can then be used to help manage the data size. The review of literature on existing score-based and series-based approaches shows that there are research gaps and open questions on the effectiveness of series-based assessment using smaller groups of tasks, the use of expert data and similarity measures for such small groups to avoid the need for large data sets, and the suitability of different similarity measures. These gaps are the basis of our research questions and the motivation for our proposed solution and study.

3. Study design

3.1. Overview

Our literature review revealed that series-based assessment has the ability to incorporate the sequence of learning activities, but performing this assessment on large sequences may not provide proper feedback. We also noticed that similarity-based analysis with expert data could reduce the need for large amounts of data. In our previous work (Nowlan et al., 2018), we confirmed the ability of motifs to define smaller sequences for score-based assessment. In this study, series-based stealth assessment is defined as using full or partial interaction series of a learner captured during a learning session to analyze and gain insight into performance. The time series clustering approach has been used for profiling and future prediction (Edwards & Cavalli-Sforza, 1965, pp. 362–375, Peffer, Quigley, & Mostowfi, 2019). The downside of this method is that it requires a large amount of data before it can be used. Series similarity measure-based analysis (Loh & Sheng, 2014, 2015a, 2015b; Sawyer et al., 2018) is another alternative. This method is quite practical compared to machine model creation methods. The drawbacks of this method include its application to the whole series, which lacks flexibility and partial feedback, and its use of similar analysis for all activities when different activities might be better assessed using different similarity indices.

In this study, we investigate the use of motifs in series-based assessments to provide the ability to have localized feedback, and the use of expert actions (provided or approved by the instructor) and similarity measures to reduce data requirements and offer flexibility in assessment methods through different similarity indices for each motif. We aimed to answer the following research questions.

RQ1: How can a small yet meaningful series of process metrics (motifs) be used for series-based HOTS assessment?
RQ2: How can similarity analysis between student and expert motifs be used to assess HOTS?
RQ3: Which similarity indices are more effective in assessing HOTS?

The research and our questions are motivated by the need to find assessment methods that allow localized feedback and evaluation on specific (smaller) activities without the need to use a large amount of training data. Motifs and similarity with expert actions have shown potential in other cases, and as such, we proposed an assessment method based on these concepts. The proposed method allows a more direct involvement of instructors and course designers in the VE assessment process, as they can flexibly define motifs and expert actions.

Our study was performed in the context of a chemistry lab in a 3DVLE (both desktop and head-mounted display). Providing safety training before letting students into a physical chemistry lab is a mandatory step that all educational organizations need to facilitate. Traditionally, this step is done by text or video-based materials along with a question-and-answer assessment of readiness. Before entering the lab, it is important for students to learn about the dress code (e.g., safety goggles), the components of the lab (e.g., eyewash and shower and how to use them), and the correct process in case of an emergency. Due to safety concerns, creating situations such as a fire or chemical spill where students can apply their emergency reaction skills is impractical or impossible except in virtual reality.

This research was approved by our institutional Research Ethics Board, as detailed at the end of the paper.

3.2. Participants

The research team invited university students to participate and collect research data in the Winter 2021 semester. Invitations were done through dedicated social media groups the university has for research participants and also colleagues who could forward the invitation. Any university student could participate (no VR experience needed) but we particularly encouraged those from science programs and also emphasized the desired gender and ethnic diversity. Unfortunately, due to COVID-19 restrictions, we could not invite our participants to our lab to participate, which would have allowed us to capture screen video.

We also had difficulty finding participants due to COVID-19

pandemic restrictions. All educational activity was required to be performed online, and we found that students were less eager to participate in online/virtual research than before the pandemic.

Ultimately, we recruited 36 university participants, 20 male and 16 female. The average age was 25, with a standard deviation of 6.45. All participants were university students in Chemistry or another science/engineering program. On average, they had taken 5.10 chemistry courses with a standard deviation of 3.70. Almost all participants (94%) had completed previous traditional lab training; yet, according to the partner instructor, even students who passed training tended to have problems in the lab. 66% of participants had prior experience in immersive VR with a variety of games.

Due to the technical difficulty in creating motifs on the desktop app, we could only use data from the Head Mounted Display (HMD) participants, of which there were 18 in total, 10 male and 8 female students. The study participants' age range is between 20 and 52, with a median of 22.50.

3.3. Materials

Our 3DVLE prototype was built for a chemistry lab using Unity 3D, a popular game engine to build 2D, 3D, and VR games and experiences accessible on desktop, mobile, and HMDs. The virtual chemistry lab due to multiple reasons: (1) it is a typical Science, Technology, Engineering, and Math (STEM) experiential learning environment, (2) we had an instructor who was willing to partner and collaborate, (3) according to the instructor, the activities were suitable to strengthen and evaluate commonly-needed HOTS in students. The environment had three areas for general VR familiarity, chemistry lab training, and actual testing purposes. Typical labs have only one area. We took advantage of virtual facilities to have a more training-friendly area to show common lab elements (lab training). Also, since our target audience had limited or no experience using VR, we included a basic training area to help them gain experience using touch controllers. This was an important design consideration to help users acclimatize themselves and avoid potential motion sickness for some users.

Area one: This area was designed to provide basic training activities such as picking up simple objects (cubes, spheres), which enabled participants to learn how to use touch controllers (Fig. 1). Since a VR experience can overwhelm first-time users, participants were guided to interact with simple objects, travel or teleport (locomotion), pick up objects, and use help tips. This level was built in based on the learning principle of scaffolding, i.e., helping participants build the skills and knowledge necessary to navigate and interact with objects, from simple moving in the virtual environment to using chemistry equipment.

Areas two and three: While area one was for generic VR training, areas two and three offered specific chemistry lab experience. Divided into two sections separated by a wall and a door (Figs. 2 and 3), these two areas were the virtual chemistry lab. Area two was for more advanced interactions and safety training, including personal protective equipment (PPE) and safety questions. Area three was the actual lab with the simulated chemistry experiment, equipment, and scientific experiment stations known as fume hoods.

Instead of using existing 3D platforms, a new standalone app was created that could capture all student's interactions within the VR space, record the order and duration of each, and email the data to the research team at the end of each experience. All data cleaning, series creation, and similarity index calculations were done by using scripts created for this study in Python. Visual Basic was used to run the scripts. Correlation assessment was performed by using Microsoft Excel Data Analysis functions.

Through pilot testing and received feedback, the research team decided that although it could be beneficial to create a high-fidelity environment, in terms of efficiency and considering the level of skills to be acquired by the learners, such visual fidelity was not critical, as suggested by Lefor et al. (2020). Our study included expert data provided by the instructor as they saw fit for comparison. The criteria for selecting the expert data were based on the instructor's past experience.

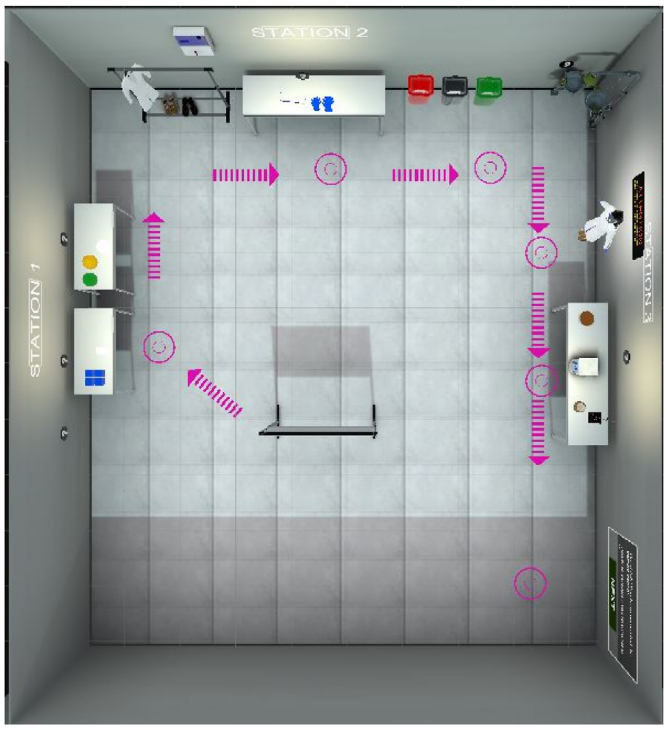

Fig. 1. Area one.

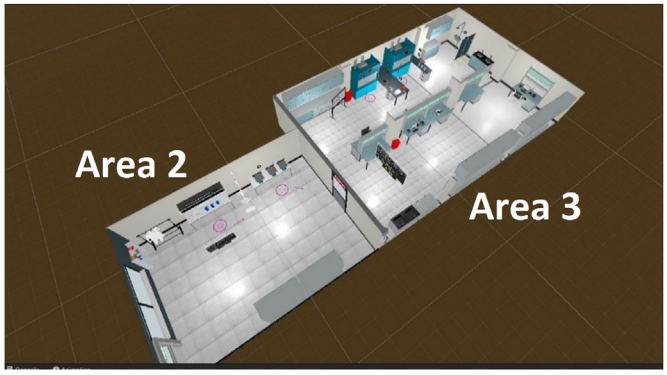

Fig. 2. Areas two & three.

3.4. Procedure

Due to COVID restrictions, the studies were done remotely. All required pieces of equipment were dropped of and picked up with proper cleaning. The participants were instructed to go through all three environments and perform the tasks. Overall, the following procedure was employed for collecting and analyzing data.

1. A three-dimensional virtual chemistry lab was designed and created, where:
   a. Learners could explore and discover the chemistry lab's dress, equipment, and components used for emergency situations.
   b. Learners could perform a virtual chemistry experiment with guidance.
   c. Learners could be faced with an emergency (fire) that they must handle with the correct protocol without any guidance.

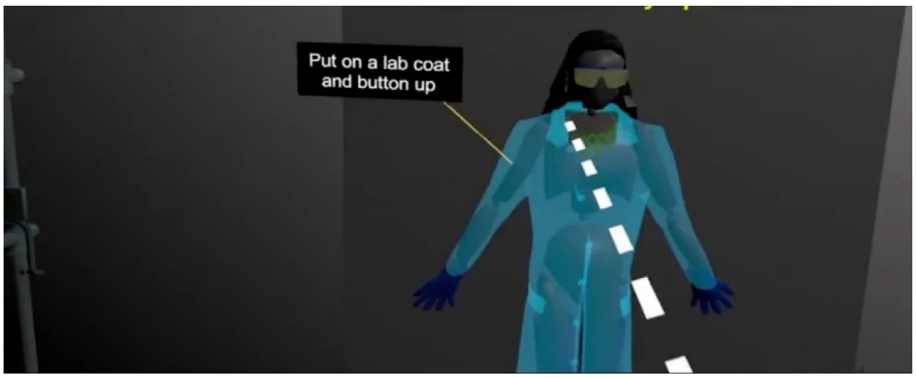

Fig. 3. Screenshot from area two.

d. Learners' digital footprints would be captured with timestamps when they:
   i. Interact with an object in the environment.
   ii. Read information.
   iii. Grab a tube, etc.
2. Participants were recruited to follow the safety training in the VLE, and then perform an experiment where they were faced with an emergency.
3. Activities that participants were expected to perform were completed by an expert.
4. All participants' and expert's full interaction series were logged.
5. Both participants' and expert's activity series were split into three skill components and a series created for each.
6. A series similarity analysis was applied to assess performance on students' activity path as compared to the expert path.
7. A correlation assessment was performed between students' similarity-based performance assessment versus the manual expert assessment based on a log file following students' digital footprint.

Participants could practice freely in areas one and two, although we told them to try everything. They were required to perform all assigned tasks for area three, as listed in Table 1.

A screenshot of a student performing an experiment is shown in Fig. 4. The system log in Fig. 5 shows a sample of the digital activity trace that was collected during students' learning sessions.

3.4.1. *Motif identification for HOTS assessment*

Loh and Sheng (2015a, 2015b) suggested that an expert path compared to novices' path similarity measure can be used for skill assessment. This approach is more flexible than a machine model training approach, a method where large-size training data is collected from many players and labeled to train the model before it can be used. In this research, we hypothesized that the similarity measure used for this assessment method should be chosen in line with the learning activity. As each similarity measure formula works differently, tests are required to determine the measure that works best for different learning activities with different learning objectives.

Gibson and de Freitas (2016) put forward the idea that learning motifs, a small group of activities meaningful as a group, to facilitate granular analysis, large patterns of action were transformed into motifs, which then became the transformed units of analysis. In our previous study (Nowlan et al., 2018), we defined motifs as an overlapping combination of activities that build four separate skills. These skills existed in 6 chambers that students visited. For this study, the instructor was looking for three skills that could be identified within three separate tasks. There was no longer any overlap (shared activities), and the pattern and order of activities for each HOTS (and the related motif) were defined by the instructor. We hypothesized that by applying different similarity index measurements on all tasks and correlating the results with the instructor's assessment, we would get insight into the suitability of different similarity indexes and the ones that fit better for different learning tasks.

Table 1
Steps and tasks to be facilitated in Area 3.

Step	Detailed Tasks
Guided Experiment	• Pull up the hood door/glass up (1/3rd) • Grab the stand (from the bench), put it inside the fume hood • Grab/put the hot plate beside the stand • Plugin the hotplate • Grab/put the oil bath on the hotplate • Turn on the hotplate and increase the temperature (140 C) • Add the materials/powders (They'll get purple solution for this experiment) • Grab and put the condenser on top of the flask • Clamp the flask - (Step where emergency will be inserted) • Turn on the water which is connected to the condenser • Turn on the hot plate and magnetic stir bar • Turn on the water tab which is attached to the condenser • Turn on the hot plate • Wait 2 h (simulated) • STOP the reaction: o Press the off/on button to turn off the heater o Turn off the device magnet knob to stop stir bar o Don't turn off the water connected to the condenser until it's completely cool down
Example of an emergency in VR lab and responsive actions	• Put a round bottom flask inside the oil bath WITHOUT clamping it • If the flask falls inside the oil bath and causes fire: o Unplug the hotplate first o Pull down the hood door/glass and let the oil-bath completely cool down o Pull up the hood door again o Remove the hotplate and pieces of broken glass from the fume hood o Place the hotplate on the bench o Dump the pieces of broken glass into the red bin o Clean up the fume hood using napkins o Use acetone to clean up the oily fume hood

The resulting motifs were selected based on three separate sections of the experience:

Section 1: The purpose of this section was for students to explore the pre-lab section and practice holding/using 3D virtual objects and donning the lab gear/dress. Students then entered the VR lab, explored the lab equipment's components, and practiced using the components. The main HOTS in use in this section was Information Collection.

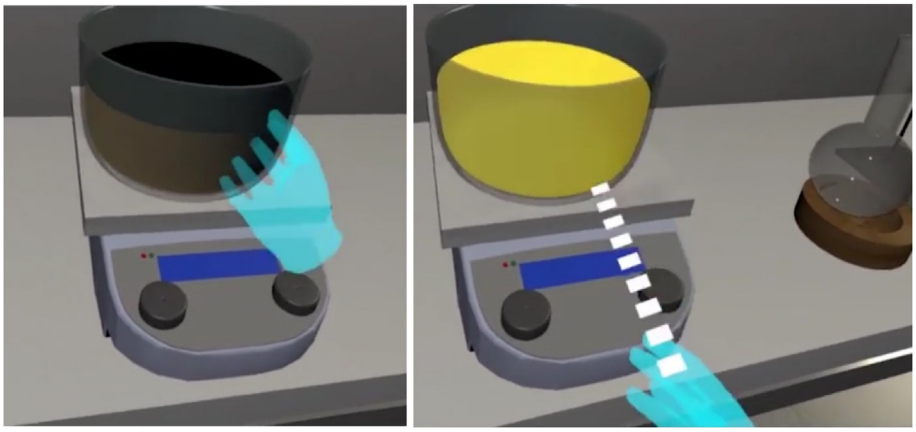

Fig. 4. Student performing an experiment in the three-dimensional virtual chemistry lab.

```
Please Find Below the Logs of the player 5, There is also a log file attached to this email.
No Name No Surname No Age Male
TIMESTAMPS,NAME,USETYPE,USED_BY,LENGTH
00:03,NextButtonATWelcome,Hover,UI Pointer,0,38
00:04,NextButtonATWelcome,Select,UI Pointer,0,00
00:05,NextButtonATWelcome (1),Select,UI Pointer,0,00
00:09,Cube (5),Select,Right Hand,1,51
00:19,gogglefinal4 (3),Select,Right Hand,2,48
00:22,HelpBtnGoggles,Hover,Right Ray Interactor,1,32
00:23,Glove (1),Hover,Right Hand,0,55
00:23,HelpBtnGoggles,Hover,Right Ray Interactor,0,66
00:24,Glove (2),Hover,Right Hand,0,66
00:24,Glove (1),Hover,Right Hand,0,98
00:25,HelpBtnGoggles,Hover,Right Ray Interactor,2,00
00:29,Waste-Bin_Green (8),Hover,Right Ray Interactor,0,94
00:30,Waste-Bin_black (7),Hover,Right Ray Interactor,1,03
00:31,Waste-Bin_Red (6),Hover,Right Ray Interactor,0,80
00:34,Oil,Hover,Right Ray Interactor,0,81
00:36,Oil,Select,Right Hand,0,77
00:39,flask_5,Select,Left Hand,0,82
00:40,flask_5,Select,Right Hand,0,90
00:41,flask_5,Select,Left Hand,0,87
00:41,tripod,Hover,Right Ray Interactor,0,56
00:41,tripod,Hover,Right Ray Interactor,0,56
00:42,flask_5,Hover,Right Ray Interactor,0,96
00:42,HelpBtnLabEquip,Hover,Right Ray Interactor,1,30
00:45,NextButton1,Hover,UI Pointer,0,43
00:46,NextButton1,Hover,UI Pointer,0,40
00:46,NextButton1,Hover,UI Pointer,2,06
00:50,NextButton1,Hover,UI Pointer,0,57
```

Fig. 5. Sample of the data captured in the application audit file.

Section 2: The purpose of this section was: for students to (i) understand the purpose of each item of lab gear/dress and be able to select the correct one with the help of inserted help messages, and (ii) understand and be able to use the lab equipment and perform an experiment with guidance from the inserted messages. The main HOTS in use in this section was Critical Thinking.

Section 3: The purpose of this section was to observe students' learning and understanding of the lab equipment and emergency response process by creating a situation where they could demonstrate these without any help. A virtual emergency was created, where students needed to demonstrate their understanding of the process to be followed without any guidance. The main HOTS in use in this section was Drawing Conclusions.

As a result, three main motifs identified in Study 4's learning session were.

- o Information Collection
- o Critical Thinking
- o Drawing Conclusions

Data points (basic metrics) were collected on each of these motifs for each student, 135 for information collection, 104 for critical thinking, and 70 for drawing conclusions. Examples of these data points are shown below.

Time	Object	Action	User Method	Duration
04:31	Glove	Select	Right Hand	4.20
06:11	manniqTorsoOnly	Hover	Right Ray Interactor	5.54
08:00	flask_5	Select	Right Hand	0.74
10:19	LabCoattorso	Select	Right Hand	3.73
16:29	HoodDoor	Select	Right Hand	1.71
30:32	HelpFH1-1	Hover	Right Ray Interactor	2.38
47:27	broom2	Select	Right Hand	10.94
49:02	Prop_BeakerAcid	Select	Right Hand	15.79
49:39	WaterTrigger	Hover	Right Hand	0.88

3.4.1.1. Identifying activity series gram number. Action series are orderly data points of activity performed over a timeline. Raw data coming from the platform log file capturing these actions was organized in a way that made sense for the analysis. Depending on the objective of the learning task, series data points were created differently. If the before and after actions were deemed imported as well as the current one, each data point was created accordingly. After this decision, an activity series was created from the action path by using each entry as an element of the generated time series (uni-gram series) or by combining two or more entries by creating multiple n-gram series. Depending on the n, series were created as unigram (n = 1), bigram (n = 2), trigram (n = 3), quadgram (n = 4).

Below is an example of an action series where the learner performed the following activity:

A, B, C, A.

The following different gram series can be created to apply similarity analysis:

Unigram series would be: **A, B, C, A**.

Bigram series would be: **A/B, B/C, C/A**.

In the educational context, we believed that having the students perform a controlled action was important, i.e., they were not randomly interacting without an overall strategy. To understand the implication of the importance of the order, we created and used both uni-gram and bi-gram series in all our comparisons. Each activity in our research was captured as an object and an action performed on it, such as LabCoattorso-Hover. Below are examples of the unigram and bigram series used for this research:

Unigram:

LabCoattorso-Hover, sliper-Hover, Q2Cube-Hover, Q1Cube-Select, Q2Cube-Select.

Bigram:

LabCoattorso-Hover/sliper-Hover, sliper-Hover/Q2Cube-Hover,

Q2Cube-Hover/Q1Cube-Select, Q1Cube-Select/Q2Cube-Select.

3.4.1.2. Recording expert/s path. A similarity analysis is performed by methodically comparing two series of data points based on the chosen index's formula. In this study, we wanted to find out how similar learners' interaction paths compared to an expert's path while demonstrating a focal HOTS for each different area.

During testing, our expert invited a competent (high-performing) student to follow the process and do self-recording. We then created an activity path for each of these sections and asked our expert to control the series to make sure it was what the instructor expected from a high-performing student.

Based on the high-performance path, we created expert unigram and expert bigram for the three HOTS motifs for comparison with students' series.

3.4.1.3. Similarity index selection. Maximum Similarity Index (MSI) is a term proposed by Loh and Sheng (2014), as the similarity index that gives the best match, to study the performance of a player in games. In experimenting with different similarity indexes as a performance measure, Loh and Sheng (2015a, 2015b) used multiple similarity indexes to profile players' playing styles. They concluded that a combination of multiple index-based similarity measurements provides the best classification in terms of profiling players' playing styles.

In this research, we also used multiple similarity indexes to assess HOTS by comparing to an expert. Our goal was to find out which one would be the Maximum Similarity Index (MSI) and if MSI would be the same for all HOTS, or different HOTS could use different indexes. Considering the many different similarity calculation methods and their advantages (Loh & Sheng, 2014; Winkler, 1999), we decided to use the following similarity indexes in our research.

i. Jaccard Index:

The Jaccard index formula calculates two series' similarity based on common elements between the two and divides that by a unified set number (Jaccard, 1912). As it has been identified as the best index for assessing similarity (Loh & Sheng, 2015a, 2015b), we decided to include this index in Study 4's assessment. It should be noted that the Jaccard index does not take repeated steps into its calculation. So, if the same trigger is used twice in the expert path (assuming it needs to be used twice to perform the activity properly), this would not be included in the similarity calculation by using the Jaccard index unless the time series was created in a way that each time a similar action performed, it is prefixed with a different number. Our time series creation script was not implemented that way.

$$Jac(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

ii. Modified Jaccard Index:

As our main objective was to find how similar students' activity path was to an expert's activity path, we decided to also calculate a modified Jaccard index with the following formula to focus on the number of common elements between students and expert within the expert path over an expert path, rather than the combined set:

$$Jac(A,B) = \frac{|A \cap B|}{|A|(A\ is\ expert\ series)}$$

iii. Cosine Similarity Index:

The cosine similarity index takes the series into vector space and then calculates the similarity between them; as such, the number of times the same element is repeated in the series makes a difference, where not only the elements should be identical but also the repetition times for series to be more similar. We decided to perform a cosine similarity-based calculation as well.

iv. Levenshtein Similarity Index:

Levenshtein Distance (Levenshtein, 1966) is also known as the edit distance metric. As per the edit distance similarity approach, the smallest number of operations, or edits, required (e.g., insertions, deletions, or substitutions) while transforming one string to another one (e.g., words) is used to quantify the two strings' similarity. The number of operations needed for transforming and series similarity are inversely proportional; the fewer operations required to transform one series to another means the more similar the series are.

There are different ways of applying an edit distance algorithm (performing edits) and calculating similarity depending on the nature of the comparison, such as (i) applying equal cost to each operation, (ii) assigning a higher cost to some of the operations, (iii) allowing transposition of two adjacent elements, (iv) banning substitution from the operation list, and (v) applying only transposition in adjacent nodes. We decided to use the basic Levinshtein algorithm, where operations of insertion, deletion, and substitution were allowed with equal weight. Levenshtein distance calculation starts by initially identifying the path most similar to both series, and then applying the necessary operations to turn one series into another through edit operations. Due to this approach, the Levenshtein similarity index is highly sensitive to the order of the series.

3.4.2. Data analysis method

There were three important steps in applying a time series similarity analysis.

1 Filtering the series of action logs to prepare for the analysis
2 Deciding on the (dis)similarity index to be used to compare with the expert series
3 Creating n-gram series

Each step is described in turn below.

1 **Filtering the series of action logs:** We decided to apply two types of filters to capture the metrics to provide information on the skill and process we wanted to assess:
   a. *Cleaning the log and removing the entries that were not important*: We instrumented the environment to capture everything, including looking at or hovering on the object or selecting the object. We collected between 400 and 500 data entry points for one student during a half-hour session. We decided that if a student looked or hovered on an object, these were unintentional activities, not necessarily performed to serve the task, so these were removed from the log.
   b. *Creating motif-based sections*: Motif creation for this study is already explained in section 3.4.1.
2 **Deciding the similarity index:** As mentioned earlier, we used the four indexes:
   i) Jaccard Index
   ii) Modified Jaccard Index
   iii) Cosine Similarity Index
   iv) Levenshtein Similarity Index
3 **N-gram series:**We performed the similarity assessments on both the bigram and unigram series.

4. Results

As mentioned in Section 3.2, due to technical problems, we could only use data from the participants with the HMD devices. The average time

spent by participants in each of the areas 1, 2, and 3 were 6.6, 11.3, and 4.2 min (std dev 5, 5, and 3), respectively.

In this study, we aimed to investigate the possibility of using filtered metric-based similarity analysis for assessing students' HOTS. To perform correlation analysis between similarity-based performance assessment and instructor's manual assessment, the following summarized steps were implemented, which were explained in the previous sections and are repeated here for completeness.

- Recording expert activity and all students' activities.
- Cleaning up the expert and students' data and preparing the activity series for each student, leaving only activity-based interactions in the log file – script base.
- Creating unigram bigram series for each student and expert; and
- Calculating similarity measures for each student with respect to the expert series for the four similarity indices identified over both the unigram and bigram series.

We performed a correlation analysis between instructor scores and similarity measures (four indexes, unigram and bigram) on three different filtered series from the full series where students demonstrated the following HOTS: *information gathering, critical thinking, and drawing conclusions* (Table 2). It should be noted that although students were engaged in more than one HOTS in all the phases of the learner activity, we made a conscious decision with the Subject Matter Expert (SME) to associate each phase with one specific HOTS that was most relevant.

The SME performed a manual unfiltered log trace-based assessment for each of these sections, along with an overall assessment. We then calculated the similarity between the learner's trace and the expert's trace with the four above-mentioned similarity index formulas for each student for each section. We performed this step both for unigram series and bigram series actions. As the final step, we performed a correlation analysis between the similarity measures and the SME's manual assessment. Table 2 shows the details of students' unigram-based similarity, instructor scores, and their correlation. Table 3 compares the correlation values when using unigram and bigram.

All similarity indexes in the similarity analysis showed high correlation when the series was created as unigram series. The Levenshtein similarity index, which is the only order-sensitive similarity measure, was also correlated when the bigram series was used for assessment.

To gain more insight into the most appropriate gram series that could be used while assessing learning activities, we investigated further. We manually changed randomly selected students' activity orders and calculated students' unigram similarity indexes. As expected, for the indexes where the calculation was based on the number of common elements (Jaccard or Modified Jaccard) or the number of common elements and the number of occurrences (Cosine), we got the same similarity measure. Only the Levenshtein index provided a measure change when the order of the activity changed, with a smaller or bigger similarity measure depending on the students' modified path being more similar or more dissimilar to the expert's path. When the order of the activity is important, we recommend using the Levenshtein similarity index with the bigram series.

It can be argued that the information collection activity was not an order-sensitive activity, and as such, the unigram series could be used instead of the bigram series. If that is the case, the Cosine similarity index provides the highest correlation between matrix-based assessment versus manual SME assessment. Cosine similarity checks the existence of the same data elements in both series, just like Jaccard similarity. Additionally, cosine similarity also checks the number of occurrences of common elements.

In our study, the instructor believed that for critical thinking and drawing conclusions, the order of the activities was important. As such, the bigram series should be used, and the Levenshtein similarity index provided the highest correlation on the bigram series.

The point of our research was not to choose the right similarity index or gram choice, as they depend on each case. Our goal was to show that the similarity-based method is useful and that the index can be dependent on the task, to be decided in each case by the experts.

5. Discussion

5.1. Reflections on findings and research approach

In this study, we applied motif-based learner vs. expert series similarity analysis for HOTS assessment. The study was motivated by the shortcomings of existing methods, such as the lack of localized feedback and the need for a large amount of training data. Our study answered our research questions and found that.

1. Defining motifs based on the dominant skill in a task is indeed a potentially suitable method for assessment. This was the key concept and foundation of our proposed approach, corresponding to our first research question regarding the use of motifs. While motifs were suggested by some researchers (Gibson & de Freitas, 2016), they were used in a very limited way and in score-based assessment. This meant that the order of activities was not considered. Also, our own limited past study (Nowlan et al., 2018) used activities that had a combined set of HOTS. That meant a motif could not be clearly mapped to a single skill. In this study, we defined three sections each associated with one task and one skill (as decided by the instructor). This arrangement meant that we could show how motifs could be used for assessment as a series of actions for one skill. Our approach is consistent with previous work but extends it naturally to series-based assessment, opening the possibility of evaluating how the learners perform actions and what process they follow.
2. Series-based assessment with motifs and similarity analysis was in line with the instructor's assessment, and so can effectively be used for HOTS assessment. This finding answered our second question about using similarity analysis on motifs. Question 2 was closely related to question 1 but more specific on the assessment approach. We chose to have them as separate questions, although we did not investigate motifs with any other assessment method. Our literature review showed that there was a gap related to clustering activities and performing the local evaluation, including comparing to expert activities (Reilly & Dede, 2019). As such, we were specifically looking for an understanding of how to evaluate motifs and if similarity-based analysis with expert data can be an effective way to perform a more localized evaluation without the use of large training data and black-box algorithms. Again, the findings showed that the combination of motifs and similarity analysis can be used as an assessment tool when expert data is used and with various similarity indices. This insight is important as it gives the initial confirmation and allows the option of further investigating multiple experts and multiple similarity indices/measures.
3. Applying different similarity measures is potentially more effective and possible when using motifs. Last but not least, finding a somewhat unexpected one as we were hoping to find one similarity measure that outperforms others but found out that there could be different ones depending on the task or skill. The literature on similarity measures already suggests the possibility of combining different indices (Loh & Sheng, 2014). So, the idea of the relative suitability of indices for different HOTS or tasks is in line with the research direction and further suggests more research to explore this relative suitability.

In our analysis, we focused on similarity-based performance assessment as opposed to machine model creation methods because we believe it is more practical and easier to apply in the education world. Below is the summary of the justification of our assessment approach.

Table 2
Correlation analysis of SME assessment versus similarity index-based calculation.
Columns 1–18 are for participants. Correlation values are between SME score and similarity measures for each of the indexes (unigram). IC: Information Collection, CT: Critical Thinking, DC: Drawing Conclusions.

		1	2	3	4	5	6	7	8	9	10
IC	**Score**	11	11	11	11	11	11	11	11	10	11
	Jaccard Similarity	0.733333	0.654545	0.705882	0.425926	0.653846	0.716981	1	0.724138	0.666667	0.741379
	Cosine Similarity	0.851843	0.791257	0.828956	0.609071	0.792629	0.835215	1	0.841516	0.800198	0.854886
	Mjaccard Similarity	0.956522	0.782609	0.782609	0.5	0.73913	0.826087	1	0.792453	0.782609	0.934783
	Levenshtein Similarity	−0.34615	−0.45055	−0.2069	−0.61039	−0.11628	−0.35165	0.4456	0.18	−0.17778	−0.58416
CT	**Score**	20	20	20	20	20	20	20	20	19	20
	Jaccard Similarity	0.639344	0.649123	0.238095	0.566667	0.639344	0.631579	0.8113	0.609756	0.580645	0.649123
	Cosine Similarity	0.781408	0.793728	0.412082	0.729372	0.781408	0.78187	0.9007	0.757576	0.737154	0.793728
	Mjaccard Similarity	0.735849	0.698113	0.283019	0.641509	0.735849	0.679245	0.8113	0.757576	0.679245	0.698113
	Levenshtein Similarity	0.06	0.117021	−0.33333	0.085106	−0.48	0.064516	0.7187	0.318182	0.153061	−0.32979
DC	**Score**	15	15	3	18	20	16	12	18	18	15
	Jaccard Similarity	0.578947	0.552632	0.333333	0.560976	0.969697	0.735294	0.4166	0.5	0.658537	0.486486
	Cosine Similarity	0.737028	0.716928	0.57735	0.719101	0.984732	0.853486	0.6154	0.67082	0.794461	0.668043
	Mjaccard Similarity	0.666667	0.636364	0.333333	0.69697	0.969697	0.757576	0.4545	0.6	0.818182	0.545455
	Levenshtein Similarity	0.366667	0.423729	−0.02273	0.390625	0.846154	0.40678	0.1960	0.333333	0.5	0.272727

- Machine model creation-based assessment requires large amounts of training data to create a model before it can be used for assessment. Therefore, it is more time-consuming as it requires a large amount of participants' data to train the model.
- Curricula always need to be adapted to learners. Changing the curriculum would mean retraining the model and collecting data to retrain the model and stops classroom teacher from using the curricula.
- Machine model-based assessment is more appropriate for an overall assessment and does not provide actionable information for learners to improve specific areas for better performance.

With a similarity-based assessment model, and with a 3DVLE-integrated data recording and assessment tool, the above-listed issues can be resolved for classroom teachers. Additionally, adding a practice mode to curricula might allow the platform to perform an ongoing assessment during the learning session over smaller sections as they are completed and potentially can provide feedback to students to foster self-reflection.

5.2. Limitations and further research

In performing this study, we had some unforeseen challenges. Our first challenge was regarding data usability. Initially, we had hoped that our instructor partner could watch video screen recordings of the learners' activities and perform a holistic assessment. Due to COVID-19, all our participants had to perform the learning activity at home. Recording the HMD screen and uploading the video posed a technical challenge to our non-technical study participants, and we could not get the videos as planned. Therefore, we had to adjust our research methodology; our SME followed learners' full log files manually and recreated/visualized students' activities to perform the assessment. Better tools for collecting and preparing the data should be investigated in future research.

The next challenge was the VR platforms. The research team planned to have two platforms for delivering the learning curriculum to research participants. As planned, our app was designed to be experienced on both head-mounted displays (HMD) and desktop screens. The participants' data was to be collected from both platforms. However, due to design oversights in the desktop app, we could only use the data coming from the HMD app. So, the data from participants who used the desktop app could not be studied, and we could not replace them with new participants with HMD due to the limited availability of users.

In our study, we also used one expert recording as the emergency activity should be performed short and well-defined without many variations. However, multiple acceptable expert paths may be possible in some cases. Future studies may consider adapting this multiple expert paths similarity assessment for a more flexible assessment, but the general approach could stay the same.

Novelty is always a potential factor when investigating emerging technologies such as VR (Chaudhry, 2021; Miguel-Alonso, Rodriguez-Garcia, Checa, & Bustillo, 2023). This factor can sway the user feedback and result in more positive responses, due to the "wow" effect or negative ones, due to the difficulty and lack of familiarity. We tried to have participants who have a mix of different experience levels with VR to avoid bias, but unfortunately, most of our participants were fairly new to this technology. We also introduced the general VR training area to limit the novelty effect. Still, we admit the possibility of some positive or negative effects, which can be explored further in the future.

We are aware of the concerns that when students know they are "being watched," they may feel additional stress. An opportunity to perform the educational activities "privately" while still receiving feedback will be a valuable addition. Regardless of this "private mode," the principles for the good use of data should be followed and improved upon. Such principles include (1) using data that is directly related to what we study, (2) direct benefit and ownership for students and instructors, and (3) consent and transparency (D'Ignazio & Klein, 2020; O'Neil, 2017).

5.3. Implications and recommendations

Overall, our research showed the potential of using motifs and expert data in HOTS assessment. But at the same time, it was a testament to the complexity of the assessment process and the need to have multiple methods. While previous work (see Section 2) was successful in some cases, our proposed method could be more useful when dealing with procedures that can be broken into clear motifs. On the other hand, even within our study, multiple similarity measures proved useful in different cases. This implies that there is no unique method that can serve as a silver bullet, and researchers and instructors should be aware of and equipped with a variety of tools as appropriate. Further research is recommended to investigate how and when these methods can be appropriate and how different sets of them can be combined.

Another important implication of our research is the emphasis on the role of instructors and course designers. Defining motifs and expert paths is clearly within the expertise of domain specialists. This important role means that (1) research and development projects on assessment methods and tools should be done in close collaboration with these specialists, and (2) the tools should be designed and developed to be easy

11	12	13	14	15	16	17	18	*Correlation*
11	11	4	5	5	5	5	8	
0.788462	0.763636	0.25	0.104167	0.258621	0.25	0.267857	0.188679	0.868652
0.881771	0.867132	0.42135	0.278639	0.425628	0.410689	0.442326	0.357599	0.89276923
0.891304	0.913043	0.304348	0.108696	0.326087	0.326087	0.326087	0.217391	0.87228291
−0.11828	−0.16495	−1.04286	−1.83019	−0.91781	−0.8	−0.90141	−1.31746	0.75789006
20	19	18	5	5	5	5	2	
0.532258	0.684211	0.035088	0.028986	0.075472	0.018519	0.036364	0.018868	0.83667506
0.699441	0.816944	0.112154	0.064752	0.274721	0.097129	0.137361	0.137361	0.84057219
0.622642	0.735849	0.037736	0.037736	0.075472	0.018868	0.037736	0.018868	0.84914705
−0.06316	−0.85417	−0.83051	−0.53521	−0.87719	−1	−0.92982	−1.03704	0.70847742
15	0	20	18	16	20	20	13	
0.52381	0	0.675	0.65	0.615385	0.714286	0.5	0.414634	0.8313596
0.687836	0	0.80606	0.787879	0.76277	0.836242	0.668994	0.591864	0.83151594
0.666667	0	0.818182	0.787879	0.727273	0.909091	0.727273	0.515152	0.92641691
0.265625	−0.4	0.298507	0.393939	0.365079	0.236111	0.263889	0.258621	0.80546362

to use for instructors. Course designers and instructors aiming to incorporate our proposed method and other similar assessment approaches are recommended to.

- define clear (and multiple) paths that the learners may take to solve the assigned problems (for example, algorithmic thinking approach for math/programming courses and procedures for labs),
- break them into smaller elements to consider how to evaluate the higher-order thinking, not as a single activity but as a series of smaller ones (for example, various steps for problem-solving and logical flow in a debate),
- consider multiple expert solutions and paths to allow flexibility in assessment.

Following up on the role of educational experts, this study confirmed the difficulty of having a variety of expertise that increases the cost and time involved in running metric-based assessments. In addition to the subject matter expert (instructor), we needed to receive assistance from programmers, content developers, and data analysts, expertise that may be too difficult and costly for course designers to hire. Similarly, various required tools may not be easily available. As such, we identified the lack of an integrated HOTS assessment framework with proper collection and analysis tools for process metrics as a major gap in this area. We propose the framework to be developed as a standalone unit to provide full flexibility and reusability and be connected with multiple 3DVLE platforms when needed. Such a framework should provide components such as a proper user interface, integration with 3DVLE platforms, data collection, assessment using various methods, visualization, and feedback. AI algorithms for assessment and providing feedback and suggestion are necessary parts of this framework that should be investigated.

6. Conclusion

This paper investigated the use of motifs and expert data in series-based HOTS assessment in 3DVLEs. These platforms, with their ability to offer remote virtual classrooms, play an important role when physical classrooms cannot be used. We showed that motifs as small meaningful elements of learning activity have a strong potential to be used for series-based assessment. We also demonstrated the value of using expert data and multiple similarity measures.

The most important implication of our research on HOTS assessment is offering practical guidance to instructors on how to set up 3DVLE-based assessment approaches that match their knowledge of the process. This is done by defining and using motifs, comparing them to expert data, and using appropriate similarity measures. Our work showed examples of how to define motifs and what measures to use.

Table 3
Similarity index-based correlation analysis. Unigram vs. Bigram.

Motif/HOTS	Similarity Measure Used	Unigram-based Correlation with Instructor Scores	Bigram-based Correlation with Instructor Scores
Information Collection	Jaccard	0.8687	0.4651
	Cosine	0.8928	0.5811
	Modified Jaccard	0.8723	0.6104
	Levenshtein	0.7579	0.7065
Critical Thinking	Jaccard	0.8367	0.4833
	Cosine	0.8406	0.5989
	Modified Jaccard	0.8491	0.5940
	Levenshtein	0.7085	0.7243
Drawing Conclusions	Jaccard	0.8314	0.3774
	Cosine	0.8315	0.4694
	Modified Jaccard	0.9264	0.5258
	Levenshtein	0.8055	0.7374

We envision that in the future, with the help of automated assessment tools, classroom teachers will be able to design more experiential tasks for their students and integrate them into their teachings. This will also help them provide individualized feedback on the process. There are several limitations in our research (listed in Section 5.2) that can be addressed in future research. Our work in this research only investigated individual learner processes and assessments. However, future research might investigate team-based learning tasks and the assessment of each member's skills.

Statements on open data and ethics

This study was approved by our university's Research Ethics Board (#113105) to comply with various requirements such as privacy and consent, proper data collection and storage, safety (especially during COVID pandemic), and ethical use of data. Inspired by current research on social and ethical responsibilities when working with user data (D'Ignazio & Klein, 2020; O'Neil, 2017), we defined the following three principles as guidelines for our research.

- Using data that directly corresponds to what is studied (no proxy data)
- Providing direct benefit and ownership for students (using data only for academic assessment)

- Consent and transparency in collecting and analyzing data (clear algorithms)

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded in parts by Ontario Centre of Innovation under Grant #33593.

Abbreviations

(3DVLE) Three-Dimensional Virtual Learning Environment
(HMD) Head-Mounted Display
(HOTS) Higher-Order Thinking Skills
(LMS) Learning Management System
(MSI) Maximum Similarity Index
(SME) Subject Matter Expert
(VLE) Virtual Learning Environment
(VR) Virtual Reality

References

Abdullah, M. H. (1998). *Problem-based learning in language instruction: A constructivist model. Bloomington, ERIC clearinghouse on reading English and communication*. Retrieved from http://www.ericdigests.org/1999-2/problem.htm.
Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., Torres, C., Haertel, G., Dolan, R., Beddow, P., & Lazarus, S. (2010). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—a foundation for research. *The Journal of Technology, Learning, and Assessment, 10*(5).
Alqahtani, A. S., Daghestani, L. F., & Ibrahim, L. F. (2017). Environments and system types of virtual reality technology in STEM: A survey. *International Journal of Advanced Computer Science and Applications, 8*(6).
Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education, 24*(4), 387–426.
Arya, A., Hartwick, P., Graham, S., & Nowlan, N. (2012). Collaborating through space and time in educational virtual environments: 3 case studies. *The Journal of Interactive Technology & Pedagogy, 2*. Retrieved from http://jitp.commons.gc.cuny.edu/collaborating-through-space-and-time-in-educationalvirtual-environments-3-case-studies/.
Azarnoush, H., Alzhrani, G., Winkler-Schwartz, A., Alotaibi, F., Gelinas-Phaneuf, N., Pazos, V., Choudhury, N., Fares, J., DiRaddo, R., & Del Maestro, R. F. (2015). Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *International Journal of Computer Assisted Radiology and Surgery, 10*(5), 603–618.
Baker, R., & Clarke-Midura, J. (2013). Predicting successful inquiry learning in a virtual performance assessment for science. In *International conference on user modeling, adaptation, and personalization*. Berlin, Heidelberg: Springer.
Bennett, H. (2003). Successful K-12 technology planning: Ten essential elements. *Teacher Librarian, 31*(1), 22.
Borgman, C. L., Abelson, H., Dirks, L., Johnson, R., Koedinger, K. R., Linn, M. C., Lynch, C. A., Oblinger, D. G., Pea, R. D., Salen, K., Smith, M. S., & Szalay, A. (2008). Fostering learning in the networked world: The cyberlearning opportunity and challenge, a 21st century agenda for the National Science Foundation. *Report of the NSF task force on cyberlearning, 59*.
Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education, 24*(4), 387–426.
Chaudhry, M. (2021). Creating effective virtual reality learning experiences: Lessons learned. In *Education and training in optics and photonics* (p. Th4A-1). Optica Publishing Group.
Chen, X., Zou, D., Xie, H., & Wang, F. L. (2021). Past, present, and future of smart learning: A topic-based bibliometric analysis. *International Journal of Educational Technology in Higher Education, 18*(1), 1–29.
Code, J., & Zap, N. (2013). Assessments for learning, of learning, and as learning in 3D immersive virtual environments. In *EdMedia: World conference on educational media and technology*. Association for the Advancement of Computing in Education (AACE).
Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). *AI in education needs interpretable machine learning: Lessons from open learner modelling*. arXiv preprint arXiv: 1807.00154.
D'Ignazio, C., & Klein, L. (2020). *Data feminism*.
Dede, C. (2007). Reinventing the role of information and communications technologies in education. *The Yearbook of the National Society for the Study of Education, 106*(2), 11–38.
Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*(5910), 66–69.
Dede, C., & Ketelhut, D. (2003). *Motivation, usability, and learning outcomes in a prototype museum-based multi-user virtual environment*. American Educatonal Research Conference.
Duncan, I., Miller, A., & Jiang, S. (2012). A taxonomy of virtual worlds usage in education. *British Journal of Educational Technology, 43*(6), 949–964.
Edwards, A. W., & Cavalli-Sforza, L. L. (1965). *A method for cluster analysis* (pp. 362–375). Biometrics.
Elme, L., Jørgensen, M. L., Dandanell, G., Mottelson, A., & Makransky, G. (2022). Immersive virtual reality in STEM: Is IVR an effective learning medium and does adding self-explanation after a lesson improve learning outcomes? *Educational Technology Research & Development, 70*(5), 1601–1626.
Floryan, M., Dragon, T., Basit, N., Dragon, S., & Woolf, B. (2015). Who needs help? Automating student assessment within exploratory learning environments. In *International conference on artificial intelligence in education* (pp. 125–134). Cham: Springer.
Fullan, M., & Langworthy, M. (2013). *Towards a new end: New pedagogies for deep learning*. Retrieved from http://www.newpedagogies.nl/images/towards_a_new_end.pdf.
Gibson, D., & de Freitas, S. (2016). Exploratory analysis in learning analytics. Technology. *Knowledge and Learning, 21*(1), 5–19.
Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36*(3), 189–213.
Griffin, P., & Care, E. (2014). *Assessment and teaching of 21st-century skills: Methods and approach*. Springer.
Hill, P. (2013). *Personal correspondence with author. Towards a new end: New pedagogies for deep learning*. Retrieved from http://redglobal.edu.uy/wpcontent/uploads/2014/07/New_Pedagogies_for_Deep-Learning_Whitepaper1.pdf.
Hopson, M. H., Simms, R. L., & Knezek, G. A. (2001). Using a technology-enriched environment to improve higher-order thinking skills. *Journal of Research on Technology in Education, 34*(2), 109–119.
Hussain, M., Hussain, S., Zhang, W., Zhu, W., Theodorou, P., & Abidi, S. M. R. (2018). Mining moodle data to detect the inactive and low-performance students during the moodle course. In *Proceedings of the 2nd international conference on big data research* (pp. 133–140). ACM.
Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist, 11*(2), 37–50.
Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.
Kasim, N. N. M., & Khalid, F. (2016). Choosing the right learning management system (LMS) for the higher education institution context: A systematic review. *International Journal of Emerging Technologies in Learning, 11*(6).
Kelman, P. (1989). Alternatives to integrated instructional systems. In *Paper presented at the national educational computing conference* (Nashville, TN).
Ketelhut, D. J., Nelson, B. C., Clarke, J., & Dede, C. (2010). A multi-user virtual environment for building and assessing higher order inquiry skills in science. *British Journal of Educational Technology, 41*(1), 56–68.
King, F. J., Goodson, L., & Rohani, F. (1998). *Higher order thinking skills: Definition, teaching strategies, assessment*. Publication of the Educational Services Program (now known as the Center for Advancement of Learning and Assessment).
Koper, R. (2014). Conditions for effective smart learning environments. *Smart Learning Environments, 1*(1), 5.
Kuang, T. M., Adler, R. W., & Pandey, R. (2021). Creating a modified monopoly game for promoting students' higher-order thinking skills and knowledge retention. *Issues in Accounting Education, 36*(3), 49–74.
Leighton, J. P. (2011). Cognitive model for the assessment of higher order thinking in students. In D. H. Robinson, & G. J. Schraw (Eds.), *Assessment of higher order thinking skills* (pp. 151–181). Charlotte, NC: Information Age Publishing.
Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady, 10*(8), 707–710. Bibcode:1966SPhD...10..707L.
Loh, C. S., & Sheng, Y. (2014). Maximum similarity index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior, 39*, 322–330.
Loh, C. S., & Sheng, Y. (2015a). Measuring the (dis-) similarity between expert and novice behaviors as serious games analytics. *Education and Information Technologies, 20*(1), 5–19.
Loh, C. S., & Sheng, Y. (2015b). Measuring expert performance for serious games analytics: From data to insights. In *Serious games analytics* (pp. 101–134). Cham: Springer.
Miguel-Alonso, I., Rodriguez-Garcia, B., Checa, D., & Bustillo, A. (2023). Countering the novelty effect: A tutorial for immersive virtual reality learning environments. *Applied Sciences, 13*(1), 593.
Młynarska, E., Greene, D., & Cunningham, P. (2016). Time series clustering of moodle activity data. In *24th Irish conference on artificial intelligence and cognitive science (AICS'16)* (pp. 20–21). Dublin, Ireland: University College Dublin. September 2016.
Nowlan, N. S., Hartwick, P., & Arya, A. (2018). Skill Assessment in Virtual Learning Environments. In *2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2018* pp. 1–6).
O'Neil, C. (2017). *Weapons of math destruction*. Crown.
Peffer, M., Quigley, D., & Mostowfi, M. (2019). Clustering analysis reveals authentic science inquiry trajectories among undergraduates. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 96–100).

Queiroz, A. C. M., Nascimento, A. M., Tori, R., & Silva Leme, M. I. D. (2019). Immersive virtual environments and learning assessments. In *International conference on immersive learning* (pp. 172–181). Cham: Springer.

Reilly, J. M., & Dede, C. (2019). Differences in student trajectories via filtered time series analysis in an immersive virtual world. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 130–134).

Robinson, D. H., & Schraw, G. J. (2011). *Assessment of higher order thinking skills*. Charlotte, NC: Information Age Publishing.

Sabourin, J. L., Shores, L. R., Mott, B. W., & Lester, J. C. (2013). Understanding and predicting student self-regulated learning strategies in game-based learning environments. *International Journal of Artificial Intelligence in Education, 23*(1–4), 94–114.

Sawyer, R., Rowe, J., Azevedo, R., & Lester, J. (2018). *Filtered time series analyses of student problem-solving behaviors in game-based learning*. International Educational Data Mining Society.

Scavarelli, A., Arya, A., & Teather, R. J. (2021). Virtual reality and augmented reality in social learning spaces: A literature review. *Virtual Reality, 25*(1), 257–277.

Schmidt, B., & Stewart, S. (2009). Implementing the virtual reality learning environment: Second Life. *Nurse Educator, 34*(4), 152–157. https://doi.org/10.1097/NNE.0b013e3181aabbe8. PMID: 19574850.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction, 55*(2), 503–524.

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 311–321). New York: Springer.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education, 80*, 58–67.

Snow, E. L., Jacovina, M. E., & McNamara, D. S. (2015). Promoting metacognition within a game-based environment. In *Proceedings from international conference on artificial intelligence in education* (pp. 864–867). Springer International Publishing.

Spector, J. M., & Ma, S. (2019). Inquiry and critical thinking skills for the next generation: From artificial intelligence back to human intelligence. *Smart Learning Environments, 6*(1), 1–11.

Van Voorhis, V., & Paris, B. (2019). Simulations and serious games: Higher order thinking skills assessment. *Journal of Applied Testing Technology, 20*(S1), 35–42.

Veenman, M. V., Bavelaar, L., De Wolf, L., & Van Haaren, M. G. (2014). The online assessment of metacognitive skills in a computerized learning environment. *Learning and Individual Differences, 29*, 123–130.

Warburton, S. (2009). Second Life in higher education: Assessing the potential for and the barriers to deploying virtual worlds in learning and teaching. *British Journal of Educational Technology, 40*(3), 414–426. https://doi.org/10.1111/j.1467-8535.2009.00952.x

Winkler, W. E. (1999). *The state of record linkage and current research problems*. Statistical Research Division, US Census Bureau.