Assessment of Higher Order Thinking Skills in Virtual Learning Environments

by

Nuket Savaskan Nowlan

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Information Technology

Carleton University

Ottawa, Ontario

Abstract

This research addresses the assessment of Higher-Order Thinking Skills (HOTS), such as metacognition, reflection, and problem-solving, in Virtual Learning Environments (VLEs). We particularly focus on the use of process metrics, their combinations, and various analysis methods that allow VLE platforms to perform automated HOTS assessments. Traditional learning assessments rely mostly on outputs and are not suitable for HOTS assessment that requires process observation. As a result, it is a challenge for learners and educators to identify the areas of weakness for customized help when it comes to HOTS. Our objective is to understand the requirements of a VLE-based HOTS assessment framework and explore what process metrics can be used and how they can be analyzed to offer insight into learners' HOTS development.

To achieve the above objective, a series of four studies were performed within this research. Study 1 was an initial exploratory investigation that suggested 3D VLEs as a possible HOTS fostering platform though associating their unique affordances to the requirements of common learning theories. This study was a motivational activity and initiated our research. Study 2 was performed on a text-based VLE and provided new insight into how aggregated process metrics can be used to represent student attention and participation, which are linked to HOTS. Study 3 focused on identifying 3D VLE process metrics and their alignment with HOTS components. Study 3 results suggested that the rich data coming from a 3D VLE, and the combination of process metrics as small groups (motifs) and time series, can offer more insights about HOTS. Finally, Study 4 employed motifs and time series-based similarity analysis on process metrics for performing HOTS assessment during learning tasks in a 3D VLE. Study 4 investigated task compatibility with four different similarity indexes, and findings suggested employing different similarity indices depending on the learning tasks.

Overall, the studies conducted within the scope of this research provided supporting evidence of the possibility of automated HOTS assessment on VLEs using process metrics. They showed the value of motifs (small yet meaningful series of process metrics) as a measure for HOTS. However, they suggested that there is no single method, and different learning tasks might use different data analysis strategies.

Acknowledgments

It is so hard for me to believe that I am at the finishing line. Looking back reflecting on kind, very knowledgeable, wise individuals that have supported me through all these years, it takes almost a village for one to get a PhD.

I like to thank Peggy Hartwick for being my research partner first and then becoming a dear friend. Her being with me on this journey made the path a lot easier, meaningful, and fun.

I like to thank the School of Linguistics and Applied Language Studies (SLaLS), John Osborne, and Randall Gess. Thank you for believing in this research and supporting the proposed 3D Virtual Carleton initiative.

Hossain Samar Qorbani, thank you for sharing your talent and creating a special app for us to use in our research. Very much appreciated!

Maryam Abdinejad, it was indeed a pleasure collaborating with you on our research. Your professionalism and attention to detail was impressive.

Thank you very much Lois Frankel for reading my thesis in advance and providing the most valuable feedback on how to improve and better word some of my thoughts and findings.

My friend and project partner, Walter Krumshyn, thank you for lending me your technical support whenever I needed it.

My proofreader Nina Dore, thank you for reading and rereading and providing feedback and suggestions. Writing a thesis in my second language was only possible with your ongoing support.

Above all, I am extremely grateful to my supervisor, Ali Arya for his continuous support, wisdom, invaluable guidance, and patience during my PhD study. His immense knowledge on a wide range of areas and plentiful experience have encouraged me in all the time of my academic research, allowing me to continue my study. It is so rare to find a supervisor, who instills the authority and respect to force you to work hard when needed, but also offers kind words and understanding when you need them as well. I cannot thank him enough.

John, Isinsu, Yasmin; all becomes meaningful when it is shared with you. Thank you for being in my life. Love you.

Finally, I like to dedicate my thesis to my parents Ismet and Hikmet Savaskan for teaching me the value of hard work and the importance of knowledge creation. I will be forever missing them.

Table of Contents

Table of Figures

List of Tables

Glossary

Active Learning	A broad range of teaching strategies which engage students as active participants in their learning during class time with their instructor.
Artificial Neuron Network	A computational model that mimics the way nerve cells work in the human brain.
Augmented Reality (AR)	An interactive experience of a real-world environment where the objects that reside in the real world are enhanced by computer-generated perceptual information.
Basic Process Metrics	Metrics showing user's specific action, e.g., picking up an object, moving into a defined separated space.
Computer-Based Assessment	Assessment performed by a computer based on a predefined coded algorithm
Computer-Based Learning Environments (CBLE)	Computer-created platforms designed to support learning. Also referred as Virtual Learning Environments (VLEs).
Combined Process Metrics	Metrics that bring together multiple basic process metrics to provide better insight on a complicated activity
Constructivist Learning	A learning model/theory based on learners constructing knowledge through interaction as opposed to passively receiving information
Experiential Learning	Engaged learning process where students learn by doing and reflecting on experiences.
Head-Mounted Display (HMD)	A small display or projection technology that is integrated into eyeglasses or mounted on a helmet or hat
Higher-Order Thinking Skills (HOTS)	Logical, reflective, and creative thinking activated when individuals encounter unfamiliar problems, uncertainties, questions, or dilemmas.
Motif	Small yet meaningful series of learning actions in 3DVLEs to demonstrate a process. Generally, basic

	metrics collected on motifs form process metrics for that small unit of task.
Multi-dimensional Process Metrics	Process metrics capturing more than one type of HOTS information
Moodle	Modular Object-Oriented Development Learning Platform, http://www.moodle.org
Output Metrics	Metrics capturing results and output data.
Performance-based assessment	Assessment performed by observing performance throughout the learning process.
Problem- and inquiry-based learning	Teaching method in which problems are used as the vehicle to promote student learning of concepts and principles (as opposed to the direct presentation of facts and concepts). PBL can also promote the development of critical thinking skills, problem-solving abilities, and communication skills.
Process Metrics	Metrics capturing the type, order, and quantity of interactions to provide information on HOTS demonstrated. Process metrics can be defined for a single HOTS or can capture multiple HOTS information as in combined process metrics.
Series similarity analysis	Calculation of the overlapping percentage between two sets of data by employing a similarity index of choice.
Similarity index (measure)	In data mining, a distance with dimensions describing object features. If the distance between two data points is small, there is a high degree of similarity among the objects (and vice versa).
Subject Matter Expert (SME)	A person who is an authority in a particular area or topic.
Situated Learning	The concept that learning occurs within authentic context, culture, and activity and that it is widely unintentional.
3D Virtual Environments (3DVE)	A virtual representation of three-dimensional (3D) space.

3D Virtual Learning Environment (3DVLE)	A 3DVE for learning purposes.
21st Century Skills	The mental skills required in an increasingly competitive 21st century environment, such as drawing conclusions from vast amounts of information.
Virtual Reality (VR)	An immersive or non-immersive representation of a fully virtual space. In this dissertation and in recent usage, VR and 3DVE are used interchangeably.
Virtual Learning Environments (VLE)	Computer-created platforms designed to support learning. Also referred to as CBLE

1 Introduction

"I'm calling on our nation's governors and state education chiefs to develop standards and assessments that don't simply measure whether students can fill in a bubble on a test, but whether they possess 21st century skills like problem-solving and critical thinking and entrepreneurship and creativity."

President Barack Obama,

Remarks to Hispanic Chamber of Commerce, March 10, 2009

1.1 Background

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; Aldrich, 2009; Almond et al., 2010). The demands of the 21st century (such as critical thinking, problem-solving, and drawing conclusions) have especially reinforced the importance of Higher-Order Thinking Skills (HOTS), which are the "critical, logical, reflective, metacognitive, and creative thinking activated when individuals encounter unfamiliar problems, uncertainties, questions, or dilemmas" (King et al., 1998). Resnick (1987) claimed that HOTS are the hallmark of any successful education and have long been the standard for elite education (i.e., only offered to privileged groups). However, changing economic and social conditions, such as the offloading of repetitive work tasks to computers and competition from the world population due to globalization, have made it essential for everyone to develop these skills.

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). Current output-based assessment approaches are not suitable for assessing HOTS (Fullan, 2013). Output-based approaches focus on the result (output or outcome) of the learning process, which 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*) (Griffen & Care, 2014). Using process metrics provides information on the learning methods and strategies that learners use throughout tasks and thus better capture HOTS development (Greiff, Wüstenberg, & Funke, 2012).

Unfortunately, observing each individual learner performing a task is extremely time-consuming and requires significant resources. To address this problem, researchers have turned to *computer-based assessment* as a way of capturing rich information about the learning process (Burleson, Muldner, Rai, & Tai, 2014). Computer-based platforms designed to support learning are commonly called Computer-Based Learning Environments (CBLEs) or Virtual Learning Environments (VLEs). They support various forms of assessment, generally referred to as computer-based assessments. The most used VLEs are Learning Management Systems (LMS[1]) such as Moodle[2] and BrightSpace[3]. 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.

A particular type of emerging 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 (Nowlan et al., 2011). 3DVLEs allow avatar-based interaction and navigation of virtual spaces that can simulate real ones or visualize novel spaces (Arya, 2012). They have been increasingly attracting the attention of scholars interested in skills development due to their ability to facilitate simulated experiential learning (Schmidt, 2009). They also have the potential to observe learners throughout the process and track and assess a significantly wider range of user activities (Loh, 2015).

While new VLE (especially 3DVLE) platforms are emerging with sophisticated features to design rich experiential curricula, to the best of our knowledge, no VLE platform available on

[1] Also knows as Learning Content Management System (LCMS)
[2] https://moodle,org
[3] https://www.d2l.com/

the market yet offers digital tracking and built-in collected data analysis capabilities for process metrics. It can be argued that 3DVLEs will not reach their potential as a mainstream educational technology until process-based assessment methods and instruments, along with a design framework, are developed. Such tools would allow educators to take advantage of the unique ability of 3DVLEs to collect and analyze process metrics. They can also benefit non-3D VLEs to the level that they can collect process metrics. Considering the importance of using process metrics, the challenge of using VLEs for HOTS development is twofold:

- Identifying and collecting specific metrics as an indication of HOTS and their components.
- Identifying the algorithms and methods of data analysis to perform HOTS assessment.

The reliance on process-based metrics and data also raises concerns about the ethical use of data, which is not currently properly addressed in virtual environments literature and industry (Moon, 2017; Aleksieva-Petroya, 2019).

1.2 Problem Statement

Throughout an educational experiential learning session, a student employs multiple HOTS to successfully complete the task. In a typical output-based assessment, multiple questions would be asked about different sections of the context to find out the student's knowledge gaps, which would provide the student with feedback on areas for improvement. In a HOTS assessment, however, feedback should include both areas of weakness and strength for the student to focus on. This requires the collection and analysis of granular data (process metrics) from a HOTS perspective. Researchers have studied VLE-collected metrics for assessing learners' knowledge gains and HOTS development and proposed notable skills profiling and assessment methods (Shute, Kim, 2015; Loh, Shen, 2014; Snow at. al 2015, Sawyer at.al., 2018). Several of these methods are summarized in the following sections. However, what process metrics to use, how to collect them, and what methods to use to analyze them are still open questions. Most existing VLEs do not provide proper data collection and analysis tools, and using the limited existing functionality requires technical support, which is not convenient for instructors. Through an extensive literature review (Chapter 2), we identified that using VLEs to foster HOTS requires addressing two essential problems:

(1) Understanding of the type of metrics and analysis methods to use for providing insight to/about learners and support their HOTS development.

(2) Integrating the required data collection, assessment, and visualization tools within an easy-to-use framework for instructors and course designers.

The primary goal of this research is to directly address the first problem (Chapters 3-6) and provide initial insights and recommendations for the second one (Chapter 7).

Collecting and tracking digital learning data raises many ethical and social issues. User data can provide more customized services but can also lead to abuse, discrimination, and privacy breaches (D'Ignazio and Klein, 2020; O'Neil, 2016). Data scientists and privacy advocates have highlighted the need for "doing good with data" and offered design guidelines for those working with data collection and analysis tools such as Artificial Intelligence (AI) algorithms in various fields from advertisement to policy making (D'Ignazio & Klein, 2020). The ethical uses of learners' data in VLEs, on the other hand, are less investigated. It is, therefore, our intention to offer general ethical guidelines for using VLE metrics in performance assessment, or what we[4] refer to as "the good use of learning data". This "good use" encapsulates using "good data" (as opposed to, for example, "big data") and "doing good with data".

While our primary focus is on 3DVLEs due to their extensive data collection abilities, some of our research findings can be used for other VLEs to the level they allow data collection. To allow some level of generalization, our studies have considered different types of educational subjects (English language learning and Chemistry) and also 3D and non-3D VLEs. We focus on university courses for the studied subjects and use the terms *learner* and *student* interchangeably, but we expect that our findings can benefit other forms and levels of learning as well, although a discussion of such benefits is beyond the scope of this dissertation.

1.3 Research Approach

In this research we used mixed-method approach. In term of data collection both qualitative (Study 1) and quantitative (Study 2 & 3 & 4) approach were used. Our research methodology was cyclical

[4] Throughout this document, and unless specifically defined, the words "we," "us," and "our" refer to the author, her supervisor, and any research partners for individual studies. The supervisor provided guidance for the whole research process. Individual studies had partners as subject matter expert or technical developer whose roles are explained in related sections.

and reflective. We were inspired by Action Research methodology with some similarities and differences. Action research can be described as "an approach in which the action researcher and a user (client) collaborate in the diagnosis of the problem and in the development of a solution" (Bryman and Bell,2011). Our research methodology was similar to the action research in the sense that we started with a problem of assessing HOTS in virtual learning environments and designed and applied proposed solutions together with our users (classroom teachers in our case). Our study was cyclical and reflective as we went through cycles (studies) while reflecting on findings of each to design and move on to the next. Our research was different from the action research in the sense that although research overall objective remains the same, identifying process metrics and analysis methods, the content of the course, and study-specific research questions changed in each cycle. We were very careful validating our choices of course content and specific research questions with educational theories studied in detail in Study 1 on each cycle to ensure our alignment with the theories guiding us through and our connectedness to overall research objective.

To address the main problems of our research, we performed a literature review to gain insight into what has been done so far for HOTS assessment and what possibilities exist. As discussed in the next chapter, the majority of existing methods use units of action by users, such as accessing a resource or viewing an object, and apply some form of scoring for these actions to calculate performance indices. As we see later, this approach lacks flexibility, requires a large amount of learning data, and does not provide feedback on HOTS. A more suitable approach has been suggested and used by some researchers that includes a series of actions as the basis of assessment (Loh and Shen, 2014). While providing certain flexibility and the ability to compare to an expert without much learning data, as we discuss in Chapter 2, series-based methods still lack proper feedback on learning elements as it provides overall single score similarity assessment. The proper definition of series and the analysis methods for them also need more research.

While VLEs can collect and offer *basic process metrics*, i.e., data on single elements such as triggering an event, picking up an object, or accessing a resource, it is hard to associate such metrics individually with learning or HOTS. Existing literature, as discussed in Chapter 2, considers these metrics with or without attention to their order. In this dissertation, we proposed the use of *combined process metrics* that bring together multiple basic metrics. To solve the metric identification aspect of the research problem, we borrowed the concept of a *motif* as a small and purposeful sequence of action in VLEs (Gibson & Freitas, 2016) and used it as the basis of the

process metrics for assessment. Such a sequence is an example of a *time series*, which is a set of items that are sequential in time. Time series have been investigated by researchers in machine learning and data science and can be the temporal values of the same variable or a sequence of actions (Ahadi et al., 2015; Aghabozorgi, 2015). We used motifs as a special case of time series where users's actions change over time and create a series of components with stamps. We show that such motifs can demonstrate competency on a unit of learning or HOTS. For example, a user who performs picking up an object, manipulating it, dropping it, and looking at a results screen with particular timing may form a motif related to a part of an experiment in a chemistry lab. Such a motif can also be created by only collecting information related to activities over a logically related time frame, such as reading a document or zooming in to check the reading on a machine, to assess information collection skills. Motifs allow instructors to define activities and their related patterns that are meaningful for learning, and so can be used for assessment. Our motif-based approach provides flexibility and feedback on specific parts of the learning process. It also allows for the definition of data collection systems that are transparent and can be used directly for the benefit of the learner.

Other combined metrics used in our research included temporal values of a single metric and *aggregated process metrics*, which are multi-dimensional items made of related basic metrics. For example, we use attention and participation metrics, which are aggregates of a few basic metrics happening in the same time period that together show a user's attention to something or participation in an activity. In addition to motifs, we explored the use of such aggregated metrics for HOTS assessment.

To address the analysis aspect of the research problem, we used similarity analysis (Jaccard, 1912) to compare motifs and other time series generated from student activities to previously recorded expert data as a method of assessment. Jaccard originally created a (dis)similarity measurement function to perform analysis in phytology (Jaccard, 1912). This statistical function quantifies the similarity of two objects such as a text string, DNA sequence, audio/video file, or photographic images. Over time, similarity analysis has become very important for data mining and machine learning (Loh, 2016). We investigated different distance measurement methods for similarity and proposed appropriate ones for cases with larger or smaller data sets and for different HOTS. We also used clustering for time series of aggregated metrics to see if they could be used for HOTS assessment.

Our goal is to show that by using combined process metrics, we can develop assessment methods that do not rely on large amounts of data, can be flexible based on different curricula, and offer feedback on specific parts of students' performance. Overall, we investigated what educational activities can be defined to help with HOTS development, what process metrics can be collected for these activities, what analysis methods can be applied to them, and what principles can be followed to ensure "the good use of data". More specifically, we tried to answer the following research questions:

1. What process metrics can be used as indications of specific HOTS?
    a. How can VLEs implement educational activities based on learning theories such as experiential and situated learning?
    b. How can individual basic metrics be used to assess HOTS?
    c. How can combined process metrics be used as an aggregated HOTS assessment tool?
    d. How can small yet meaningful series of process metrics (motifs) be used for HOTS assessment?
2. What analysis methods can be applied to process metrics to provide insights on HOTS?
    a. How can a combination of time series of aggregated process metrics and clustering provide a way to predict learning and academic performance?
    b. How can scoring elements of a motif be used to assess HOTS?
    c. How can similarity analysis between student and expert motifs be used to assess HOTS?
    d. Which similarity indices are more effective in assessing HOTS?

In addition to these research questions, and through proper literature review, we aimed to establish principles on how data collection and analysis of learning process metrics can be done within a series of guidelines for "the good use of data." Our goal was to ensure our proposed methods follow these principles.

To achieve the above objectives, we performed an extensive literature review and four user studies. All studies followed the approved ethics protocols as required by Carleton University Research Ethics Board. Figure 1 provides an overall view of these individual studies and their related research questions. While Studies 1 to 4 responded to various research questions, our

critical literature review resulted in identifying a set of three principles for the good use of learning data. A summary of the four studies is shown in Figure 1.

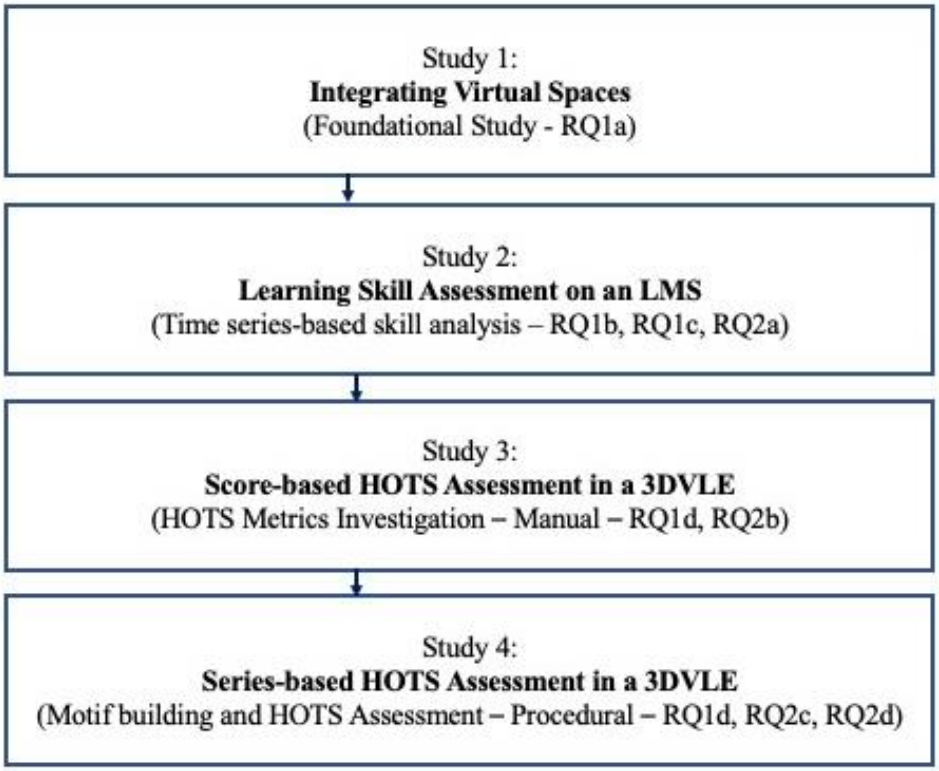


Figure 1. Research phases and studies

- **Study 1**: **Integrating Virtual Spaces with 21st Century Learning**

  This initial exploratory study helped answer Research Question (RQ) 1a. It investigated how learning theories can be used to guide space and task design for 3DVLEs. This study was our primary motivator to focus on HOTS development and assessment methods, as it demonstrated the gaps and potentials in the area. It was done in partnership with another PhD student in the School of Linguistics and Language Studies who acted as the educational domain expert. The course was chosen based on previous collaboration with the instructor on 3DVLEs and because it would allow us to create experiential and situational learning tasks.

- **Study 2: Learning Skill Assessment on an LMS**

  This study of text-based VLEs (LMS) answered RQ1b, RQ1c, and RQ 2a. It aimed to identify aggregated process metrics and the use of time series and clustering methods to predict student

success. The aggregated metrics were attention and participation. The LMS case was chosen to allow investigation of larger amount fo data and a 2D VLE.

- **Study 3: Score-based HOTS Assessment in a 3DVLE**

  This study answered RQ 1d and RQ 2b. It explored extended process metrics available in 3DVLEs and the use of motifs for HOTS assessment. Combined with video recording, the study was done in a generic 3DVLE used for language training. We manually defined multiple areas as motifs and associated them with various HOTS in a language course. The study was done in partnership with another PhD student in the School of Linguistics and Language Studies who acted as the educational domain expert. This study was a follow up study of the Study 1, going back to 3DVLEs, but using process metrics.

- **Study 4: Series-based HOTS Assessment in a 3DVLE**

  This study answered RQ1d, RQ2c, and RQ 2d. It examined full interaction data in a science lab implemented in a customized 3DVLE. The study used motifs by students and experts and compared a variety of similarity-based assessment methods. All metrics collection (metrics suggested by study 3) and motif creation were performed automatically with custom designed scripts. Assessment of motifs was also automated by employing custom scripts. We again compared the automated assessment with that of an instructor. The study was done in partnership with two other PhD students, one in the School of Information Technology at Carleton University and one in the Department of Chemistry and the University of Toronto, who acted as the 3D developer and educational domain expert, respectively. Having challenges in data collection in the previous studies, it was decided by the research team to create a special application that allows us to collect all the data we need. The content of the application and the subject of the course were changed from the previous study (language to chemistry) due to the availability of subject expert and suitability of 3D content for the new topic.

Upon completion of our studies, we reflected on our findings and what we had learned from the studies to develop initial insights into what a more comprehensive framework for HOTS assessment should be. This reflection (presented in Chapter 7) offers a preliminary design for a standalone HOTS assessment framework and identifies its recommended components, i.e., data collection, assessment, and visualization, to be an easy-to-use tool for instructors and course designers. This framework can be the subject of further research.

1.4 Contributions

The research described in this thesis resulted in the following contributions to the fields of 3DVE and HOTS assessment:

- Showing that basic and combined process metrics can be used to assess student performance
    - Study 2, RQ1 and RQ2
    - We also showed that using time series of these metrics and clustering them can help with assessment and be an indicator of student success, so they can also be used as a feedback mechanism.
    - This research suggests that VLE platform metrics can be grouped in a meaningful way to gain new insight into learners thinking skills.
    - **Publication:** Nowlan, N. S., Shafiq, M. O., & Arya, A. (2019), Are You Paying Attention? Assessing Student's Attention and Participation on Learning Management System., 11th International Conference on Education and New Learning Technologies (EDULEARN 19), pp. 8872-8881.
- Proposing combined metrics in the form of motifs (small time series of actions)
    - Study 3, RQ1 and RQ2
    - We showed how scoring motifs can correlate to the instructor's evaluation and as such, can be used for HOTS assessment.
    - This research suggests that 3DVLE platform metrics are more meaningful and provide HOTS insight when it is analyzed over a meaningful sequence of action
    - **Publication:** Nowlan, N.S., Hartwick, P., and 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
- Proposing motifs and similarity-based indices for HOTS assessment in 3DVE
    - Study 4, RQ1 and RQ2
    - We showed that different similarity indices might be appropriate for different HOTS
    - This research suggests that automated series similarity analysis can be used to assess learners' activity when it is compared to expert activity

- **Publication:** Qorbani, H.S., Arya, A., Nowlan, N. and Abdinejad, M., 2021, March. Simulation and Assessment of Safety Procedure in an Immersive Virtual Reality (IVR) Laboratory. In 2021 IEEE Conference on Virtual Reality and 3D User Interfaces, Special Session on Education.
- **In-Process Publication**: Nowlan. N.S., and Arya, A. (2022). Motif-based analysis of learning process metrics in 3D virtual environments for student performance assessment.

- Offering a better understanding of the relationship between educational theories and the design of educational activities in a 3DVLE
  - Study 1, RQ1
  - We provided design recommendations for educational activities and highlighted the need for HOTS assessment as a future direction.
  - **Publication:** Hartwick, P., & Nowlan, N. S. (2018). Integrating virtual spaces: Connecting affordances of 3D virtual learning environments to design for twenty-first-century learning. Integrating Multi-user Virtual Environments in Modern Classrooms (pp. 111-136). IGI Global.
- Identifying consent and transparency, use of non-proxy data, and direct benefit to the students as the main criteria and principles for "good use of learning data"
  - Literature review and theoretical analysis

1.5 Dissertation Outline

In Chapter 2, the existing literature on HOTS and assessment requirements are reviewed along with metric-based assessment in VLEs to identify the research gaps and establish a basis for the present research.

Chapters 3 to 6 report on the studies performed as part of this research. Each of these studies built on top of what was learned in the previous ones and added new elements in order to answer our research questions. These chapters all start with an overview of the study's goals, theoretical basis, and research questions. They then use the APA standard (https://www.scribbr.com/apa-style/methods-section) and explain the study design in terms of participants, material, and procedure (including measurements). Thi sis followed by experimental results and discussion.

Chapter 7 reflects on the findings of four studies, discusses the overall limitations and lessons learned, and offers initial insights on a comprehensive and integrated HOTS assessment framework.

Chapter 8 provides concluding remarks and recommendations for future research.

2 Related Works

In this chapter, we review the most relevant concepts and research work related to HOTS assessment in VLEs. We start with a general understanding of HOTS and their assessment, then review their development and assessment methods in VLEs. We conclude with a gap analysis that is the basis of our research questions and goals.

2.1 Defining and Developing Higher Order Thinking Skills

Schraw and Robinson (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. Anderson and Krathwold (2001) revised Bloom's (1956) well-known thinking skills taxonomy to argue that creating, evaluating, and analyzing skills are HOTS, whereas applying, understanding, and remembering are lower-order thinking skills. Resinick (1987) suggested that HOTS can be recognized in action when a person is situated in an unfamiliar situation to find meaning or structure.

The value of HOTS is well recognized at the local and international levels and is referred to within educational policy documents. Organizations and partnerships promoting essential 21st-century essential skills include the Partnership for 21st Century Learning (P21)[5] and the Global Partnership for New Pedagogies for Deep Learning (NLDP)[6]. These organizations highlight the importance of HOTS and list them as key skills required by 21st-century citizens, and note above mentioned challenges. The updated edition of Ontario's 21st Century Competencies[7] aims to provide focus and direction to Ontario school boards and mentions HOTS as must-have skills in the "The way we think" section (see Appendix A).

HOTS teaching methods are being developed by diverse groups of educators around the globe, tapping into a range of approaches and platforms (Pearlman, 2010). Due to their abstract nature, HOTS can be more challenging to teach than, for example, scientific information. In their

[5] https://education-reimagined.org/resources/partnership-for-21st-century-learning/
[6] https://deep-learning.global/
[7] http://www.ontariodirectors.ca/CODE-TLF/docs/tel/21_century_appendixC_only.pdf

book *Assessment of Higher Order Thinking Skills*, Schraw and Robinson (2011) identified five main challenges of teaching HOTS:

1. Defining and translating effective thinking into an educational curriculum.
2. Adapting teaching methods to individuals' unique dispositions.
3. Promoting a learning culture that fosters higher order thinking.
4. Transferring newly learned skills across different domains.
5. Identifying programs that are conducive to developing HOTS.

Many educational researchers have investigated why the traditional education system is failing to develop HOTS. Aldrich (2009) cited the well-known argument, "You cannot learn how to ride a bike from a textbook" (p. 3). In other words, competencies such as communication, creativity, and innovation require strong situational awareness and practice to be developed and cannot be acquired through reading or the listening of lectures that the traditional educational system favours.

Constructivist learning theory (Hein, 1991) posits that students construct their own knowledge and understanding of the world through experience and reflection on those experiences. Piaget, a prominent child development researcher, stated that learning is an internal process that involves experience and reflection (Smith, 1985). Bruner, who made significant contributions to human cognitive learning theory, similarly stated, "learning is an active process in which learners construct new ideas or concepts [through their experiences] based on their current and past knowledge" (Galindo, 2014; Bruner 1960). In his book *The Process of Education*, Bruner emphasized that students are active learners who construct their own knowledge; hence, the main purpose of the education should be to facilitate thinking and problem-solving skills, which will help students to become independent learners (Bruner, 1960; McLeod, 2008). Constructivist teaching strategies are intended to provide an environment (or a situation) where students can have the experiences required for knowledge construction to take place (McLeod, 2008).

Educational theorists have suggested that learning strategies informed by constructivist theory merit special consideration in the 21st century learning for their alignment with fostering HOTS (Fosnot, 2013). Lai (2008) connected the constructivist approach to strategies for developing 21st-century skills; specifically, constructivist learning strategies such as problem-based learning, cooperative learning, and formative assessment are methods recommended for developing 21st-century skills. As discussed in Section 2.3, researchers have suggested that virtual

learning environments offer the potential for implementing the guidelines from learning theories such as constructivism and experiential learning, to address some of the existing challenges in HOTS development. In our foundation study (Study 1, Chapter 3), we review key learning theories—behavioural, cognitive and constructivist—and highlight connections between 21st-century learning and constructivist theory, such as active learning, learning by doing (e.g., role play, scaffolding, problem-based), and collaborative learning. In this study, we found that VLEs (especially 3DVLEs) provide a good way of implementing constructivist learning strategies.

2.2 HOTS Assessment

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).

Schraw and Robinson (2011) argue that HOTS can be assessed through 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. An examiner watching and taking notes about the students collecting information from various sources, making decisions, and performing different actions based on new events (such as dealing with a fire or chemical spill in the lab) can be particularly time-consuming when performed for a whole class of students.

Computer-based assessment, with its capacity to capture rich information about students' learning process, may help educators with this task. Students' answers to a traditional knowledge assessment question only provide information on whether they know the answer, whereas computer-based assessment can provide information about why and how students came up with their answers and how they can correct them if they are incorrect. Emerging technologies such as VLE/3DVLE enable such practices through offering simulated space and digital tracking

capability (Borgman et al., 2008; Dede, 2009; Warbuton, 2009). We discuss some strengths and gaps in the use of VLEs for HOTS development and assessment in the following two sections.

2.3 HOTS Development in Virtual Learning Environments

In this section, we briefly review VLEs (particularly 3DVLEs) and serious/educational games as a typical application of 3D virtual environments. We then review common HOTS development approaches in VLEs.

2.3.1 Virtual Learning Environments

Broadly speaking, a virtual learning environment (VLE), including the commonly used *learning management system*, stores educational documents and facilitates teachers' and students' communication and assessments (Jenkins, Browne, & Walker, 2005). 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, 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.

Duncan et al. (2012) concluded the following: (1) the use of VLEs for supporting education is widespread and increasing; (2) there are numerous studies on VLE usage in education; and (3) 3DVLEs are mainly used for collaborative or simulation-based education. While non-3D VLEs with text and 2D content are now a standard part of university education, universities and other educational organizations have been experimenting with 3DVLEs over the past decade. The hardware and software advances caused a jump in interest around 2010, as evidenced by the popularity of tools such as Second Life (Warburton, 2009; Schmidt and Stewart, 2009). Despite this interest, 3DVLEs have not been widely used to support education, and interest has slowed significantly since it peaked in 2012 (Reisoglu, 2017). Lower prices and technical improvements made immersive Virtual Reality (VR) systems more popular in the last couple of years, which resulted in a new surge in the use of generic 3D virtual environments such as Mozilla Hubs[8] and

[8] https://hubs.mozilla.com

Virbela[9] for educational purposes (Scavarelli et al., 2019) as well as specialized 3DVLEs such as LearnBrite[10] and Labster[11]. The lack of customization, the need for technical skills to use and set up, limited interactivity and accessibility, and the lack of assessment tools are among the shortcomings of these tools (Scavarelli et al., 2019). These shortcomings together with missing advanced features such as realism and intelligent environments and agents can be identified among the reasons for the slow adoption of 3DVLEs in educational contexts (Arya et al., 2011).

2.3.2 HOTS Development in 3DVLEs

Spires (2008) noted that today's students were born into social and educational environments where digital technologies are pervasive; as such, they bring different skills, needs, and interests to the classroom compared to previous generations. They need to develop knowledge and skills in a complex and ambiguous problem-solving landscape as well as know-how to collate multiple information sources in multiple collaboration streams (Spires, 2008).

Spires (2008) mentioned two model projects that used a 3DVLE platform: Dede's (1990) River City project, where students investigated ecological disorders, and Lester et al.'s (2007) Crystal Island project, where students solved a biological mystery, such as a multi-legged frog, through environmental investigation and chemical lab testing.

Dede and his colleagues developed River City (1990) with the claim that HOTS are best developed when:

- Learners construct knowledge rather than passively ingest information.
- Sophisticated information-gathering tools are used to stimulate the learner to focus on testing hypotheses.
- There is a collaborative interaction with peers, similar to team-based approaches underlying today's science teams; and
- Evaluation systems are used to measure complex higher order skills, rather than the simple recollection of facts.

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

[9] https://www.virbela.com
[10] https://www2.learnbrite.com
[11] https://www.labster.com

learning. Dede (2007) argued, “Fortunately, emerging information communication technology that enable immersive, collaborative simulation now offer the capability to implement situated learning environments in classroom settings” (p. 19). The central notion of active, experiential, and situated learning is that learning is an active process whereby cognition is shared, and learning is achieved through collaboration and co-participation that fosters higher-order thinking skills (Dede, 2007).

Kelman (1989) identified 3DVLEs as a potential environment to foster HOTS development. In a study by Hopson, Simms, and Knezek (2001), students’ self-reports indicated that the 3DVLE resulted in increased motivation, creative tendencies, inclination towards exploring the unknown, perseverance, and taking individual initiatives. As noted in a highly cited article by Roschelle, Pea, Hoadley, Gordin, and Means (2000), “Although active constructivist learning can be integrated in 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).

Affordances are described by Molka-Danielsen et al. (2012) as “...a quality of an environment which allows an individual to perform an action” (p. 3). Noting this, researchers have identified three important affordances that distinguish VR and 3DVLEs from other virtual platforms: (1) 3D visualization of physical space (realistic or not) ; (2) immersion (full engagement), related to the feeling of presence and embodiment; and (3) synchronous interaction and mediation through tools and collaboration with others (Steffen et al., 2019).

In a well-designed game, players need to apply many HOTS to achieve the objective of the game and be successful (Shute, 2014). Therefore, a good game offers efficiency to players to develop HOTS in an engaging activity. VLEs that use game elements such as ongoing scoring feedback, competing against other players, or the game itself, can be referred to as serious educational games (SEGs; Ciftci, 2018). By engaging in activities that demand high focus and decision-making, gamers can improve many core transferable skills that overlap with HOTS (Green, 2010). For example, multiplayer games in virtual environments can force gamers to use skills such as teamwork, social connectedness, decision-making, planning, and resourcefulness (Galarneau, 2005). Further, a common gaming element, such as providing immediate feedback to the player that fosters engagement, can offer alternative suggestions to traditional evaluation processes (Belotti, 2013). Thus, in our studies, we used gamified elements such as gates that open when the correct answer is provided (e.g., Study 3).

Research on how VR affordances can be used to implement learning and educational strategies is very limited (Dede, 2007, Scavarelli et al., 2020). Our Study 1 aimed to address this shortcoming through an exploratory investigation.

2.4 HOTS Assessment Methods in Virtual Learning Environments

Azevedo (2005) pointed out that any educational technology's effectiveness is limited if a deep assessment is not performed. In terms of assessment methodology, most studies on the effectiveness of computer-based learning tools and serious educational game measure learning gains via a pre-test/post-test design in line with traditional assessment systems where learners are under the observation of a teacher or facilitator (Azevedo, 2005, Smith, 2015).

In a meta-analysis published by Smith (2015), researchers identified categories of data collection to validate the learning effectiveness of VLEs and SEGs, including: (i) focus groups, (ii) questionnaires, (iii) direct observation in the field, (iv) direct observation in a controlled environment, and (v) indirect observation (Smith, 2015), as shown in Figure 2. Smith's investigation suggested that, out of 510 data collection techniques used in 299 studies, only 10 techniques used field data that was collected during activities. As a result, Smith stated that current data collection strategies are more aligned with traditional educational disciplines, such as assessing knowledge gains, and missing the opportunity to collect and follow students in learning activities (i.e., in situ collection) to identify areas for development and thus increase the efficiency of the learning activity (Smith, 2015). These strategies are called "black box" approaches (Loh, 2015), illustrated in Figure 3.

On the other hand, most 3DVLEs use original game engines, such as Unreal (https://www.unrealengine.com) or Unity (https://unity.com/). As a result, many 3DVLEs support gaming functions such as building narratives, setting the stage, assigning roles and characters to players, managing game rounds, submitting and accessing documents, text, and voice chat functions, and creating belief with maps and graphics, along with collecting metrics on user positions and every interaction. Therefore, gaming platforms offer the built-in potential for collecting these metrics. Non-3D VLEs such as LMSs also offer limited levels of activity data such as files accessed or objects viewed. Alonso-Fernandez et al. (2019) noted that log data such as simple values from interactions (e.g., completion times and scores) and more complex information such as the type of failures or exploration strategies are important to assess the effectiveness of

learning methods. Griffin and Care also highly recommended the collection of process (field) data to ensure that as much educationally relevant information as possible is captured (Griffin & Care, 2014).

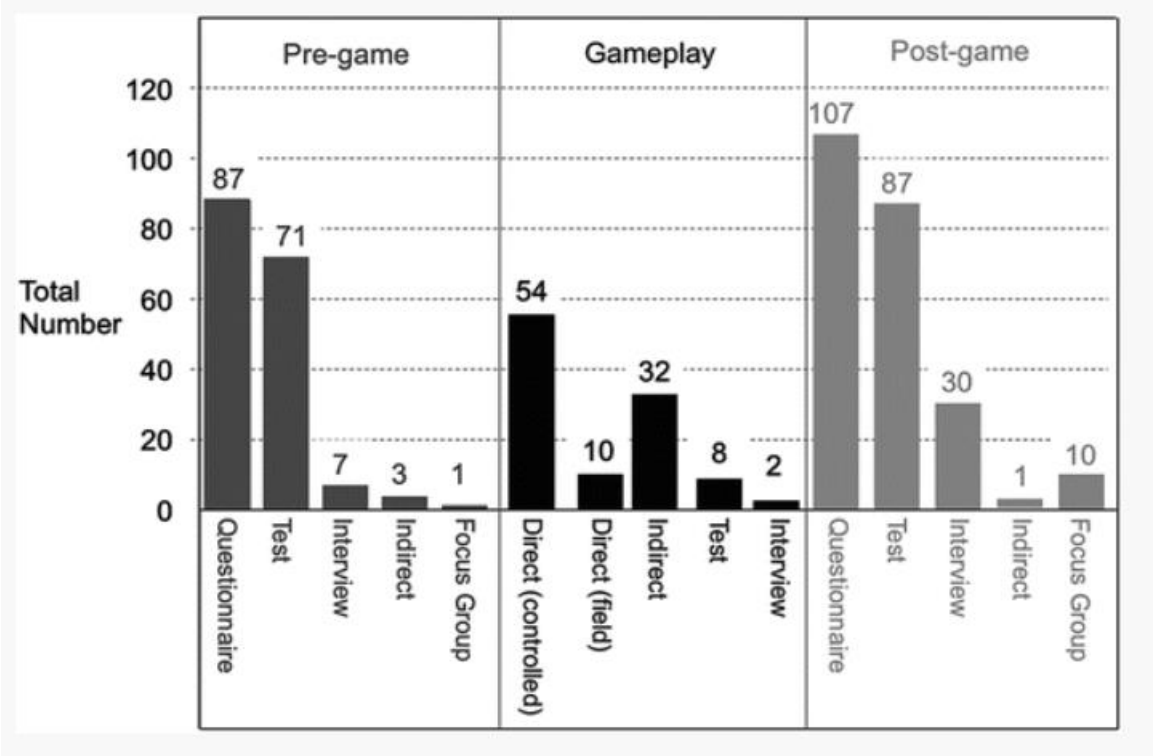


Figure 2. Number of specific data collection techniques used per phase of study (Smith, 2015)

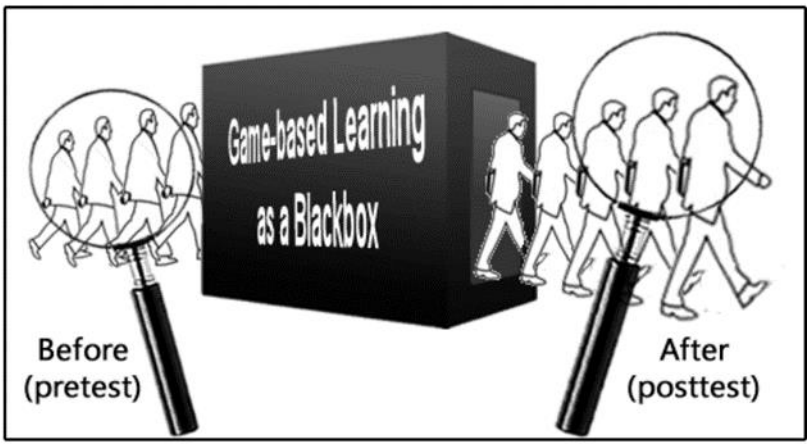


Figure 3. Black box testing (Loh, 2015)

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. It is not intended to be "hidden" from students as ethical concerns require consent, but the assumption is that the student work follows its normal flow and observations are in the background and not affecting the performance. SA is a relatively new concept and researchers are working on identifying metrics and methods to perform this form of assessment (Qian, Clark, 2016). SA approaches in existing literature can be organized into the following two categories:

1- Score-based *Stealth Assessment*, where each interaction/activity is scored based on its weight (importance) in the learning task; and
2- Series-based Stealth Assessment, where interactions are assessed as a time series.

In the following two sections, we provide more information on both processes' metric-based methods, specifically for assessing HOTS, although as the metric-based assessment field matures, researchers aggregate the results from multiple assessment methods to increase the accuracy.

2.4.1 Score-Based Stealth Assessment

Score-based stealth assessment methods are used when learners' interactions with the VLE are scored with learning activity compatible scoring logic, 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.

As mentioned earlier (Section 1.3), we refer to metrics collected for a single activity/interaction as *basic metrics*. Scoring-based assessment approaches usually use these basic metrics, as opposed to more complex and combined ones, due to the simplicity of scoring, in the following ways:

- Basic metrics can be used with equal weight and scored equally.
- Basic metrics can be weighted as their type and/or location and scored differently.
- A script can be applied on interactions to derive process-related information.

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 provide them with support when needed, thus helping them to improve their meta-learning skills. In their study of the learning processes of 11-, 13-, and 14-year-olds, Veenman et al. documented those player actions they believed to represent players' thinking processes (see

Table 1). All logged actions were then automatically scored as learning. For example, (Number.exp) reflected the skill of drawing conclusions and was recorded as a positive indicator of meta-learning skillfulness, where more experiments indicated more complete learning. Thinking time between receiving the outcome of a former experiment and taking action was registered as a positive indicator of outcome evaluation, reorientation, and planning for the next experiment. Correlation results showed a stronger relationship between metacognitive skills traced through user action captured in log files than 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.

Table 1. Player action log entries and metrics (Veenman et al., 2014)

Label	Description
Logfiles measures	
Number.exp	Total number of experiments performed.
Thinktime	Time in sec. elapsed between receiving the outcome of a former experiment and the first move in the next experiment, accumulated over the experiments.
Scrolldown	Frequency of scrolling down to earlier experiments.
Scrollup	Frequency of scrolling back to later experiments.
Votat.pos	Number of transitions between experiments in which only one variable is altered.
Votat.neg	Mean number of variables changed in transition between experiments, minus one.
Unique.exp	Number of unique experiments performed out of 48 possible unique experiments.
Variation	Mean proportion of investigating the different levels of the five independent variables.
Trace measures	
Systematic	Score (0–4), rated from traces of learner activities, indicating the extent to which a participant revealed (idiosyncratic) systematical patterns of experimentation.
Complete	Score (0–4), rated from traces of learner activities, indicating to which extent all variables were equally varied over experiments, both singularly as well as in combination with other variables.

A medical application of score-based stealth assessment was reported by Weiss et al. (2013). A medical simulator called AngioMentor mimics a modern angiographic suite and provides realistic feedback in carotid artery stenting. A total of 33 interventional cardiologists

(eight novice, 15 intermediate, and 10 experienced) completed 82 simulated procedures. Experienced interventional cardiologists generated a scoring algorithm that classified surgeons as novice, intermediate, or experienced practitioners. The algorithm used basic metrics collected by the simulator and calculated content/process awareness metrics in three broad categories: technical performance, medical management, and angiographic results. Each metric was assigned a score to represent the metrics' relative clinical importance. The objective was to use basic metrics collected by the simulator software to calculate the metrics (process metrics) used for the assessment. Some of the calculated assessment metrics included:

- Crossing lesion with 0.035-in. wire before filter placement (per incident)
- Crossing lesion with catheter before filter placement (per incident)
- Filter undersized to the vessel landing zone
- Predilation balloon length shorter than part of lesion that is >80%

The results of the study suggested that simulator-collected basic metrics can be used to create process metrics, which can be used to validate surgeons' skills and determine their level of proficiency (Weiss 2013).

Similarly, 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. Figure 4 shows three tiers of metrics, some provided by the NeuroTouch simulator as basic metrics, and some derived from the basic metrics.

Although scoring different type 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. For example, while performing an experiment at one point in the curriculum might be evidence of understanding, it might be totally redundant to another point. To include positional information in each interaction scoring, we can create both task (context of interaction) and the interaction combined into one data item, such as "opening a box during task x", rather than just "opening a box" and treat this differently than "opening a box during task y". However, that means the number and combinations of data items will increase.

Figure 4. Organization of metrics used in the study by Azarnoush et al. (2015)

This increase in data, which can also be caused by a large number of students and tasks, has motivated researchers to investigate Machine Learning (ML; Jordan & Mitchel, 2015) methods to study students' learning task interactions. Jordan and Mitchel (2015) note that Machine learning is a data-intensive branch of Artificial Intelligence (AI) that addresses the challenge of improving computers automatically. Thanks to new learning algorithms such as Artificial Neural Networks (ANNs) and the availability of large amounts of data, machine learning, and its applications have grown rapidly in the last few years in many areas, from health care and education to financial modeling and marketing.

2.4.1.1 Using Machine Models to Facilitate Score-Based Stealth Assessment

Scores that each interaction receives can be analyzed based on a rubric and aggregated, or each interaction along with its weighted score can be fed into a machine learning algorithm and a model can be trained to group students. This approach was used by Sabourin et al.'s (2013) study of Crystal Island, a 3D virtual island where learners investigate, collect information, and analyze data using an experimental testing device to solve a biological mystery, as described in Rowe et al. (2009). Researchers found that students in the high group of Self Regulated Learning (SRL)

also had the highest knowledge gains. Investigating the correlation between students' in-game interactions and their SRL skills, Sabourin et al. (2013) used 10-fold cross-validation and 49 selected features (basic metrics) to train machine algorithms that successfully profiled students into self-regulated groupings (low, medium, high).

Interestingly, although highly self-regulated learners were not more likely to solve the mystery, they did demonstrate significantly higher learning gains based on the comparison of before and after curriculum tests that were taken (Sabourin et al., 2013). This finding raised concerns about performance evaluation based on result actions, such as solving a mystery and answering a question.

Another machine learning training model was reported by Shute and Kim (2014) for the assessment of problem-solving skills. A Bayesian Network (BN) is a type of probabilistic graphical model that can represent a multitude of relationships between a set of variables in a system (Loh, 2015). Shute and Kim (2014) helped to further explain BNs with their study of a 3D game-based stealth assessment based on evidence-centered assessment design (ECD). According to the ECD framework, the first model to build is a competency model, which determines the competencies on which learners will be assessed. The second model is an evidence model, which consists of (1) evidence rules that convert work products to observable variables, and (2) a statistical model that defines the statistical relationships between observable and competency variables. According to Shute and Kim (2014), observable variables (interactions) provide evidence relative to a student's competencies (such as problem-solving skills). However, to gain insight from this evidence, interactions along with the point in the game that it happens (evidence) should first be mapped to related competencies, as shown in Figure 5.

Once interactions' values are defined, Shute & Kim and collaborators decide how to score them and establishes the relationship between each interaction and the associated levels of competencies. After that, the BN is constructed to accumulate the data. The BN graphically demonstrates the conditional dependencies between different variables in the network. It is composed of both competency model variables (e.g., problem-solving and its four facets) and associated observables that are statistically linked to these competencies. A separate BN needs to be constructed for each level because the observables change across levels.

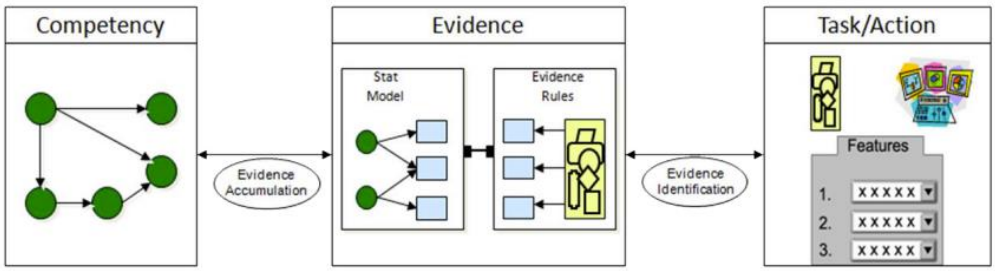


Figure 5. Three main models of ECD (from Mislevy, Steinberg, & Almond, 2003)

It should be noted again that the BN method does not only investigate type of interaction but also the location of the interaction to identify the competency value of the observables. Figure 6 shows Shute's Bayesian Network; at the top there is Problem solving skills, second layer shows the sub component of these skills, finally the lower yellow level shows the individual observables related with the connected sub skill component. The Figure 6 shows the statistical relationship between two indicators (observables), I12 and I37, in Shute and Kim's (2014) study. Shute and Kim compared their findings from the above-described stealth analysis with Greiff and Funke's (MicroDYN, 2009) results of a complex problem-solving assessment, a real-world performance-based assessment approach, and found strong correlations.

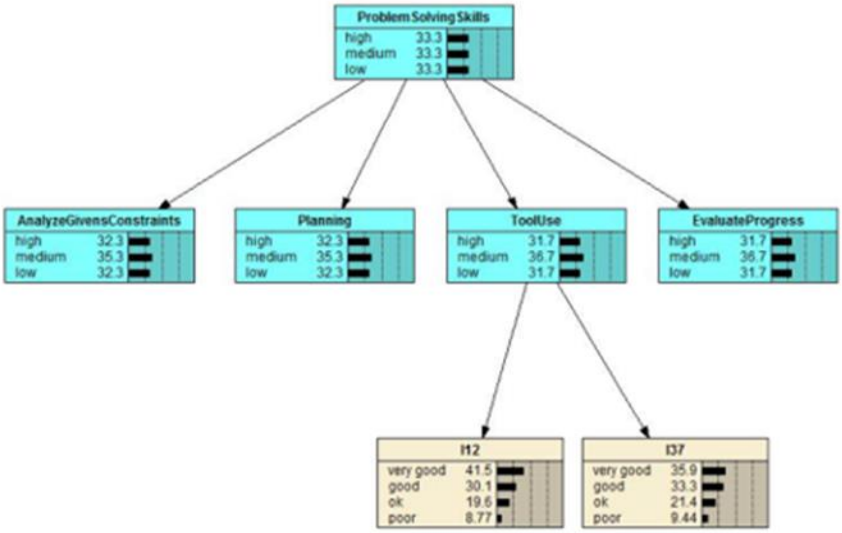


Figure 6. Bayesian Network of problem-solving probabilities (Shute, 2011)

After Sabourin's (2013) initial study (summarized at the beginning of this section), she then used a static and dynamic BN approach to predict students' SRL using the same data from

the previous machine learning study. The static BN represented the components of the process but in a static way, while the dynamic BN incorporated time to represent the cyclical nature of SRL processes. For each model created for this analysis, Sabourin (2013) explored whether the additional, position-based interaction diversity offered additional predictive power. To train the different machine algorithms, Sabourin used three different feature sets: All Features, totaling 55 features before feature selection is applied; Feature Selection, where 49 attributes and occurrence features were identified during feature selection; and Feature Selection + Event Feature Creation, i.e., contingency and patterned contingency features that were created through the differential sequence mining approach. Sabourin's results indicated that the BN significantly outperformed both baseline measures and the naïve classifiers (Figure 7). This result points to the importance of analyzing the interactions (observables) based on position and meaning.

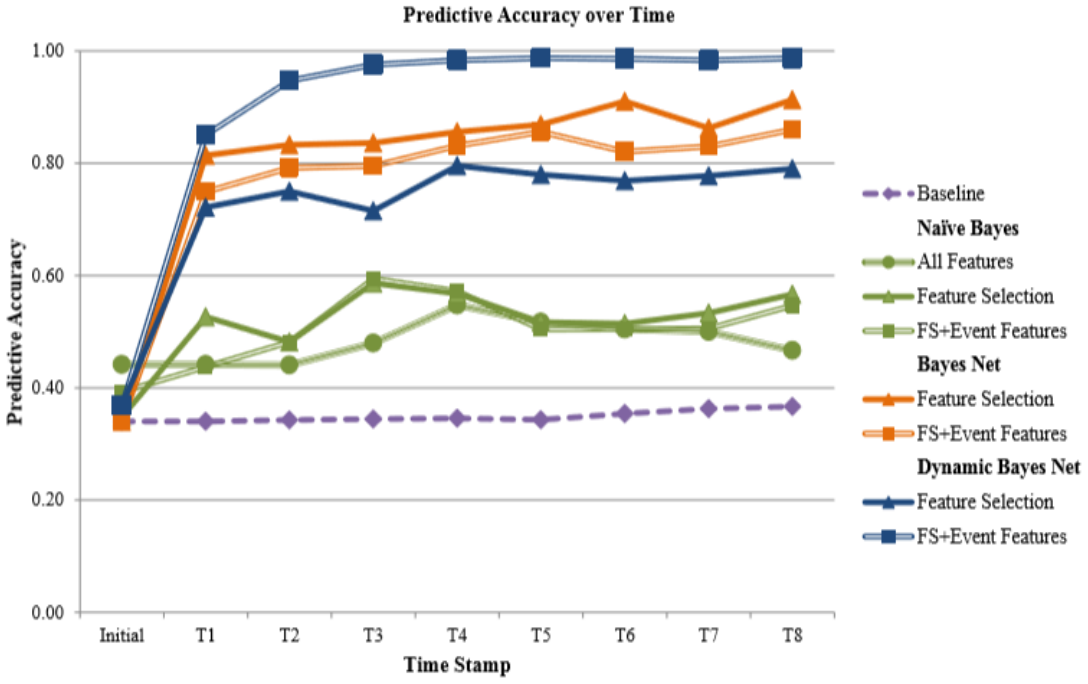


Figure 7. Predictive accuracy of metacognition over time (Sabourin, 2013)

There are some limitations to the BN-based approach. First, it requires 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. Second, it is a challenge to assign a practical, real-life meaning to the BN and many other machine learning (especially deep learning) results and interpret their probabilities as actionable insights that can

improve performance (Conati et al., 2018; Loh, 2015). In other words, although this type of prediction approach can identify learners who might perform low, they do not provide an area of focus to improve their future success.

The field of Intelligent Tutoring Systems (ITS) has strong connection with machine models and the decision-making process to provide guidance, and thanks to research in the field, successful techniques have been developed based on user models to support students in learning activities (Freedman et al., 2000; Conati et al., 2018). Conati et al. (2018) argued that it is critical for these systems to explain their decisions to learners to gain trust and identify students' weaknesses for further development.

As stated by researchers such as Loh and Conati, providing explainable assessment is critical and interpreting any machine learning-based performance assessment has limitations in the educational world in regard to interpreting the probabilities as valuable insights for the learner for improvements. Any assessment result should be explainable to learners so they can understand their areas of improvement, and this lack of explainability and interpretation is a major issue with machine learning-based approaches. Another downside of machine learning-based assessment, as seen in all examples, is the need to have huge amounts of data to train the machine model to reach acceptable accuracy.

To make the assessment results more explainable, researchers have considered data patterns instead of single metrics (Baker and Clarke-Midura, 2013; Gibson and de Freitas, 2016). A notable investigation of students' inquiry skills assessment based on collected metrics was performed by Baker and Clarke-Midura (2013). They investigated data collected from 52,000 students from 1,985 middle schools over a 5-year period on Rowe' Crystal Island, a 3DVLE that hosted an inquiry-based biology mystery. The collected log files were further distilled for analysis, producing a set of 48 semantically meaningful features or metrics (see Baker and Clarke, 2013, for list). Baker' and Clarke's study showed that time spent in stages of an activity or actions performed can be used to predict learners' successful response only if it is interpreted along with the process. It is not simply any time spent, or any activity performed that can give the end result performance indication, this information has to be used and make sense within the activity.

Log files collected from Baker and Clarke's (2013) Crystal Island biology learning activities were later used by a team of two researchers, Gibson and de Freitas (2016), for further data analysis. In their study, Gibson and de Freitas used data mining methods (clustering, machine

learning, symbolic regression, etc.) along with causal explanations to identify meaningful patterns for metric usage for automated performance assessment. Their study demonstrated that transforming data for data mining involves both reductions moves and intermediate pattern aggregations. In particular, 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. Analysis methods such as clustering, or machine learning could only be applied only after the motifs were identified and related metrics were used. Despite these initial studies, there is very limited research on the design and use of motifs and other combined process metrics, especially for HOTS. Motifs link score-based assessment to series-based assessment as an alternative, discussed in Section 2.4.2.

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) argued that logged metric data can be used to train a machine learning algorithm to assess students "on the fly" and provide feedback. Floryan et al.'s study used Rashi (Figure 8), a simulation system used in biology for teaching symptom-based medical diagnosis. Students using the system can get information about the patient being investigated through voice, text, and video-based information files, and through asking questions and performing tests. Their task is to diagnose the patient through investigation.

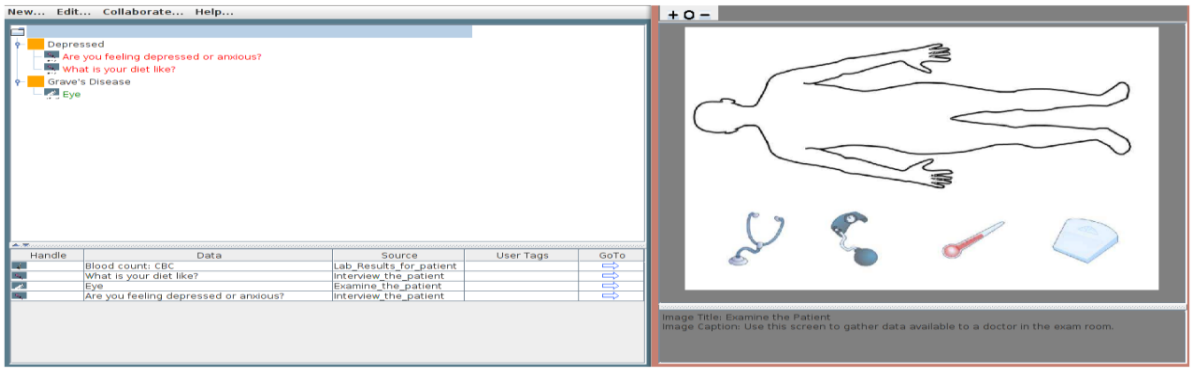


Figure 8. A screenshot from the subset of the Rashi system (Floryan et al., 2015)

Floryan and the research team used (1) basic features and (2) expert knowledge-based (EKB) features to train the machine learning algorithm. EKB metrics in this system were created

by a simple text matching scheme to detect when students were utilizing biology concepts that experts have agreed were related to the case. After training the algorithm, the researchers compared the performance assessments by subject matter experts and by machines in real-time while watching students use the system (Figure 9).

BASIC FEATURES

Feature Name	*Description*
Num Hypos	Number of hypotheses the student considered
Num Rels	Number of relationships the student created
Num Data	Number of data observations collected by the student
Relations Per Hypothesis	Number of relations divided by num. of hypotheses
Percent Exam	Percent of data collected through physical examination of the patient
Percent Interview	Percent of data collected by interviewing the patient
Percent Lab	Percent of data collected by running lab tests
Time Examination	Percent along time scale in which student primarily performed physical examinations
Time Interview	Percent along time scale in which student primarily interviewed the patient
Time Lab	Percent along time scale in which student primarily requested laboratory tests

EKB FEATURES

Feature Name	*Description*
Hypo. Match Score	How well student hypotheses match nodes in the ekb
Rel. Match Score	Score for how well student relationships match the ekb
Num. Data Associated	Amount of data collected related to a hypothesis but not connected by student
Lab. Justification Matches	How well justifications for ordering lab tests match to the ekb
Scratch Pad Matches	How well contents of student scratch pad match the ekb

Figure 9. Basic features and EKB features used to train the machine algorithm (Floryan et al., 2015)

A close correlation between the machine assisted model and the human observation model on two different diagnosis scenarios in Floryan at al.' study is shown in Figure 10.

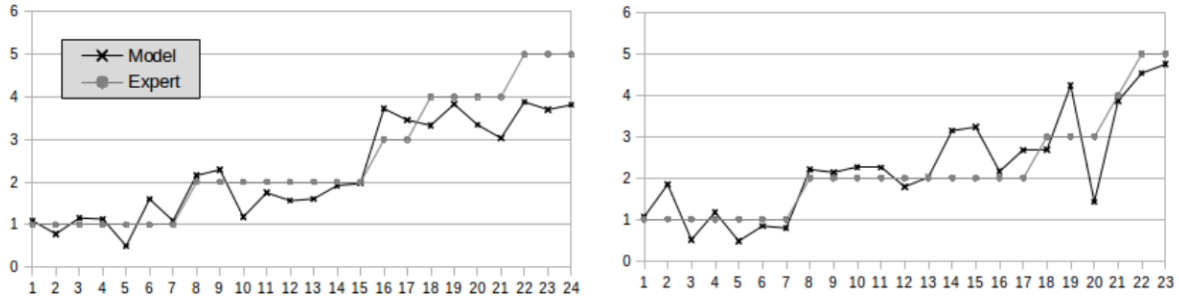


Figure 10. Example assessments (two cases) given by the machine assisted model and human expert observer (Floryan et al., 2015)

Floryan et al. (2015) provided supporting evidence that using both basic metrics and scripted calculated-interpreted metrics, such as EKB metrics, increased the accuracy of the

assessment. Based on their study, one can suggest that the use of expert knowledge can be helpful for HOTS assessment, as well.

2.4.1.2 Score-Based Performance Investigations on Learning Management Systems (LMS)

The widespread adoption of learning management systems (LMS) has provided a way to collect a wealth of metrics for educational organizations to mine and use to develop teaching and learning strategies and advise and allocate resources (Brown, 2011). LMS are also widely used as a platform for delivering and assessing exams, thus offering further means to collect student performance information and customize digital learning content.

Moodle is a widely used LMS platform. It is an open-source platform that can collect a wealth of metrics on students' activity. A study by Romero and his colleagues (Romero, López, Luna, & Ventura, 2013) provides a good example of this line of study, where the relationship between students' online discussion forum activities and their pass/fail performance in the course was investigated using Moodle collected data.

A group of researchers using K-means clustering algorithm reported that general lack of activity is the main characteristic of students with the tendency to fail (Hussain, Hussain, Zhang, Zhu, Theodorou, & Abidi, 2018). Mwalumbwe and Mtebe (2017) developed a tool to predict students' performance using LMS data. Their study found supporting evidence that peer interaction was the most significant factor in predicting students' performance. Carezo, Esteban, Santillan, and Nunez (2017) studied students' procrastination behaviors in computer-based learning environments using a predictive apriori algorithm (Scheffer, 2001). The study highlighted the importance of time management in students' performance.

Overall, these studies show that despite the limited data compared to 3DVLEs, text-based learning management systems can also offer opportunities for collecting metrics and performing assessment. This suggests that HOTS assessment can be done in LMS environments, provided we can define proper process metrics associated with HOTS.

2.4.1.3 Summary of Score-based Assessment

To summarize, score-based stealth assessment uses individual interactions (basic metrics, in some cases with different weights), along with coding these interactions differently based on their location to perform the performance assessment. Treating each category of basic metrics regardless of their location being activated presents a meaning problem, as opening a door is not the same in task A versus task B. Alternatively, creating a location-based interpretation makes the

rubric costly, timewise, and not flexible. Although instructor-created rubric/script methods of assessment are also being used by the researchers, the trend of creating machine models by feeding all the collected data is becoming popular. However, machine model assessments provide a single overall result. Thus, it is difficult to provide actionable feedback to students on identified strengths and weaknesses and requires trained data to be collected before the curriculum is used in the classroom. Analysis of partial actions is necessary for detailed feedback, which is somewhat addressed by series-based methods, and our proposed use of motifs tries to offer a more effective solution still.

Additionally, score-based assessments calculate the volume of the metrics in use and make judgments based on volume, not the order in which they are performed. If the number of data points is too big, machine learning models are used. Machine learning methods can also include pattern matching and the same action can be counted differently based on its location (Shute, 2011), which also leads to series-based assessment, as discussed in the next section.

Other challenges of using a score-based assessment approach include the lack of flexibility and the amount of training data. Once a performance script is tested through manual assessment correlation, or an AI model is trained, the curriculum should remain loyal to this version for the assessment to be accurate. This is due to the fact that patterns of action are not easily defined by instructors.

2.4.2 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, 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 new metrics to assess learners' meta-skills in order to help game designers assess, provide individual support, and foster metacognitive skills development. 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. If students were observed not performing well on tasks,

they were asked if they would like to receive tutorial help. If they accepted help, a pop-up tutorial message gave reminders about their status in the task and strategies they could use to move forward.

Snow et al. (2015) observed that students who chose to receive help via tutorial messages performed significantly better compared to students who did not request help. To arrive at this finding, the researchers analyzed log data and defined metrics for students' meta-skill development, where the assumption was that students who performed better would demonstrate higher metacognitive skills as demonstrated by their interaction in the gaming user interface. Three distinct types of stability methodologies—random walks, entropy analysis, and Hurst exponents—were used within iSTART in the form of stealth assessment. These analyses provided the researchers a means of unobtrusively assessing how students behaved and learned within the environments.

Three stability components calculated on students' interaction series revealed the following insights:

- *Random walk*: High-ability students tended to gravitate more towards identifying mini games where they could freely identify learning strategies, whereas low-ability students interacted most frequently with generative practice games that provided more guidance.
- *Entropy*: Students who engaged with the game in more controlled and strategic ways had lower entropy scores.
- *Hurst exponent*: Students who demonstrated more controlled and strategic action in the game activities scored higher (0.5 on the scale of 5).

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, and 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 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). There are different methods of statistically measuring similarity/dissimilarity between two data points. As a result, a similarity measure (similarity index) might be different based on the applied methods of calculation. Loh and Sheng's (2014) research provided supporting evidence that the Jaccard similarity index (or JACC; Jaccard, 1912) is the best metric for discriminating expert players from novices based on their action sequences (see Chapter 6 for a full discussion of similarity index-based performance assessment). However, in a serious gameplay field, there might be more than one expert and different paths to perform a task; for this, Loh and Sheng (2014, 2015) suggested using multiple similarity indices for different experts and using the maximum similarity index (MSI) to identify specific players' expertise level. Loh and Sheng's suggestion provides a practical metric for 3D platforms to be created such that a learner-user's actions can be measured against an expert's (or multiple experts) following different paths. This approach offers a single metric and easy-to-measure practicality that is not offered by score-based approaches.

While a similarity metrics-based performance assessment approach has its advantages, identifying the components of HOTS is still a challenge, as it only provides an overall performance assessment rather than a performance component. It is important to identify students' skill profile dimensions and competency levels in order to provide the required support when and where it is needed most. The similarity metric-based approach does not analyze or identify the source of the lack of the skill that causes a learner's performance to be not as good as that of an expert.

A recently published paper that used series-based analysis in assessing performance assessment is Peffer et al. (2019). Peffer and colleagues focused on students' scientific practices, such as inquiry, where 106 undergraduate students' activities were recorded on Invasion of the Grackles SCI, an online simulation platform. In a virtual biology lab, students were given complete autonomy as to the generation of their hypothesis (H), how many and which tests (T) they performed and in what order they chose to perform them in, and whether to seek outside information (I), either through the in-simulation library or via the Internet. Students had autonomy to decide when they were finished and generate their final conclusion (C) as a result of their scientific inquiry based on the evidence they collected through their tests. Students' click stream data (virtual biology lab interaction points) was collected and merged to generate a string of actions performed by each user. Peffer et al. then used clustering analysis (discussed more in Study 2,

Chapter 4), which is a statistical method that groups similar data points and identifies groups among them (Cavalli-Sforza, L.L., 1965). Using clustering analysis, Peffer et al. found that non-science majors who performed simple investigations (smaller number of steps) tended to cluster together and biology majors who performed complex investigations also tended to cluster together. The researchers concluded that a clickstream analysis of inquiry practices is sufficient for distinguishing different groups of novices, namely non-science and biology majors (Figure 11).

	Features	Cluster			
		A	B	C	D
Centroid	Maximum of Repeated T	1.9	2.0	1.9	5.3
	Maximum of Repeated I	0.5	3.1	9.9	0.3
	Transitions After Initial H	2.3	4.7	2.4	2.4
# of	Biology Major	10	8	5	19
	Non-STEM Major	30	12	11	10
	Simple Investigation	26	8	6	2
	Complex Investigation	12	8	9	24

Figure 11. Clusters derived from student's clickstream (Peffer, Quigley, & Mostowfi, 2019)

Applying clustering methods to categorize users based on their interaction is not necessarily a performance assessment. However, this approach provides insights into the groups of different academic investigation skill profiles and along with the group an individual student belongs to.

In a similar study, Sawyer, Rowe, Azevedo, and Lester (2018) used Crystal Island (an immersive virtual biology learning system), as a learning platform to record students' cumulative actions (e.g., conversations with virtual characters and reading books) and create filtered time series for each student for assessment. Figure 12 shows the filtering process from action sequences to filtered time series.

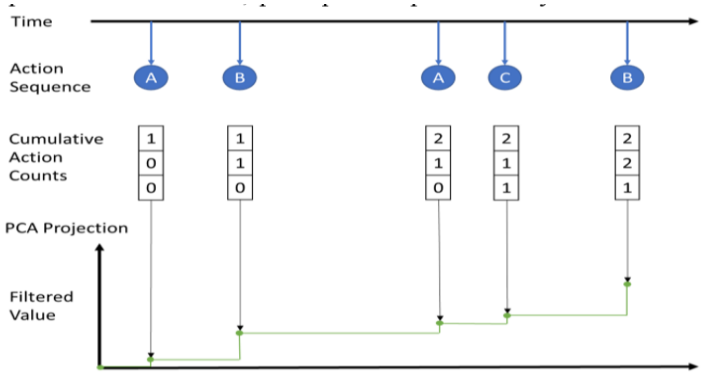


Figure 12. Filtering process from action sequence to time series (Sawyer et al., 2018)

The filtering process shown in Figure 12 assumes three actions (A, B, and C) and a sequence showing the number of occurrences for each. It takes action A as principle action and starts the sequence as 1 0 0 (i.e., action A has happened). When action B happens, the sequence becomes 1 1 0 and so on. When the action A happens again, the sequence starts with 2. For comparison data, students' before and after test scores were used as a normalized learning gain (NLG). It should be noted that students' game scores showed a high correlation with students' knowledge gains measured by before and after knowledge tests. Figure 13 shows the trajectory of students' time series along with the golden path series that belongs to an expert, whereas students' trajectories move away from the golden path, their learning gain lowers. Sawyer et al. (2018) suggested that comparing students' problem-solving path to an expert's problem-solving path might provide strong insight into students' problem-solving skills.

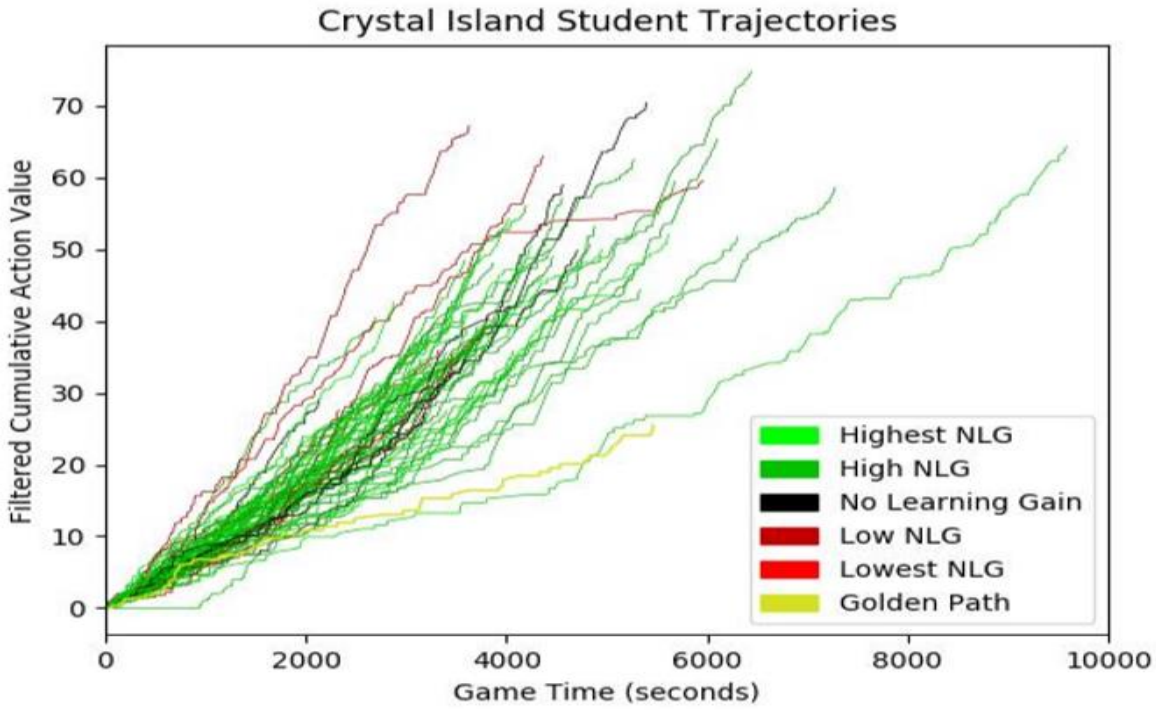


Figure 13. Expert Solution Path in Crystal Island (Sawyer et al., 2018)

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 (Figure 14). 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.

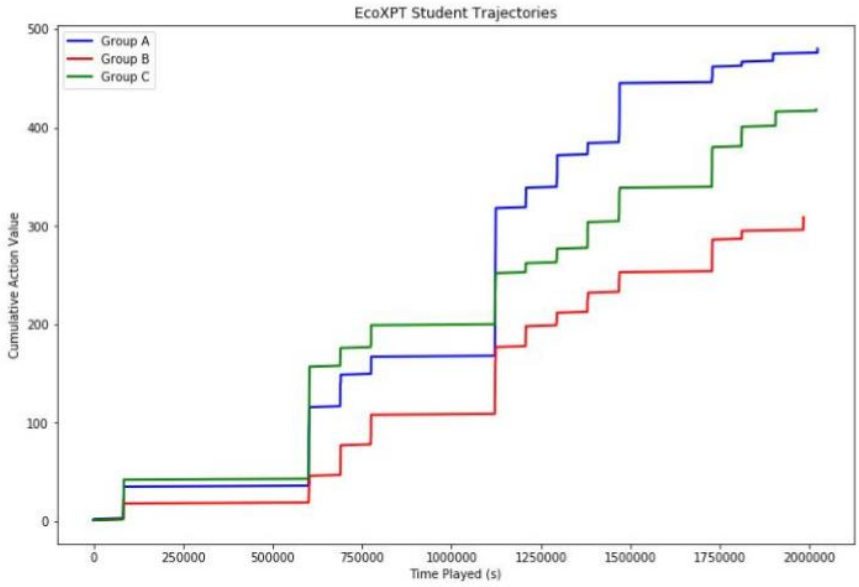


Figure 14. Three exemplar group trajectories (Reilly & Dede, 2019)

2.4.2.1 Summary of Series-based Assessment

In this section, we have provided examples of series-based performance assessment methods. In summary, series-based assessment uses the methods of comparing learner series of actions with typical/expert one, and therefore has the advantage of not requiring large amounts of data to train the model as in machine learning approaches. Educators using series-based analysis need to be careful about the challenge of defining a "golden" (typical or expert) path where there might be multiple paths to follow to solve a problem. Comparing multiple paths requires further investigation. All the series-based performance assessment applications reported in the literature so far have been implemented with the assumption that a learner's full path can be analyzed with the same approach, without any variation based on task characteristics. As such, they may provide an overall assessment without identifying learners' development areas. Treating full series and applying the same analysis method introduces the risk of not using compatible analysis methods for each task and failing to identify students' specific weaknesses.

Regardless of the concerns noted in the previous paragraph, studies summarized in this section provided supportive evidence that students' thinking process can indeed be followed via

scoring learners' actions or analyzing students' full interaction series and can help educators predict students' performance and/or provide the opportunity for in situ guidance. More studies are needed to help instructors in offering more specific and targeted assessments identifying weaknesses and strength in students. This might be achieved by performing assessments on smaller tasks or elements of an activity (for example, a motif). Such a detailed and specific assessment will allow more detailed feedback. We investigate such motif-based approach in this dissertation.

2.4.3 Social and Ethical Consideration in Using Students Data

Advances in hardware and software technologies have made it possible to collect large amounts of data from computer users. The analysis of large data sets to identify patterns and other meta information, sometimes referred to as Data Mining, has been around for decades (Chung and Gray, 1999), but is rapidly growing as the result of the pervasive and ubiquitous nature of computer-based digital products (desktop, mobile, wearable, or embedded). The growth in data sets and their analysis methods is commonly referred to as Big Data Analytics (Tsai et al., 2015). Analyzing the collected data is helping many businesses and contributing beneficially to tasks from policy making to commercial recommendations, and, recently, education (Romero & Ventura, 2013). However, this storage and processing of possibly sensitive data raise serious social and ethical concerns, such as privacy and good use. Methods that allow the knowledge extraction from data, while preserving privacy, are known as privacy-preserving data mining (PPDM) and are being explored by researchers (Kevin, 2017; Mendes and Vilela, 2017).

The introduction of Artificial Intelligence (AI) methods such as Machine Learning to data mining (and data science, in general) has empowered data owners to perform large scale pattern analysis in a much larger scale and shorter times (Jordan & Mitchell, 2015). These advantages are particularly evident when Deep Learning methods are used. These methods are based on Artificial Neural Networks (ANN) with multiple layers and can process a large amount of data with unknown structures through supervised and unsupervised learning methods (Goodfellow et al., 2016). The complexity and hidden nature of these algorithms, while useful in unlocking many patterns in the data, creates a transparency problem as the data analysis tool acts as a black box that no one can properly understand (O'Neil, 2017). This lack of transparency makes it difficult to understand and judge the decisions made by and based on deep learning methods.

Additionally, data analysis methods very commonly use proxy data. Common examples are gender as a sign of interest in certain stereotypical products, instead of direct data on interest, or neighborhood as a sign of education level or cultural background. Lastly, the data is owned by, and analysis commonly benefits, people other than those whose data was collected—for example, businesses instead of customers (O'Neil, 2017; D'Ignazio & Klein, 2020). These structural concerns with big data analytics can cause problems such as data bias, inappropriate reinforcement and feedback, stereotyping instead of personalizing, and confusing correlation with causation (Harford, 2014; O'Neil, 2017).

D'Ignazio and Klein (2020) tasked data scientists with thinking about norms and structures where that data will be used for good. As we are all connected digitally, all our activities can be collected and used for various reasons, e.g., to push for a product or predict votes. Pointing out that data science is a form of power and power can be used for both good—e.g., expose injustices, improve health outcomes—and bad—e.g., discrimination, policing, and surveillance—this power should not be offered blindly. Through what they refer to as *data feminism*, D'Ignazio & Klein (2020) provide seven thinking points for data scientists to consider:

1. Examine Power: Data scientists should ask questions such as the following to identify if the work only enforces existing power structures: Who does the work (and who is pushed out)? Who benefits (and who is neglected and harmed)? Whose priorities are turned into a product (and whose are overlooked)?
2. Challenge Power: Data scientists should look for creative ways to challenge power structures to create a better future.
3. Elevate Emotion and Embodiment: Data scientists should consider how their work supports oppressed groups and minorities, where data feminism also supports elevating suppressed emotions and embodiment to foster change.
4. Rethink Binaries and Hierarchies: Data scientists should make sure that no groups or categories are overlooked, that all are counted and considered.
5. Embrace Pluralism: Data scientists should recognize that complete knowledge comes from synthesizing multiple perspectives, by starting to disclose your methods, your decisions and your own positionalities, all should be up for discussion and improvement.

6. Consider Context: Data scientists should work hard to understand the context of the data collected, social structures, power imbalances around the data collection, and who might have conflicts of interest.
7. Show your/others work: Data scientists should give credit where it is due to show the true cost of the data work.

Digital data collection allows educational organizations to collect data whenever a student accesses the VLE. Student data collected through various academic activities can help educators make connections that lead to insights and improvements. This is what traditionally happens when teachers observe their students and evaluate their work. But the notion of big data analytics in education inherits the issues mentioned above and has concerns such as "technical challenges, ethics and privacy, digital divide and digital dividend, lack of expertise and academic development opportunities to prepare educational researchers to leverage opportunities afforded by Big Data" (Daniel, 2019). Slade and Prinsloo (2013) identified three categories for ethical issues in learning analytics: (1) location and interpretation of data, (2) informed consent, privacy, and identification of the data, and (3) management, classification, and storage of data. Although the (good) use of data (D'Ignazio & Klein, 2020) can be added as a fourth category on its own.

When it comes to education, several authors (Bienkowski & Feng, 2012; Campbell, DeBios, & Oblinger, 2007) have referred to the obligation that institutions have to act on knowledge gained through analytics. Although universities traditionally evaluate students based on academic performance and behaviour, with analytics, the level and scope of assessment can change and become deeper, such as identifying students who might fail based on the category of the profile they fit into (Picciano, 2012). Some universities use students' digital data to create predictive analytics reporting frameworks (Ice, Diaz, Swan, Burgess, Sharkey, Sherrill, Huston, & Okimoto, 2012). It is important that this predictive analysis is used to help students and not to their disadvantage. The University of Texas, Arizona and Harvard University have their own adaptive analytics technology to adjust course content for individual learners based on the collected learning analytics (Parry, 2011, 2012). At Harvard, for example, students can be paired in real time for better collaboration based on in-class data collection (Parry, 2011). All these EDM activities should be driven by companies with ethical policies to inform students and be performed with students' consent.

Scholars have noted the potential conflict between the collection of students' behavioural learning data and confidentiality. Rubel and Jones (2016) argued that it is not clear whether the benefits of collecting learning analytics justify students' loss of privacy. In general, they suggest that any digital data collection system should offer controls for different levels of access and institutions should offer justification and ethical consideration for collecting students' data (Rubel & Jones, 2016).

After collection, there must also be policies and safeguards in place to make sure that individuals cannot access the data in ways that are not clear and "relevant to the purpose for which they are to be used, and to the extent necessary for those purposes" (Watermand & Bruenning, 2014, p. 90). Such safeguards should include data encryption, limited authorization, and limited access to the data (Attaran et al., 2018).

The above discussions demonstrate the importance of not only collecting, accessing, and analyzing the student data properly but also using it for "good purposes" that benefit students directly. O'Neil (2017) summarizes the dangers of using data science and AI algorithms in three categories:

1. **Data**: They use proxy data, i.e., a data item is used for decisions related to something else. For example, home address as an estimate of financial stability.
2. **People**: They do not directly or indirectly benefit the people whose data is being used. For example, they are used by corporations that own the data for their commercial benefit, the data is shared without consent or not protected, and the methods are generalized to work on larger populations even if the data is from smaller groups.
3. **Process**: They are not transparent and work as a black box. So, it is hard to know why they make a decision.

Suggestions such as the seven points by D'Ignazio & Klein (2020) can help manage these dangers. Together, these two studies provide a comprehensive set of guidelines for data scientists and researchers. For the purpose of using learning data for student assessment, we define the following three principles based on O'Neil's three categories of dangers and the possible solutions by D'Ignazio & Klein:

1. **Using data that directly corresponds to what is studied**. This principle addresses O'Neil's point 1.

2. **Providing direct benefit and ownership for students**. This principle addresses O'Neil's point 2.
3. **Consent and transparency in collecting and analyzing data**. This principle addresses O'Neil's points 2 and 3.

These principles directly address D'Ignazio & Klein's points 1 and 7 by identifying and benefiting the real owner of the data, i.e., students, and points 5 and 6 by using the data items that are directly related (requiring multiple data depending on the context). D'Ignazio & Klein's points 3 to 4 relate to including under-represented groups. While this is an important subject, it is not within the scope of this research and requires a more thorough investigation on the topic of inclusion in educational environments. We expect such an investigation to offer new principles or expand the ones we presented.

2.5 Gap Analysis

VLEs are receiving attention from scholars for their ability to enhance educational experiences (Dalgarno, B., & Lee, M. J. ,2010, Dede, C.,2009, Georgiadis, K., van Lankveld, G., Bahreini, K., & Westera, W. , 2018, Gratch, J. & Marsella, S. , 2005).This ability positions them, particularly 3DVLEs, as an effective platform for experiential and situational learning, which are recognized good strategies for HOTS and 21st century skills development. However, without assessing students' skill development, it is hard to show that VLE materials are indeed effective and thus justify the cost of usage and development (Shute, 2011; Sabourin, J. L. 2013; Arroyo, I., Woolf, B .P.; Burelson, W., Muldner, K., Rai, D., & Tai, M. , 2014). Assessing learning is important from two perspectives: (1) to offer real time feedback and customized materials to learners for their further development, and (2) to evaluate the effectiveness of the educational environment and improve where it is needed. Yet, research on assessing the learning and skills development that occur as a result of activity-based curriculum developed on 3DVLEs is still in the early stages (Dede, C, 2009; Reilly JM, 2019; Sabourin, J. L. 2013; Loh & Sheng, 2015), particularly when it comes to HOTS assessment. Traditionally, learning assessment in VLEs has been measured through comparing "before" and "after" assessments (Loh, 2012). However, this method of assessment is not helpful for measuring students' HOTS, which are demonstrated throughout the learning process and should be assessed when and where they are demonstrated

Code & Zap, 2013; Shute & Kim, 2014; Schraw and Robinson 2011; Borgman et al., 2008; Warbuton, 2009).

To date, scholars have used stealth assessment (SA; Shute, 2011) to assess HOTS. SA is categorized into two main groups, as described in previous sections of this chapter. The first category is "score-based SA" (see 2.4.1), where learners are assessed through basic metrics. Preferred assessment methods are applied later once the metrics in use are ready; a simple scoring addition based on algorithms/script or machine learning model creation technique through different machine learning algorithms can be used. Approaches used in creating scripts or algorithms for the full learning activity, which cover each interaction, are quite time consuming and overwhelming, especially when 3D gamified and explanatory aspect of the 3DVLEs are considered. Therefore, usually a simplified approach is used, where each interaction is scored similarly regardless of the location, which can make assessment simple but misleading. For machine learning-based assessment methods, there is a need for reliable training data. Further, any modification to the activity at a later phase requires the script to be changed and for the machine learning algorithms to be re-trained (Floryan, 2015; Shute, 2011). Although by creating machine model accuracy of the learner's performance, assessment can be increased, it is difficult to provide actionable insight to students as a result of this assessment (Loh, 2013). So, machine learning model-based algorithm has two main disadvantages: (i) requiring lots of training data to start the assessment, and (ii) does not provide students with actionable insight as feedback.

Rather than analyzing learners' individual interactions within the 3DVLE platform, some scholars have instead opted to analyze learners' interactions as an activity series that provides information on learners' overall processes. Analyzing overall student path instead of analyzing individual interaction provides context-based insight to analysis. We refer to these methods as "series-based SA" (see 2.4.2). For example, Snow (2015) calculated the stability constant of the interaction series of learners' interactions in the 3DVLE to evaluate learning skills. Loh (2012) compared learners' VLE interaction strings with experts' interaction strings, using a similarity index to provide assessment. Sawyer et al. (2018) and Reilly et al. (2018) also used a learner interaction series to assess performance. However, a series-based approach only provides a single assessment (compared to the expert's series) and, just like any machine model-based assessment, makes it hard to identify specific areas that need development. We believe assessment should be provided on specific skills and elements to be useful for skill development. Assessment methods

proposed in the literature have given promising results in terms of using VLE analytics for HOTS assessment. However, none of these methods have yet reached the level of maturity and practicality to be used by classroom teachers to assess HOTS components.

Shortcomings of the VLEs performance assessment methods can be summarized in the following main points:

1. There is limited research on how to implement educational strategies in 3DVLEs.
2. There is limited research on the use of LMS data for learning and especially HOTS assessment.
3. There is a lack of understanding about the type of metrics to assess HOTS.
4. There is a lack of understanding about the analysis methods that can be used for the different learning tasks and HOTS component the learning task fosters.
5. There is no comprehensive framework to allow for the proper integration of required components for assessment (metric and methods) to be used easily by instructors.
6. The ethical use of student data is still under-investigated and not a major concern in the design of computer-based assessment systems.

Not having informative, easy-to-use VLE assessment methods and metrics, along with a supporting framework, limits teachers' ability to use and configure VLE activities, modify them to fit their needs, and provide a personalized assessment to their students. Our research aims to address the first four points above while offering some insights on points 5 and 6.

3 Study 1 : Integrating Virtual Spaces with Twenty-first Century Learning

While the scope of this dissertation included VLEs in general, it was motivated by our past experiences in 3DVLEs and a belief in their unique capabilities for offering insight into the learning process. The broad research aim was to explore how VLEs—beginning with more common text-based and 2D VLEs to more advanced 3DVLEs—can be used to collect and analyze learning process metrics for the assessment of HOTS. As such, the first study we performed (Study 1) acted as the foundation for our investigation by clarifying the feasibility of our end goal and motivating the three subsequent studies.

Study 1 examined 3DVLEs' role as a learning space with respect to learning theories. Adapting to 21st-century learners' needs, the classroom is not just a physical space, nor is teaching a mere transfer of knowledge that takes place within that space. Rather, the classroom is a mixed media space where knowledge is co-created and learned collaboratively by students and teachers/facilitators. The delivery of 21st-century post-secondary education, especially during and after COVID-19 pandemic, includes a myriad of online spaces, including multi-user 3DVLEs. This study is a response to the need to understand how the affordances of 3DVLEs contribute to general learning and how to develop and assess meaningful learning activities in 3DVLEs as a novel learning platform. This need was identified in our literature review (Section 2.3.2) and such an understanding is foundational for the design of educational activities. The results of Study 1 prepared us for the collection and analysis of process metrics in 3DVLEs as done in the additional studies presented in this dissertation.

Study 1 used qualitative data in the form of observations, but it had an exploratory and informal nature, without using rigorous qualitative research methods. A proper and comprehensive investigation of how learning theories can guide 3DVLE design was beyond the scope of our exploratory study. Here, we only aimed to see the potential and motivate further research. Still, this study offered valuable insight and was essential in future studies.

The findings of our Study 1 were reported in a book chapter, as listed in Section 1.4.

3.1 Overview

Study 1 was designed to address our first research question:

RQ 1a. How can VLEs implement educational activities based on learning theories such as experiential and situated learning?

In this section, we briefly review the context and theoretical background of this study.

3.1.1 Study Context and Objectives

This study was done in partnership Peggy Hartwick, a PhD student who was also an instructor in Carleton's School of Linguistics and Language Studies. The study and the learning activities were designed to be part of an advanced-level English for Academic Purposes (EAP) course at Carleton University (Figure 15). Generally, EAP courses are for students who have an English as a Second Language (ESL) requirement as determined by a recognized English Language Proficiency test (e.g., CAEL, IELTS). Students achieving a mid-range score are placed in one of three levels of the EAP program, which is designed to refine academic English skills and language proficiency. While the course took place on the physical campus twice a week for three hours (pre-pandemic), a major component was a Research Project worth 40% of the course grade and including 16 separate activities, four of which took place in the 3DVLE. Students accessed the 3DVLE from a lab on campus during regular class time. Throughout the term and as part of the Research Project, students were expected to demonstrate evidence of personal learning and growth based on their experiences doing preliminary research related to a commercially sustainable initiative with the overarching theme of "sustainable development" and interacting in and with the 3DVLE. Over the 12-week course, students were evaluated according to whether their chosen initiative met its sustainable development goals according to suitable indicators derived from teacher-assigned sources.

As an exploratory study, the objective of Study 1 was to consider the learning theories and affordances of the 3DVLE technology, design sample educational activities, and qualitatively review the learning outcomes, and then establish initial high-level design recommendations for 3DVLEs. We aimed to answer research question 1a:

1a. Can VLEs implement educational activities based on learning theories such as experiential and situated learning?

3.1.2 Learning Theories

Study 1 was rooted in learning theories that have guided 20th-century teaching practice—namely, behaviorism, cognitive theories, and constructivism.

Educational behaviourists believe that the correct instructional stimuli will elicit the desired learning outcomes, and so teaching emphasizes practice and performance (Ally, 2008; Harasim, 2012). Cognitive learning theorists, meanwhile, do not entirely reject the notion of stimulus and

response, but instead seek to understand the intervening mental processes. Whereas behaviorists reject the unobservable components of thinking and learning, cognitivism views the mind as a machine and information processor (Ally, 2008; Clark, 2001; Harasim, 2012) and understands learning and knowledge in terms of dynamic schematic representations of concepts that an individual shapes in their mind.

In contrast, constructivist learning theory views the learner as an active participant in their own learning process and the teacher as a facilitator of knowledge-making. Knowledge is continuously restructured throughout the learner's experiences (Piaget, 1959; Harasim, 2012). Coinciding with the social movements of the 1970s, constructivism considers learning more holistically and as inseparable from the learner and social environment.

While the above-mentioned theories help account for how people learn more generally, Dede (2003) called for a new model of education to better suit 21st-century learners who are competent digitally and have high collaboration and critical thinking skills. For instance, active learning and learning-by-doing (such as role-play), as well as scaffolded, experiential, and collaborative learning (Harasim, 2012) are all based on the premise that learning is an active process and the result of socially and culturally embedded interaction. Importantly for 3DVLEs, activity theory extends beyond basic bi-directional interaction to consider the complex social context and the technologies in and with which subjects act within situation. The central notion of situated learning is that learning is an active process whereby cognition is shared, and learning is achieved through collaboration and co-participation. Brown, Collins, & Duguid, 1989; Lave & Scardamalia and Bereiter's (2006) commented on the range and diversity of learning approaches, including situated cognition and social constructivism connected with situated learning.

In keeping with Siemens (2005) and Scardamalia and Bereiter's (2006) push for new learning theories, Dede (2007) pointed out that models of education need to shift to keep up with a changing landscape of teaching and learning as emerging technologies are used in and as the classroom. His focus was on preparing 21st-century learners with skills such as critical thinking, problem-solving, and entrepreneurship. He challenged educators and institutions to transform education in terms of what we already know about learning and cognition and claimed this transformation should include delocalization of the classroom community as technologies help to distribute knowledge outside of the traditional classroom space. While he agreed that traditional pedagogies are still important, Dede argued that situated learning is the best theoretical approach

to technology enabled collaboration: “Fortunately, emerging information communication technology that enable three-dimensional, collaborative simulation now offer the capability to implement situated learning environments in classroom settings” (p. 23). The central notion of situated learning is that learning is an active process whereby cognition is shared, and learning is achieved through collaboration and co-participation (Brown, Collins, & Duguid, 1989; Lave & Wenger, 1991; Rogoff, 1990). The assumption is that knowing *how* and knowing *what* are inseparable; further, it is through the collective act and use of authentic physical and social spaces that learning occurs, where “Learning and acting are interestingly indistinct, learning being a continuous, life-long process resulting from acting in situations” (Brown, Collins, & Duguid, 1989, p. 33).

3.1.3 Affordances of 3DVLE

Affordance, as the noun derived from the verb afford, was first used by Gibson (1979). Gibson explains the meaning as what environment provides or furnishes for the user. The key concept of an affordance is understood by investigating it with the observer. A fridge door can be just a visual in one learning environment and can be turned into an active object that needs to be explored in another. Gibson summarizes that the observer “controls the perception of affordances (selective attention) and also initiates actions.” (J.J. Gibson, 1982)

In this section, some of the key affordances of 3DVLE are summarized. The visual and spatial richness of 3DVLEs afford opportunities for experiential learning shaped by other users and by artifacts or 3D objects in the environment.

3.1.3.1 3D visualization of space

3DVLEs provide the ability to visualize a space that realistically mimics physical space or offers non-realistic variations to it. The depth of space and range of artifacts contribute to the perceived realism of space, affording learners experiences in both the visual and physical realm and facilitating social interaction and construction of knowledge. The representational realism of space (the degree to which the space mimics the physical experience) in 3DVLEs is one of the most significant and distinguishing characteristics of these learning spaces (Dalgarno & Lee, 2010). This realism allows learners to explore and understand the environments that could not be available to them otherwise. On the other hand, 3DVLEs can visualize abstract or imaginary environments (such as a stylistic illustration of how the brain works through analogy to a building

with different rooms) to help understand hard-to-grasp concepts. Together, these digital spaces have provided novel opportunities for learning and knowledge building (Savin-Baden, 2008).

In Carleton Virtual, a market area and three different types of residence were created along with other authentic laces such as clinics, where students can explore and have discussion on the meaning. Investigating purposefully designed living spaces, such as healthy snacks in the fridge, and vegetable garden in the backyard provides an opportunity to interpret the meaning of these artifacts and self-reflection on learners' values and behaviours.

3DVLEs can be designed to replicate authentic geographical places like London and Madrid (Ibáñez, Garcia, Galan, Maroto, Morillo, & Kloos, 2011; Milton, Jonsen, Hirst, & Lindenburn, 2012) or authentic physical spaces like a classroom, archeological dig site, or nightclub (Arya, Hartwick, Graham, & Nowlan, 2012; Milton et al., 2012).

3.1.3.2 Immersion

Immersion in VR/3DVEs is related to full engagement and, to some degree, to the sense of presence (being there). The ability of the environment to fully immerse the user reduces distractions and allows the simulation of real experiences that are required for experiential and situated learning. Immersion helps with the sense of flow ("being in the zone") that is shown to increase productivity and enjoyment (Csikszentmihalyi et al., 2018). However, such immersion should not interfere with the students' ability to critically analyze the environment. Simulating the real life, should encourage students to observe, think, analyze, and judge, and not "get lost" in the environment, due to the so-called "awe factor" or other reasons (van Limpt-Broers, 2020; Quesnel & Riecke, 2018). The affordance of immersion also includes the user's ability to personalize their own avatar according to traits like hair and skin colour, height, and weight, resulting in a 3D self (Dalgarno and Lee, 2010). The ability to customize personal avatars can create association and lead to a sense of presence, as through avatars users can perform gestures, such as a hand-wave or shake, or bow, and even make facial expressions and head movements (Peterson, 2006).

3.1.3.3 Interaction

In VR/3DVE, users can interact with objects in the environment or other avatars socially by speaking, which creates a high presence and positively impacts the performance and effectiveness of the task (de Freitas, 2006, 2008). The ability to control interactions through avatars, body gestures, and other mechanisms helps users behave in ways that are familiar and comfortable to them and creates familiar learning experiences that require a shorter adjustment

period (Biocca, 2014). As Kim et al. (2014) reported, more players connected to the avatar strong physical reaction (heart rate, behaviour effort) were received. This connectedness allows educators to foster learning from events such as a chemical spill in a virtual lab that could not be possible in real life or would be less effective by just watching a video.

3.1.4 Implications

In summary, the above-mentioned affordances (authentic space creation, immersion, and interaction) provide opportunities for synchronous interaction and mediation through tools and collaboration with others. These affordances make it possible to contextualize 3DVLEs as learning spaces where learning task design and creation are based on the learning principles from the previously mentioned theories (particularly constructivism and experiential/situated learning), including negotiation of meaning and development of HOTS as a result of external social processes, influences, and interaction (Harasim, 2012; Kaptelinin& Nardi, 2006; Prensky, 2006; Vygotsky, 1986). Further, learning opportunities can be mediated not just by the space but through virtually rendered artifacts and non-player avatars, which are not controlled by a user, but rather purposefully designed for the activity (Dalgarno & Lee, 2010; Warburton, 2009).

The conceptualization of the classroom has changed with the provision of online spaces and due to the learning affordances of these spaces, but how do these perspectives inform space and instructional design? The authors suggest that these unique learning contexts may have the potential to facilitate learning for some learners in a blended context by creating alternative learning spaces. Similarly, the potential for distance language learning in these spaces is significant, as supporting evidence put forward in studies involving virtual worlds and online language education (Kim et al., 2012).

3.2 Participants

Participants in Study 1 were ten students in an EAP course instructed by our research partner, Peggy Hartwick. Recruiting students was done through a process approved by Carleton University Research Ethics Board in September 2016. Student participation was voluntary and had no impact on student grades, as neither of the researchers had access to participant names until winter 2017, after final grades for the course had been submitted and the window for contesting grades was closed. This process was clearly explained to the students.

Volunteer participants from three sections engaged in the same lessons and assessments as their non-volunteer counterparts. The ages of participants ranged between 18 and 22. All ten participants were studying concurrently in their degree program at the same university and came from a range of first language groups but were primarily Chinese and Arabic speakers.

3.3 Materials

This study took place in Carleton University's 3DVLE, Carleton Virtual, which had been used for language learning for over five years (Arya, Nowlan, & Sauriol, 2010; Figure 15). Over this five-year period, the VE had evolved to include new spaces, including an Orientation Maze. Each iteration of the space was determined by the design team and involved matching teaching objectives to space design and identifying how the space would help learners achieve outcomes that might not otherwise be possible in a physical classroom, such as exploring an archeological digging site and the way it should be layered (Arya et al., 2011). This included determining the options, virtual agents, and objects needed for successful task implementation. The tasks designed and described for this study made use of five following locations in the 3DVLE:

Figure 15. Carleton University Virtual Campus

Orientation Maze: The *Orientation Maze* is a virtual path that progressively guides students through the functions of the platform so they can be technically and spatially prepared to engage in upcoming tasks. The maze is designed to be followed without a facilitator. By the end

of the maze, students will have successfully performed the most common features of the 3DVLE platform, thereby reducing technical and spatial barriers in future activities.

Career Island: *Career Island* represents a self-sustainable island in which learners explore various professional roles represented by non-player avatars positioned throughout the island in virtual rooms reflecting their role, such as a nurse in a small clinic (Figure 17) and a marine biologist in a research lab (Figure 18). The virtually rendered rooms further reflect the careers through animated artifacts, such as a patient in the clinic. The non-player characters (agents), if clicked on, provide information about the type of person who might be suitable for a career in that field and typical challenges.

Residences: In addition to Career Island and the Orientation Maze, the 3DVLE has two main functional areas, a virtual university campus and a downtown area with shops, a café, and three houses. Each of the houses in the Residences section can be designed and textured to reflect the eating, consuming, and cultural habits of the hypothetical inhabitants.

Campus and Rainbow Stage: The *Campus* area and buildings are designed around a central courtyard and replicate, for the most part, a university. The area includes a library, virtual classrooms, and offices. To take advantage of the expansive space and users' perception of being outside, more recently, an open-air classroom called *Rainbow Stage* was added. This space is ideal for group work as it includes six breakout stations complete with work surfaces and contained sounds.

Carleton Virtual started on the Avaya Engage platform (no longer operational) with a classroom and open area simulation. Over the years, more on and off-campus areas were added to support immersive 3D language learning activities. The platform currently uses Virbela (http://virbela.com), which has similar capabilities as Avaya Engage but limited data collection tools. The self-learning activity used in this study is a tutorial step for users prior to performing the other course-related activities.

3.4 Procedure

During regular class time, students developed language and academic skills through activities such as reading and learning about related sub-topics like sustainable indicators and the pillars of sustainable development. In addition to thematic content, students learned and practiced academic skills, such as citing and referencing sources and reflective writing.

The goal of the study was to observe students within a learning task in the 3DVLE with active and experiential learning theory in mind. There were five interrelated and sequential tasks, the latter four of which played out in the 3DVLE described above. The tasks were designed to target students' content knowledge regarding an understanding of their majors and disciplines and thematic topic; language in terms of topic-specific vocabulary; and meta-cognition regarding their ability to make connections between experience and knowledge. All tasks were designed to determine whether the student had achieved the desired learning outcomes according to the task design and affordances of the 3DVLE.

The tasks, described below, are the result of the design team's more than four years of experience and modifications. The recursive process allowed team members to systematically account for the affordances of space and characteristics of how people learn, such as practice, collaboration, and interaction. Tasks were intended to elicit critical thinking or problem-solving, to be demonstrated through actions and written language.

Activity One did not take place in the 3DVLE but it helped students develop an awareness of their interests, values, and possible major. It was also a necessary step to guide Activity Three by providing vocabulary related to personality traits and career preferences. Activity One was done in the physical classroom during class time and consisted of students taking a modified version of the Holland Code Assessment to identify personality traits, skills, and career preferences.

Activity Two: Spatial and Technical Orientation Learning Environments. As previously mentioned, the Orientation Maze was designed to give students a spatial and technical orientation to the 3DVLE before tackling other tasks. It was also used to determine which types of strategies students relied on to efficiently complete an activity (e.g., asking peers for help or going back and starting the maze again). Finally, students were asked to reflect on their personal learning strategies by answering the question, *Which strategies did you use to complete Activity Two (circle as many that apply)?* Possible responses were: (1) I asked another avatar by using my headset and voice; (2) I went back to the maze; (3) I worked with another avatar; (4) I followed/watched a peer avatar; (5) I did not use any strategy(ies); (6) Other; (7) I gave up/logged off.

Activity Three: Explore Your Career Learning Environments. In Career Island, the virtually rendered rooms further reflected the careers through animated artifacts, such as a patient in the clinic (Figure 17). This environment, while not authentic, was intended to contribute to a sense of immersion and real-time presence. From a constructivist perspective, task instructions

assumed that learners would actively collaborate and be resourceful by taking advantage of the objects and tools within and outside the environment in order to successfully complete the activity.

Figure 16. Screenshot of nurse's clinic on Career Island

Figure 17. Screenshot of biologist's lab on Career Island

At the end of the Activity 3, students were asked to explain how their experiences interacting in the career spaces related to what they had found out thus far about their own choice of major. Students were also asked to respond to two questions that targeted critical thinking. For example, students who selected the nurse career and who explored the clinic were asked to respond to the questions: (1) *What do you think is wrong with the female patient in the clinic?* And (2) *Have you ever cared for a sick animal?*

Activity Four: Defining Sustainable Development Learning Environments. In addition to Career Island and the Orientation Maze, the 3DVLE has two main functional areas, a downtown area with shops, a café, and three houses, and a virtual university campus. Each of the houses in the Residences section was designed and textured to reflect the eating, consuming, and cultural habits of the hypothetical inhabitants. As such, walls, refrigerators, furniture, and level of cleanliness were textured to suit the learning task and outcomes. Students explore these spaces and talk about the personality and environmental habits of the residences of the house. This experience would not have been possible in a traditional classroom space; this would have necessitated a tour of the physical library or other campus buildings.

Activity Five: Applying Sustainable Development Indicators Learning Environments. The Market (Figure 19) design was based on design team members determining the need for any tasks related to food, shopping, and health. In connection with the theme of sustainable development it was expected that as students interact with the space, they notice variation in products and costs, as well as ethical considerations such as fair-trade items, thus prompting critical thinking and problem solving. These skills are compatible with previously identified 21st century skills.

In Activity Five, students were asked to explore a virtual representation of Ottawa's Byward Market area (Figure 19) in groups. As they explored this area, students were tasked with answering questions related to sustainable development using a collaborative writing surface. The affordance of real-time voice allowed for peer feedback. In real terms, successful completion of the activity depended on students' collaboration, freedom to move and create, and a certain level of proficiency using a collaborative writing tool and moving and exploring in the 3DVLE.

During each activity, the teacher took observational notes, both within the environment and during verbal discussion reflection notes. Students also filled out written reflections after each task. At the end of the study, observation notes were compiled by the teaching and research team

and lessons learned and recommendations were collected and reported. Some of the students' feedback on their learning tasks was captured as well.

Figure 18. Carleton Virtual market area

3.5 Results

Overall, this study found that VLEs can implement educational activities based on learning theories such as experiential and situated learning (research question 1a). In the following sections, we briefly review the learning outcomes, some design recommendations, and directions for future research. It should be emphasized that this was an exploratory study and did not involve rigorous data collection and analysis. As a pilot project, we relied only on informal observation and assessment by the instructor, and no rigorous data collection and analysis method was used.

3.5.1 Learning Outcomes

While we did not collect any formal data in this exploratory study, informal observation of the students by the instructor and anecdotal results suggested success in achieving these outcomes. It should be noted that the first activity was outside the 3DVLE, and the second one gave students a general introduction to the 3D environment and navigation within it.

Activity 3 used *Career Island,* where learners explored various career options. Here is a comment from one participant:

"The second career that I chose is nursing because I want to help people and know how to treat sick people, and give them the right medicine. According to Holland Code, the nursing area

of study of social but I am not a social person but I like this career. According to 3D lesson, the nursing in the video talk about how to help people and you need to be attention to details."

This comment demonstrates her awareness of her personality traits as she justifies how nursing is a suitable career choice, although she is not social.

Primary demonstrated outcome for activity 4 was the use of formal and content-specific language., and also problem-solving and the ability to think critically based on experiences in and within the space. One student's comments clearly demonstrated the value of movement, exploration, and experience in the campus space: *"Firstly, we can use the new type of energy. The sunlight is enough, so it is able to install some solar panel to collect solar energy."* In this case, the student linked their experience in the 3DVLE to one of the sustainable initiatives from their class readings and course content.

Demonstrated outcomes for activity 5 included evidence of topic-specific vocabulary and content knowledge in the shared written document. The ability to achieve outcomes successfully included demonstrating connections between experiences in and because of the space and due to collaborative thinking and problem solving. Ideally, students could account for their experiences by including references to the input of peers and the space. In response to the prompt *("Identify possible trade-offs, as defined in the reading, in terms of access to food."),* one group responded collaboratively:

> *'Access to food is much easier in Downtown because of the shops available. Furthermore, people who live in downtown can try multi-cultural food and it is more convenient, but the people living in other areas may have access only to specific food.'*

In this response, the students demonstrated connections between their experiences exploring the space and worked collaboratively in producing a relevant response. Also, there is evidence of attempts at using comparative language and examples, which are academic conventions expected at this level.

3.5.2 Design Recommendations

While the subject of designing 3DVLEs and the virtual experience is beyond the scope of this thesis, Study 1 resulted in some general and preliminary recommendations for such design. The researchers suggest that task design should take advantage of such things as the authentic representational quality of 3DVLE spaces and their ability to provide the simulated experience of promoting content and skill development by means of enabling users to interact with the tools and

artifacts available in the space. These are characteristic of learning theories addressed earlier and align with the 21st-century learning trends and constructivist perspectives. The implication is that learners are free to explore, gather, select, and evaluate information anytime and from anywhere (Ally, 2008). Importantly, 3DVLEs are living spaces that have the potential to continuously adapt to the design team's experience; thus, a good platform should support all above-mentioned features, as well as be easy to modify and deploy on a server that students and instructors can access without difficulty.

The experience in a 3DVLE should be motivating and engaging (Ally, 2008; Clark, 1994). As such, in addition to designing tasks that foster interaction and collaboration, the 3DVLE experience should include high-quality audio as a critical component. This was evident from the participants' mention of "attention to details." Further, based on the affordances of authentic space and immersion, and participants' emphasis on "people", a chosen 3DVLE platform should support avatar customization for students to personalize their own avatar, along with non-player characters to interact with. Virtual spaces should be populated with non-playing characters that are animated and include voice to create realistic opportunities for engagement.

3.6 Discussion

3.6.1 Reflection on Findings

Despite its exploratory nature, Study 1 was essential and successful in motivating our research. The primary insight from this study was that 3DVLEs, through their unique affordances, have the ability to implement various educational strategies. This motivated us to explore the subject further to discover what is required (and missing) within the field of 3DVLE. The study was limited to one topic, a small group of students, and a limited number of educational activities. As such, and through related literature review, we discovered various opportunities not only for educational developments in 3DVLE but also for further research to explore how to achieve different learning strategies.

The second important insight that came out of this study was the importance of observation. We realized that 3DVLE allowed the instructor to "see" what the learners were doing and have a better understanding of what they could or could not do. Considering the limitation of human resources (instructor's time and ability to observe all learners), the value of process metrics in 3DVE (or VLE, in general) emerged as an important insight. This was confirmed by study 1's

reliance on human observation and the growing literature on computer-based assessment, as seen in Chapter 2, and made us realize that prior to implementing different learning methods, the educators need to have proper ways to observe and assess the learners. Our main research objectives and questions emerged from this insight and formed the basis of our next studies.

Finally, the third overall insight from Study 1 was that while there are many activities happening within the virtual environment (such as opening a door, looking at an object, conversing with an agent, etc.), these actions are very detailed and low-level. Their exact role in achieving learning objectives can only be understood when we group them together and see a pattern of actions. This introduced the notion of combined process metrics vs. basic process metics that we explored throughout the other studies, investigating different forms of combining basic metrics.

Overall, Study 1 began exploring the answer to our research question 1a:
How can VLEs implement educational activities based on learning theories such as experiential and situated learning?

We realized that using various affordances of 3DVLE, we have the ability to design interactive and engaging experiences that allow learners to perform actions within environments that resemble real-world situations. Such an ability allows the implementation of various learning strategies and potentially offers automated and manual observation and data collection necessary for assessment.

3.6.2 Limitations and Further Research

As an exploratory effort, there was a very limited level of data collection and analysis in Study 1, as the goal was mostly to motivate further research. While the study was successful in establishing the initial insight and motivating the research by showing potential, further research is needed in the areas of the curriculum and task design as well as assessment. We particularly suggest the use of analytics in combination with learning activities to offer the ability to perform an assessment. The use of process metrics for assessment is the main point of our research and the next three studies, as we believe without assessment, learning activities are not complete.

Research should appreciate and understand the role of the environment as a mediator in the learning process and use 3DVLE analytics to measure the duration and frequency of interaction to understand the impact of space on performance. As such, the remaining parts of this thesis are dedicated to the *learning analytics*, defined as “the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing

learning and the environments in which it occurs" (LAK '11, 2017). In particular, we aim to investigate the type of process metrics (including combinations) and analysis methods that can help with the assessment of HOTS.

4 Study 2: Learning Skill Assessment on an LMS

Our foundation study suggested the importance of observation and data collection throughout the learning process. As our next step, we wanted to see if detailed process metrics from a VLE (text-based or 3D) can help with such observation. Due to their detailed nature, many of these metrics (for example, opening a file or looking at an object) are too specific and low-level to be clearly associated with learning objectives. As such, a group of process metrics collectively (i.e., an aggregation of these metrics along some dimensions) could be associated with an educational concept. In this study, we investigated students' online activity patterns as recorded by Moodle[12], a Learning Management System (LMS), along the dimensions of Attention and Participation. These two dimensions have been investigated by educational researchers due to their association with student success. Identifying the importance of students' attention as an analysis factor, Narayanan et al. (2012) propose using eye-tracking system to capture that information and provide it to teachers. Users' attention is also the topic of study by Ugurlu (2014), who proposed using a portable camera and microphone system to capture that. Following students' participation through forum posts, homework being submitted, and quizzes taken, has received attention from researchers (Romero et al., 2013; Mueen et al., 2016). Aggregating metrics under two components may result in losing some information on individual metrics. However, this is just an exercise of looking at the existing data from a different perspective and does not prevent us from analyzing the basic metrics individually. To the best of our knowledge, using students' attention and participation captured through logged activities has not been investigated for learning assessment.

We defined attention and participation as accessing or viewing existing information and creating new information, respectively (Lokse, 2017). If the students do not leave any mark on the system for either instructor or other students to see, this is treated as attention. Any submission in the form of homework, forum enter, quiz etc. are treated as participation. Details of each metric grouped under attention and participation are given in the Procedure section in 4.4.We constructed them as aggregated process metrics by combining a set of basic metrics such as viewing a course file or posting a comment, as described in the following sections. We aimed to see if there was any correlation between students' Attention and Participation and their grade-based performance. Our

[12] https://moodle.org/

goal was also to investigate the effectiveness of aggregating process metrics into those two dimensions and studying their change over time as a way of gaining insight into students' future success. In addition to exploring the aggregated metrics, this study aimed to show that process metric-based HOTS assessment can be done in both text-based and 3D VLEs.

The findings of our Study 2 were reported at EduLearn-2019 conference, as listed in Section 1.4.

4.1 Overview

The ability to record students' every interaction in VLEs provides educational organizations with a valuable opportunity to anticipate students' needs and provide timely support (Montgomery et al., 2015; Brooks et al., 2014; Romero & Ventura, 2013; Baker & Clarke, 2010; Baker, 2010; Scheffer, 2001). The challenge, of course, is to interpret user data collected by the system in a way that provides students with timely suggestions for actions to improve performance.

Data Mining (DM) is a discipline of discovering novel and potentially useful information from large amounts of data (Baker, 2010). Educational Data Mining (EDM) uses methods defined by DM and improves and adopts them to educational settings to gain insight into students' learning and the settings in which learning occurs (Scheffer, 2001). There is a wide variety of current methods popular within EDM. These methods fall into the following general categories: prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment (Montgomery et al., 2015; Brooks et al., 2014; Romero & Ventura, 2013; Baker & Clarke, 2010). In clustering methods, the aim is to find natural groupings through dividing data into sets of clusters based on similar characteristics. Clustering is an unsupervised analysis method that works well on data where the outcome is not known originally or when applied to unlabeled data (Hartigan & Wong, 1979). The most popular clustering algorithms are K-Means, K-Medoids, Distributed K-Means, Hierarchical clustering, Grid-based, and Density-based clustering (Patel et al., 2016). While a full discussion and comparison of these algorithms is beyond the scope of this dissertation, the fundamental principle of clustering methods is defining data as an array of values and defining clusters as groups of data arrays that are "closer" to each other based on some measure of distance (Mori et al., 2016). Among its many applications, clustering can be used to study patterns of behaviour in students (Amershi & Conati, 2006; Beal et al., 2006).

Observing students' activity as a time series has received attention as a way to gain insight in students' study patterns and higher-order skills (Montgomery et al., 2015; Ahadi et al., 2015; Aghabozorgi, 2015; Romero & Ventura, 2013). By employing the Generalized Sequential Pattern Mining (GPS) algorithm (Mooney, 2013), Fatahi et al. (2017) successfully identified behaviours that separated learners into different learning styles (Furnham, 1996). A study by Cerezo et al. (2017) used the k-means algorithm and found four different clusters from student data collected on Moodle. The three most related variables were (1) the final marks, (2) completion time, and (3) the number of posted words in forums. Time series, with or without clustering, is an often selected way of studying the dynamic behaviours such as student activities.

Despite these applications, there are limited studies on the effectiveness of combined/aggregated metrics that turn lower-level activities such as opening a data file into high-level constructs such as attention and the importance of these low-level activities compared to participation. Such constructs are more appropriate when assessing HOTS as they demonstrate the thinking pattern of students. Basic metrics and their related actions are too small and specific to show those patterns. The need for investigating attention and participation, the gap in exploring the use of aggregated metrics, and the need for observation and thorough assessment are the motivations for Study 2.

In this study, we investigated the role of attention and participation in relation to students' performance (grades). We used multiple categories of interaction counts combined as attention and participation metrics. These newly formed metrics then build a time series over the course. Clusters of these timeseries were then formed from all students and their correlation with student performance was investigated to see if they could be used as indicators for performance and learning. Our specific research questions were:

RQ 1b: How can individual basic metrics be used to assess HOTS?
RQ 1c: How can combined process metrics be used as an aggregated HOTS assessment tool?
RQ 2a: How can a combination of time series of aggregated process metrics and clustering provide a way to predict learning and academic performance?

To the best of our knowledge, this specific investigation has never been done before on data from a fully online Moodle course. Hussain et al. (2018) used the same publicly available data

to identify students at risk of failing and features providing a strong indication of the risk of failure. They used the data as static and the K-mean clustering method. Greene et al. (2006), meanwhile, found evidence suggesting that the timing of these activities and treating Moodle data as timing series could provide additional information that provides a better model with a higher prediction rate.

The course used in Greene et al. (2006) was not a fully online course. Further, assessments mostly correlated with a pass/fail dimension rather than over different performance profile dimensions, such as survivors, early achievers, late bloomers, etc. Also, most of these studies were performed on hybrid course offerings, where lots of interaction also goes on in the face-to-face part of the course. Therefore, online interactions might be limited in terms of capturing students' full course-related activities. We believe it is important to understand the role of each component in the learning process in order to provide timely, customized feedback.

The proposed assessment method followed our guidelines for "the good use of data"; we used clearly explained algorithms, data that was directly related to what we investigated, and results that were used to benefit the students themselves.

4.2 Participants

This study did not include any participants and primary data. We used a publicly available data set of student activities. The data set used for this study was collected from a free online course (*Teaching with Moodle*; Research.moodle.net). The course was offered from the 7th of August to the 4th of September 2016. The data was anonymized and made publicly available for research purposes. Each student was given a unique, anonymized username, and each action on every unique instance of the interaction was captured accordingly. The number of students enrolled in the course was 6119, but only 12% of the students fully completed the course. In this research, we chose to perform analysis only on data from active students, defined as students who participated in at least one graded activity. No demographic information was available about the students.

4.3 Materials

We used the LMS data captured from a fully online course on Moodle (Modular Object-Oriented Development Learning Platform, http://www.moodle.org). Data was anonymized and

offered with creative commons license by Moodle Pty Ltd (Research.moodle.net). The following data was collected and offered to the research community:

- Records of each module in the *Teaching with Moodle* course.
- Records of each compilation of each activity in the course.
- Historical records of individual grades for each user and each item, exactly as imported or submitted by modules; and
- Records of entries for each "event" tracked by the Moodle logging system, and the source for all the Moodle "log" reports.

Below are the tables and corresponding records within each file:

Table 2. Moodle data files and file descriptions

File Name	File Description
mdl_badge_issued.csv	This file contains the records of all badges issued to users during the August 16th session of *Teaching with Moodle*.
mdl_course_modules.csv	This file contains records describing each activity in the *Teaching with Moodle* course.
mdl_course_modules_completion.csv	This file contains records of each user's completion of each activity in the course.
mdl_grade_grades_history.csv	This table keeps a historical record of individual grades for each user and each item, exactly as imported or submitted by modules.
mdl_logstore_standard_log.csv	This table contains entries for each "event" tracked by the Moodle logging system and is the source for all the Moodle "log" reports.
mdl_user.csv	This table contains user records.

Details of the basic process metrics list is given in the next section. To the best of our knowledge, this data has never been investigated through indexing with time to gain better insight into students' learning process.

The SQL database instructions were used to clean and organize the data. R Studio was used to perform cluster analysis.

4.4 Procedure

The digital nature of an LMS makes it possible for educational organizations to capture data that is not possible in face-to-face learning environments. The Moodle LMS captures a wealth of data and can provide information to educational organizations to understand students' needs and help them improve their performance.

Most data collected in the real-world changes over time—heart rate, temperature, students' activity, etc. Times series is a series of data points indexed (or listed or graphed) in time order (Pickup M., 2015). Time series analysis is a method of observing data and changes over time and performing analysis on this data sequence as changing data in order to extract meaningful statistics.

In this study, we proposed to understand students' needs by analysing Moodle by forming and applying time series analysis methods from two specific Moodle tables: i) mdl_grade_grades_history and ii) mdl_logstore_standard_log. The first table contained students' grades received from all activities, quizzes, workshops, forums, discussion with the timestamp, and the max grade students could receive from this activity. The second table captured all the activities for each student performed in the system, including viewing a course, checking grades, viewing a discussion in a forum, etc.

Activities performed on course modules such as Mod_quiz, Mod_page, Mod_Data, Mod_forum, Mod_lesson, etc, were classified as per their action identifier. If the action only viewing, we classified it as "Attention". All other actions—e.g., uploaded, updated, submitted—were classified as "participation".

The attention aggregated metric included the following items:

- Course content view
- Course assignment view
- Course discussion view
- Course forum view
- Course quiz view without submitting
- Students' attention to the administrative side of the course, e.g., students accessing information such as which group they are in, timeline of the course, exam dates and percentage weighting, etc.

The Participation metric included the following groups:

- Uploading an assignment
- Posting a discussion
- Asking a question
- Participating in a workshop
- Posting a forum entry

To follow students' activities over the course timeline and observe the change, we have to choose a window and aggregate. The number of activities in a day from each category was used as a value and added up to create a daily value.

With respect to performance or success criteria, the following attributes were used in our analysis as static values:

- Unique id of grade record
- Maximum allowable grade
- Final grade of the student

After splitting the data and creating time series on Attention and Participation, we applied time series-based clustering analysis to create meaningful clusters, as shown later in Section 4.5. Next, we aligned the activity-based clusters with performance-based clusters created with final marks as absolute and percentage, and looked for meaningful insight. Figure 20 summarizes the steps and analysis performed.

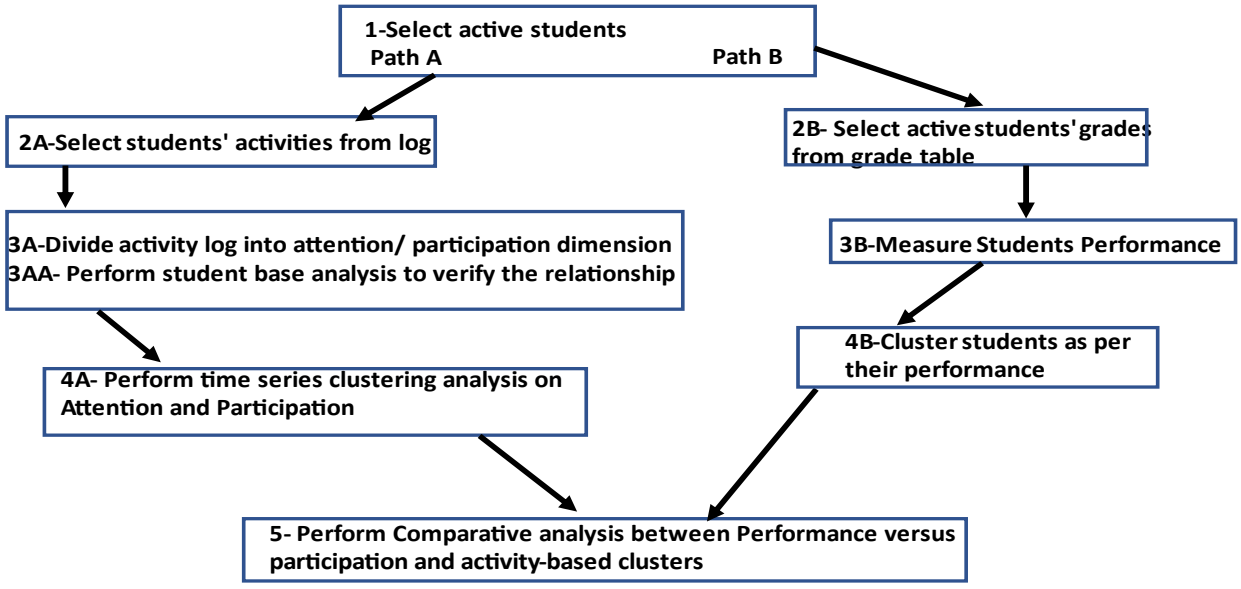


Figure 19. Study 2 Analysis Flow

4.5 Results

In this study, two branches of data preparation and analysis were performed.

i) Activity pattern analysis on the Attention and Participation dimension on one branch (identified as "A" path in the steps detailed below); and

ii) Grade-based analysis on another branch (Identified as "B" path in the steps detailed below).

Step 1- Selecting active students: Students who took at least one graded activity were considered active students and included in the analysis. Moodle file mdl_grade_grades_history was used for this step and the final grade value was checked. 1202 out of 6119 students were identified as active students.

Path A Analysis Flow:

Step 2A- Select log entries for active students: From the log file, active student Attention and Participation entries were selected for further analysis as per the list provided in section 4.4. Moodle file mdl_logstire_standard_log file was used for this step.

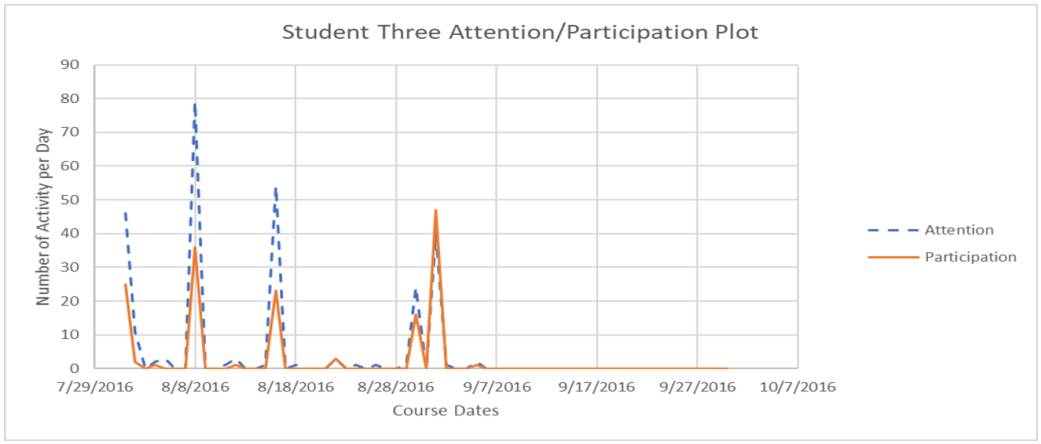
Step 3A1: Grouping activities into "Attention/Participation dimension Entries selected from log files belonging to active students were grouped into two main groups: "Attention" and "Participation". All the action entries "viewed" by students regardless of the module activity (forum, book chapter, viewing the grade etc.) were grouped under attention. The rest were grouped under participation.

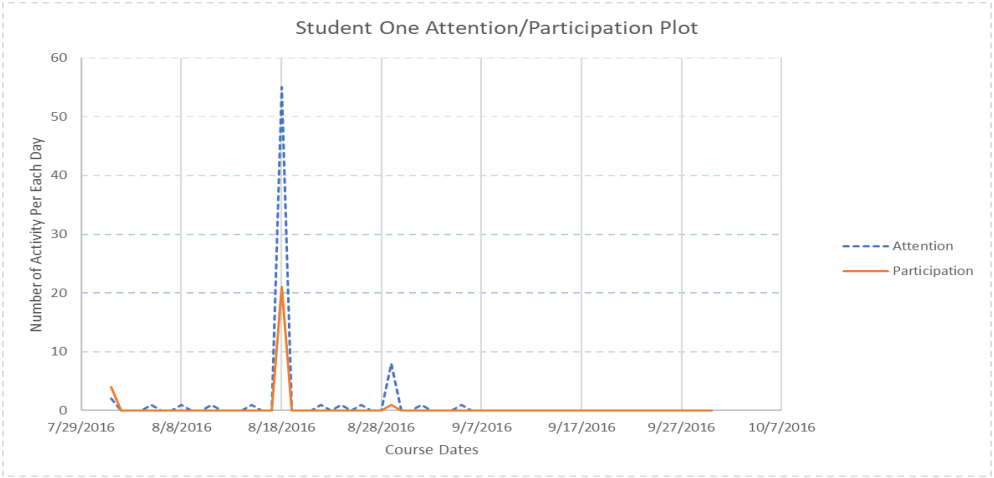
Step 3A2: Students Attention/Participation Time Series Creation & Per Student Analysis:

For the purposes of this study, we decided to perform full series analysis and used 24 hours a day as a bucket of data to create daily time series for each student (Figure 21).

After creating the time series representation, to verify the correlation between the attention and participation activity series, individual case analysis was performed on selected students. Below are sample time series plots of individual students on the Attention and Participation dimension.

Visual time series analysis on randomly selected students provided insight on how the attention and participation activity series were related. Although this visual correlation was not a necessary step for the Study 2, it is recorded here just as an additional observation. The next step was to explore if a separated cluster analysis would provide complementary insight.





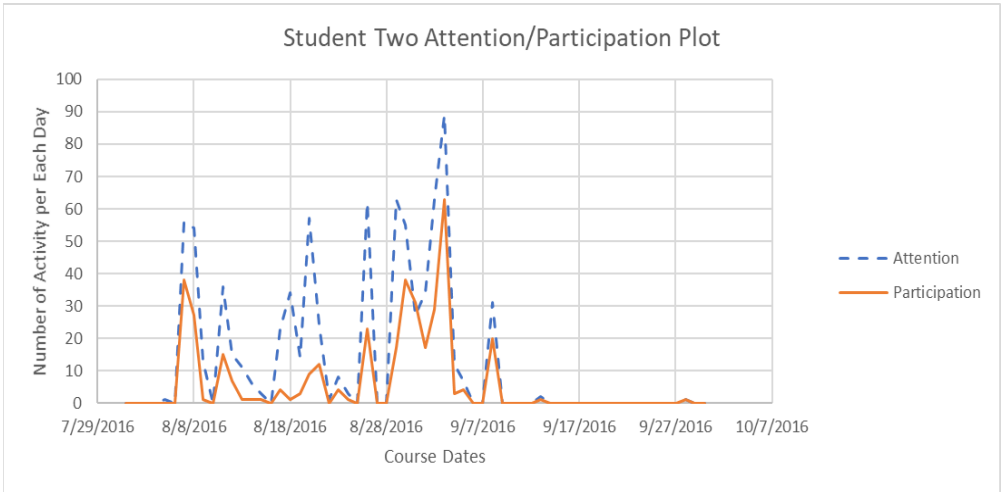


Figure 20. Student 1-2-3 Attention count versus Participation count trend

Step 4A: Applying time series clustering learning to identify emerging patterns:

Clustering is a solution for classifying enormous data when there is not any early knowledge about classes (Aghabozorgi et al., 2015). There are popular, commonly used time series clustering algorithms in the literature. In data analysis, when data is used as fixed data, not indexed to time, the analysis is then static data analysis. K-Means is one of the most widely used static data analysis techniques. It finds groups of clusters within unbaled data based on given parameters. There are two steps in the algorithm: data assignment and centroid update. Centroids can be randomly generated or provided via function parameters. Cluster centroids and mean distance between centroids and cluster points are continuously observed and updated to minimize distance.

Time series clustering is performed in four stages:

i) Time series representation: Done in Step 3a through daily bucket counts per attention/participation.

ii) Similarity and distance measure selection: In terms of distance measure, previous studies provided insight that DTW distance measure might be a good fit for analyzing learning patterns (Greene, Cunningham, 2016). Therefore, we decided to test three similar types of distance measures on time series pattern: Shape-Based Similarity measure (SBD), Dynamic Time Warping (DTW) and DTW_basic.

iii) Applying clustering algorithm based on distance matrix to classify the data: (Aghabozorgi S, Shirkhorshidi A., 2015). There are popular, commonly used time series clustering algorithms in the literature. K-Means and K-Medoids (PAM - Partition Around Medoids) are two of the most widely used static data analysis techniques when data is fixed. The *K*-means clustering algorithm is sensitive to outliers because a mean is easily influenced by extreme values. *K*-medoids clustering is a variant of *K*-means that is more robust to noises and outliers (Harting and Wong, 1979). Instead of using the mean point as the center of a cluster, *K*-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with a minimum sum of distances to other points. In this dissertation study we used K-PAM clustering. Since the structure of the data was not previously known, the usual strategy would be to run the selected algorithm several times with different numbers of clusters (k value). We experimented with the clustering with k values from 3 to 7, as we believed this

range would give us manageable numbers of groups to interpret but also meaningful divisions.

iv) Validating the clustering and choosing the clusters best fit on data through cluster validation indices (CVI). All the partitions were created by assigning multiple k values to be evaluated and selecting the partition that best fits the data. The process of estimating how well a partition fits the structure underlying the data is known as cluster validation (Halkidi et al., 2001).

We used three distance measures (DTW, SBD, and DTW_basic) and performed clustering with k values 3 to 7. Clustering validation indices were examined to find the best possible clustering. Based on our analysis, the following were the combinations giving the best CVI:

- For Attention Time Series: DTW_basic clustering with cluster number 5
- For Participation Time Series: DTW clustering with cluster number 6

Below figure are the R program outputs (Attention-based, Figure 22); Participation-based, (Figure 24), with the cluster groups and centroids for students' Attention (Figure 23) and Participation (Figure 25) dimensions for the selected best cluster validation indices (CVI) combinations and their centroids:

```
partitional clustering with 5 clusters
Using dtw_basic distance
Using pam centroids

Time required for analysis:
   user  system elapsed
  45.42    0.05    5.91

Cluster sizes with average intra-cluster distance:

  size   av_dist
1   62 14.132376
2  117 16.904076
3  331  8.761409
4   17 82.148893
5  675  4.970541
```

Figure 21. R Attention-based time series clusters in R

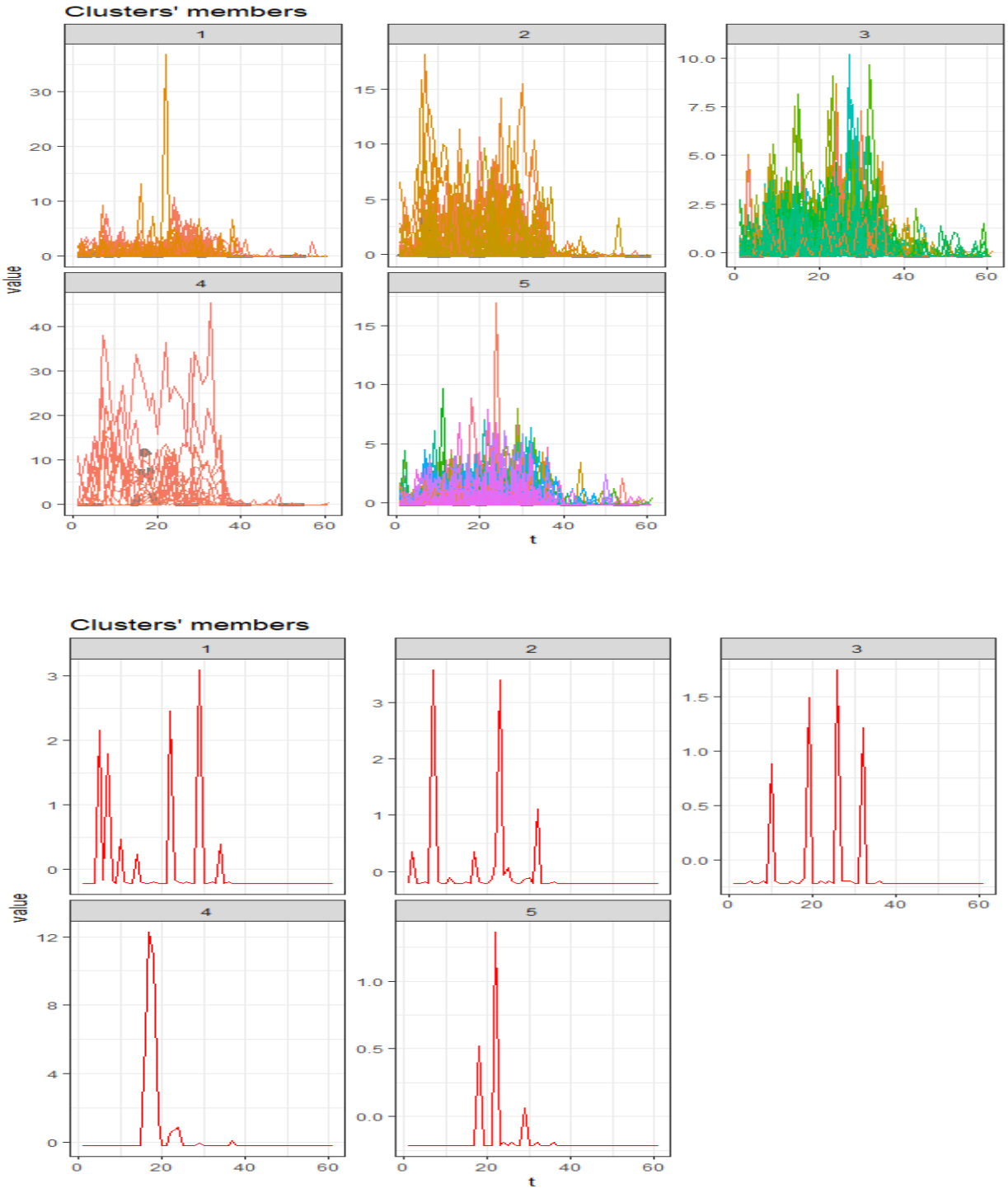


Figure 22. Students Attention-based clusters and centroid plots

```
partitional clustering with 6 clusters
Using dtw distance
Using pam centroids

Time required for analysis:
   user  system elapsed
 502.30   17.66  520.34

Cluster sizes with average intra-cluster distance:

  size     av_dist
1  173   2.0030888
2  138   5.4067923
3    9   3.0485866
4   70 12.7482172
5  295   1.1922101
6  517   0.5097114
```

Figure 23. Students' Participation based time series clustering in R

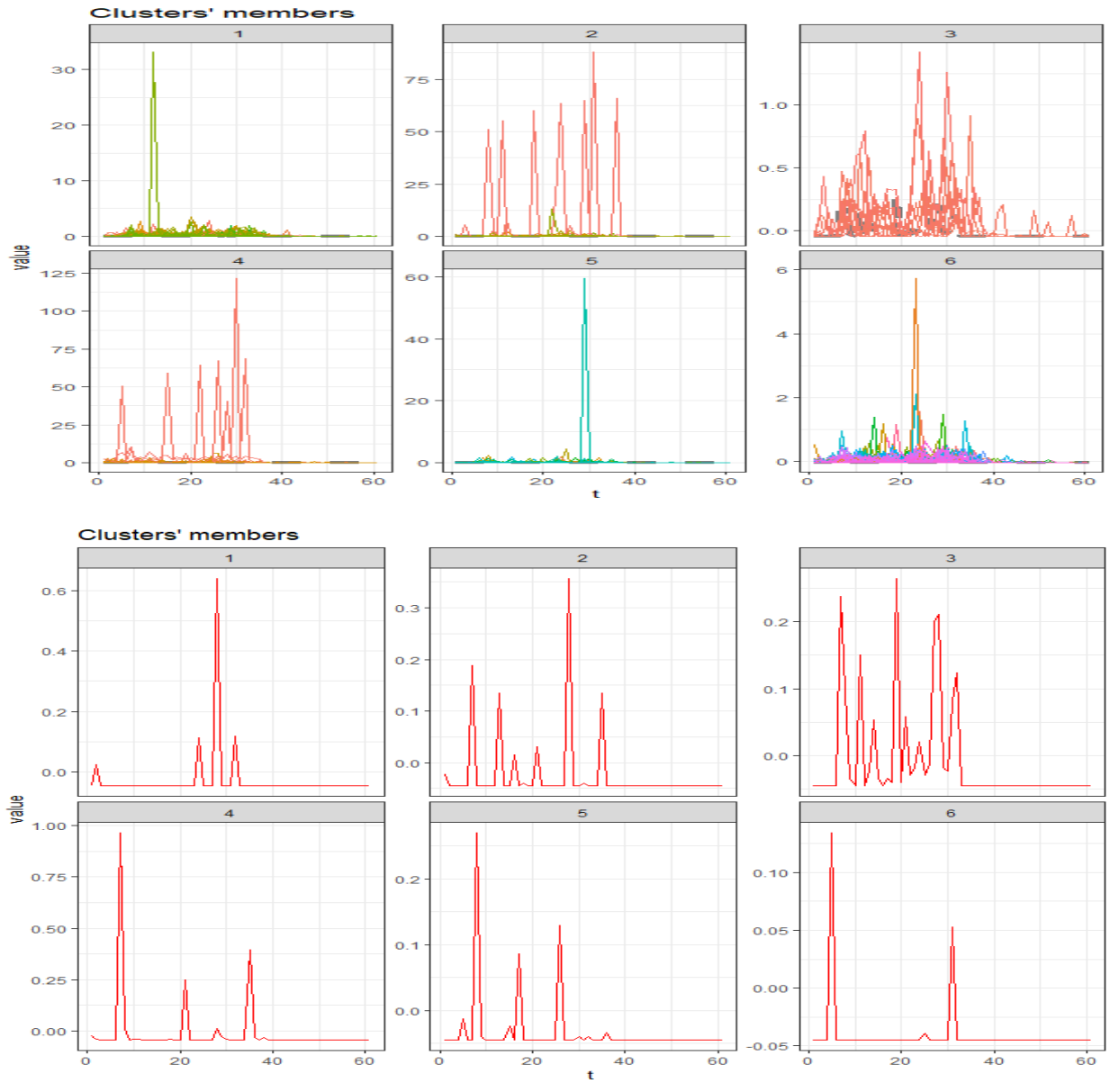


Figure 24. Students' participation-based clusters and centroid plots

Path B Analysis Flow

Step 2B: Capture active students' grades:

Study 2 "A" path analysis was on Attention and Participation dimension. After that, we moved onto analysis path "B", which was related to students' performance/grades. To do that, we filtered active students – as identified in Step 1- from the Moodle file mdl_grade_grade_history.

Step 3B: Performance measurement:

We performed grade-based clustering on two variables:

For the purpose of this study, we performed K-means clustering over two dimensions:

(1) Students' total final grades collected over graded activities; and

(2) Students' total final grade over the maximum grade they could receive over the graded activities as performance.

Step 4B: Performance Based Clustering

K means clustering was performed on the table with two variables. Below is the cluster size and means after K-means clustering was applied (Figure 26):

```
> GradeClustersFGPer <- kmeans(GradeFGandPerfor[,2:3], iter.max = 10, 6)
> GradeClustersFGPer
K-means clustering with 6 clusters of sizes 413, 348, 6, 174, 123, 138

Cluster means:
      FG_SUM       Per
1  432.30024 0.7117019
2   29.42261 0.7430419
3 1165.69671 0.8104574
4  314.59026 0.6062361
5  583.41158 0.7496664
6  175.20488 0.5214123
```

Figure 25. Final grade and performance based static clustering in R

The following are the cluster size and means after K-means clustering was applied over Final Grade and percentage performance dimensions (FGM: Final Grade Mean, Per: Percentage Mean):

i) Cl 1: S=432 FGM=432 PM=0.71

ii) Cl 2: S=348 FGM=29 P=0.74

iii) Cl 3: S=6 FGM=1165 PM=0.81

iv) Cl 4: S=174 FGM=314 PM=0.60

v) Cl 5: S=123 FGM=583 PM=0.74

vi) Cl 6: S=138 FGM=175 PM=0.52

Merging Paths, A and Path B for Alignment:

Step 5: The last step of our analysis was comparing all emerging clusters from different categories (performance, attention, and participation) with each other and aligning them. In Figure 27, the first column shows the performance clusters, the second column shows the attention clusters, and the third column shows the participation clusters. For example, the first-row alignment should be read as such: There are 320 students in the alignment, where students belong to: performance cluster 2, attention cluster 5, participation cluster 6, and so on.

To investigate the alignment, we analyzed all three sets of clusters together. This allowed us to see which clusters of attention and participation were aligned with which performance clusters. In other words, we investigated what patterns of attention/participation can be associated with a pattern of performance. Establishing such associations helped us predict performance based on attention and participation, themselves calculated as aggregated process metrics. We could have associated attention and participation separately from performance; that would have resulted in more associations but would have weakened the possibility of correct estimation, as only one parameter would be considered.

	ıalAllClCor.cluste	inalAllClCor.AttC	inalAllClCor.ParC	Count(*)
	Filter	Filter	Filter	Filter
1	2	5	6	320
2	1	3	5	179
3	1	5	5	133
4	6	5	5	109
5	4	5	6	98
6	1	2	2	60
7	4	3	6	58
8	5	3	2	53
9	5	2	5	37
10	1	1	2	32
11	6	3	2	21

Figure 26. Emerging cluster groups alignments and counts

4.5.1 Cluster Alignments & Interpretations

In this section, we will label the clusters that emerged in the performance-based clustering to better interpret the study outcomes.

Performance Cluster # 1: *"Comfortable".* Cluster grade mean is 432, and size is 413. This group is quite comfortable with the content of the course and managing well.

- **Best Alignments:** Attention clusters 3, 5, 2 and 1 and Participation clusters 5 and 2. Both these Attention and Participation clusters show high-volume activity counts on cluster centroids.

Performance Cluster #2: *"Failing".* Cluster grade mean is 29, and size is 348. This group was not engaged and failing; unfortunately, they were the second biggest group.

- **Best Alignments:** Attention cluster 5 and Participation cluster 6. These cluster centroids showed the least number of spikes and smallest activity volume on their centroids. Specifically, Attention cluster 5 demonstrated these factors during the initial two weeks of the course.

Performance Cluster #3: *"Experts".* Cluster grade mean is 1165, size is 6. This group seemed to find the course content far below their knowledge level as their high grade demonstrated.

- **Best Alignments:** Attention cluster 5 and Participation cluster 5, with the size of 6. It was hard to make a generalization on a group size of 6.

Performance Cluster # 4: *"Average".* Cluster final grade mean is 314, and size is 174. This group is called average to reflect their average grade / performance. Please note that the mean grade for the Comfortable group is 413, Expert group is 1165.

- **Best Alignments:** Attention cluster 5 and Participation cluster 6. We speculate that with more attention and activity actions, this group could have easily increased their grades, please note Attention cluster 5 again.

Performance Cluster # 5: *"Doing very well".* Cluster mean is 583, and size is 123. This group was doing very well with a cluster mean well above the class mean yet lower than experts which has the mean grade 1165.

- **Best Alignments:** Attention clusters 2 and 3 and Participation clusters 2 and 5. These Attention and Participation clusters showed evenly distributed multiple spikes with steady on-going activities over the course duration.

Performance Cluster #6*: "Struggling"*. Cluster grade mean is 175, size is 138. Instructors may have liked to provide customized help to this group. 109 students were in this trouble grade-based group.

- **Best Alignments:** Attention cluster 5 and Participation cluster 5. Participation cluster 5 is showing high volume of activity counts. We see Attention cluster 5 in this low performance alignments again.

With a unique characteristic, Attention cluster 5 is coming up on Performance cluster 6 alignment segment. Attention cluster 5 centroid shows minimal to no activity in the early period of the course. Although Attention cluster 5 pairs with Participation clusters 5 and 6 in alignment activity, these participation clusters did not demonstrate such unique characteristic differences compared to other participation clusters allowing us to identify the difference. This observation suggests that "Attention" activities in the early period of the cluster may provide a stronger indication of students' engagement and performance in the course, more so than participation activities.

4.6 Discussion

4.6.1 Reflection on Findings

The approach of logically grouping students' interaction data on an LMS as "Attention" and "Participation" provided the opportunity to analyze data with respect to two different aggregated metrics/dimensions.

By creating different profiles (clustering) based on students' attention and participation interaction frequencies, we were able to align students' performance-based profiles. We have seen Failing and Struggling clusters, 2 and 6 align with Attention cluster 5. This analysis provided evidence on two related but different topics.

- As a general educational subject, we offered supporting evidence that students' attention (e.g., identifying content, checking administrative information) in the early phase of a course might be a better indication of their future success than participation activities (e.g., submitting mandatory feedback to a message).

- For the specific subject of using VLE process metrics, we demonstrated that combining basic metrics into aggregated ones and using clustering of their time series can be used for assessing and predicting academic performance.

Analyzing students' "Attention" data might provide valuable information in terms of providing information on content efficiency and students' overall interest. In this way, instructors may choose to follow students' interests in real-time and provide more similar content or bring students' attention to the content that did not get much attention.

Our study provided initial responses to three research questions on basic metrics, aggregated metrics, and the use of time series and clustering as an analysis method. While basic metrics are too specific and low-level to be clearly associated with HOTS, the study showed that they could be used for HOTS assessment through aggregated items such as attention and participation. Our study also demonstrated that there is a potential correlation between clusters of attention/participation and academic performance. This means those aggregated metrics and their time series can be used as indicators of academic performance and learning, and that time series combined with clustering is a potentially valuable analysis method for HOTS assessment; i.e., once we identify to which cluster a student's attention and participation belong, we can make assessment and predictions for their academic success and needs.

4.6.2 Limitations and Further Research

Our study had its own limitations that require further research to fully answer the research questions. We used the alignment of attention/participation clusters with academic performance to show the possible associations, and we used the number of common students as a sign of alignment of two clusters (same students belonging to the clusters). This is a limited and simplified approach. Further analysis is needed to establish a proper correlation between process metrics and performance, but we believe the results are enough to suggest the potential. Also, our time series in Study 2 were only related to the changes in values of one item over time. HOTS generally involve a process of multiple actions performed sequentially to solve a problem. Further research is needed to include an analysis of such time series and offer feedback to students on different elements of their actions. Finally, it should be noted that this research data was collected on a non-credit, volunteer-based course. It is highly recommended to run similar research on a credit course to compare the results.

Our next study (Chapter 5) investigates HOTS assessment in an experiential situational learning task performed in a 3DVLE. This study followed the idea of combined process metrics for computer-based observation and assessment that was developed through Studies 1 and 2, and applied it to 3DVLE. We also used a more systematic correlation analysis between process metrics and instructor evaluation to address the shortcoming of Study 2 regarding the alignment mentioned above.

5 Study 3: Score-Based HOTS Assessment in a 3DVLE

In Study 2, we demonstrated the use of combined/aggregated metrics. Study 3, described in this chapter, extended the notion of combined process metrics and time series to 3DVLEs that offer a wider range of data. We also extended the use of time series to sequences of multiple actions as opposed to different values of one item over time. This allows us to see and study the pattern of behaviour that can be associated with HOTS. Study 3 divided an educational activity into smaller, yet meaningful, elements referred to as *motifs*. The use of motifs allowed us to create more flexible assessment methods by combining sequential actions into a single item. Motifs can be defined through a collection of different actions based on the instructor's judgment, can have different lengths, and rather than success/fail results, they can offer more specific feedback to students on parts of a task. In Study 3, we used score-based assessment by assigning scores to motifs, followed by a correlation analysis of motif scores vs. instructor evaluation. A series-based alternative was the subject of Study 4.

The findings of our Study 3 were reported in IEEE CIVEMSA conference, as listed in Section 1.4.

5.1 Overview

The concept of motif as a small yet meaningful set of activities has been suggested by Gibson & Freitas (2016) but not properly investigated in existing literature as an assessment tool. After investigating aggregated basic metrics over the full course in the previous study, we concluded that we need to investigate sequences of different actions, (1) to associate them to HOTS and (2) to offer more granular feedback. Motif-based assessment has the potential to help in two areas: the lack of flexibility in assessment design as motifs can be defined using different action combinations and have different assessment methods, and the lack of more granular feedback to students as they can be focused on parts of learning activity. Using motifs allows us to address the issue of explainability raised by some researchers (Conati et al., 2018; Loh, 2015). Small and granular assessment items can be focused on specific parts of a learning task and so offer feedback on each part, to make the assessment result easier to interpret, explain, and understand. Study 3 focused on the use of motifs and used a 3DVLE to allow for a richer set of metrics. We continued the concept of comparing computer-based assessment with academic performance but used a more

systematic correlation approach, as suggested by our Study 2 findings. Study 3 was performed in the context of an English as a Second Language (ESL) course for international students using the same Carleton virtual environment as in Study 1.

This study used a score-based assessment on small learning motifs to evaluate students' HOTS. To start, we defined the following HOTS, specified by the instructor as relevant:

- Communication
- Learning from mistakes
- Engagement
- Drawing conclusions

Six virtual chambers (areas) were designed with various tasks all involving the above skills. Motifs were defined as specific actions corresponding to those skills. The primary goal was to assess above skills through the calculated process metrics, and compare themto the instructor's assessment, which was done holistically and subjectively. We also included the final grade of the students as a data item to investigate any correlation with the skill assessment. Due to technical limitations, our virtual environment was not capable of recording some events, such as reading an information board. This resulted in manual observation and recording, and then manual scoring.

Correlation analysis was performed on the following pairs of the above data:

- Instructor's subjective assessment of the skills (through watching students' videos) vs. basic (individual) metrics from the environment.
- Instructor's subjective assessment of the skills (through watching students' videos) vs. assessment of the skills using the motif-based scores.
- Year-end grade vs. basic metric.
- Year-end grade vs. motif-based scores

The study used a learning maze and learning challenges designed as chambers for an advanced level English for Academic Purposes (EAP) course at Carleton University, running on *Carleton Virtual* 3DVLE used in Study 1. This environment was selected as it was the most easily available option and was being used by our partner instructor. As per the Carleton Virtual design objective, it is flexible to be used in multiple courses and purposes. The course was originally in-

person but later added online components as per the course instructors' choice. This study focused on the online part, only. The online component was divided into six tasks, each incorporating the above four skills (and their corresponding motif) at different levels as described in Section 5.3. Different combinations of basic metrics within each motif were used to represent the related HOTS, as guided by the instructor.

Our research questions for this study were as follows:

RQ1d: How can small yet meaningful series of process metrics (motifs) be used for HOTS assessment?

RQ2b: How can scoring elements of a motif be used to assess HOTS?

Peggy Hartwick was the instructor of the EAP course, and learning tasks were created collaboratively by Peggy Hartwick and Nuket Nowlan. Peggy Hartwick also performed the video-based assessment. Nuket Nowlan was the designer of the 3DVLE and Study 3, and she performed the data analysis via manual and log-based investigation and executed the correlation-based analysis. Similar to Study 2, we used our guidelines for good use of data by applying well-explained algorithms and directly related data, and investigated the use of our findings only to serve students themselves.

5.2 Participants

Students in the EAP course were given the option to complete either the 3D virtual learning task or an alternative learning task, such as an in-person presentation. As such, the number of study participants was limited.

The study was approved by Carleton University Research Ethics Board. As per the approved ethics protocol, to make sure that students did not feel forced to participate in the study, a 3rd party individual collected the study participation consent forms until after the course was over and all students' grades had been submitted. At that point, data analysis was conducted using only the data belonging to consenting participants.

Of the 10 students in the course, eight volunteered to participate in the study. Participants were from China (N=4), Middle East (N=2) and Indonesia/Philippines (N=2). Their fields of study included business, computer science, history, architecture, mechanical engineering, and biology.

The small sample size was a limitation of the study in terms of generalizability and also because statistical analysis on demographic data could not be performed.

5.3 Materials

The Carleton Virtual environment was used in this study to facilitate the learning activity. Carleton Virtual is described in detail in section 3.3 (Study 1). As a reminder, the Avaya Engage platform includes features such as real-time document up-loadable boards and interactive clickable objects. The platform also captures user gestures, interactions, talk-time, and location in a log file.

Study 3 was conducted specifically in the *Orientation Maze* part of the virtual campus. The three-dimensional, self-regulated learning path employs instruction boards to teach students about typical features of the VLE platform. The instruction boards provide explanations about related features and suggest activities to students to improve their understanding (Figure 28).

Figure 27. Experiential learning activity in Carleton Virtual

At the end of each chamber of the path, there is a related puzzle that students are recommended to solve to advance to the next section. Students are able to advance without solving the puzzle by jumping over a gate, if they so choose. The maze has six chambers, each of which requires students to (1) read some instructions, (2) solve a puzzle to get a code, and (3) use the code to open the exit door and reach the next chamber. Examples of puzzles include using the zoom feature to see a bird on a mountain and selecting the bird to open the chamber door. By the

end of the maze, students will have successfully performed the most common features of the 3DVLE platform, thereby reducing technical and spatial barriers in future activities.

The following are the puzzles/challenges in each chamber that were used as motifs in this study to evaluate students' HOTS:

Chamber 1:

- Read the instructions on the three boards placed around the chamber, and follow the given instructions:
    - Walk backwards and turn on two lights with special triggers.
    - Walk sideways.
    - Jump over the door using special instructions provided on the last board.

Chamber 2:

- Read the instructions on the board, practicing zooming in and out:
    - Zoom in on a mountain far away to see the picture.
        - After zooming in on different places and reading the instruction on those, students see the bird on a mountain far away to see the picture.
    - Click on the bird that was seen in the last activity to open the door.

Chamber 3:

- Read the instructions on the board:
    - Open the web browser in the environment and follow the instruction to get a file, download and open it.
        - The file just downloaded will give a passcode to access new instructions in a specific location.
    - Open the door using the new instruction.

Chamber 4:

- Read the instructions on the board:
    - Open the web browser in the environment and follow the instruction to get a file, download and open it.
        - The file just downloaded will give dress code instructions to use for avatar customization.
    - Get a screenshot and upload your dress on a file on the environment dropbox to receive a code.

- Open the door using the code.

Chamber 5:

- Read a story to figure out how to send a text to another person in the environment.
  - Send a private text message and receive a quote.
  - Through searching on the internet, find to whom that quote belongs.
  - Open the door using the name of the person as the code.

Chamber 6:

- Read the instructions on the board:
  - Teleport into a room in Carleton Virtual.
  - Read the board in the new room to get a code
  - Open the door and complete the maze.

Correlation analysis was performed using R statistical tool.

5.4 Procedure

During the study, students were provided with information to log into Carleton Virtual. They found themselves at the beginning of the Orientation Maze with instructions to look at a series of boards posting guidance about how to solve problems by themselves. All of the students' interactions were collected within the log file by the system; the students also recorded themselves throughout the learning session.

During the data analysis phase, four different types of data were analyzed:

1. Students' activity-based metrics (collected by the system), such as movements and triggering events
2. Additional metrics related to students' activities that were not collected by the system automatically (collected by the researchers). The technology platform for Carleton Virtual at the time of this study (Avaya Engage) was limited in terms of data collection. As such, some activities such as "reading a board" were not detected by the system. The research team "simulated" these metrics by watching videos of the students' activities in the virtual world and identifying these metrics. All students' sessions were screen-captured, and videos were available.

3. Students' holistic HOTS assessment given by the class instructor by watching students' videos. Appendix C shows an example of rubrics used by the instructor. These assessments were done for the whole experience and on four skills of Communication, Engagement, Learning from Mistake, ad Drawing Conclusions.
4. Students' year-end grades.

Group 1 above (system metrics collected by the VLE platform) were extracted from the VLE log file for analysis due to their relevance to required tasks:

- Spoke to other (sec);
- Change volume (location) and unique volumes visited.
- Gestures performed during the activity.
- Total triggers used; and
- Unique features used (calculated through a script).

The groups 1 and 2 data were used by the researcher to calculate the scores for four skills based on VLE data. Examples of group 2 data included:

- Reading the instruction board (for 10 seconds, minimum);
- Opening the door successfully.
- Advancing by jumping over the gate as the opening was unsuccessful

Following the guidance and information provided by the Canadian Partnership Assessment Index as guidance on 21st-century competencies[13], the following overlapping skills were observed and assessed through a combination of metrics as defined by the instructor:

- *Communication*: "Reading the instruction board" plus "performing the activity suggested on the instruction board".
- *Engagement*: "Performing the activity suggested on the instruction board"

[13] http://www.c21canada.org/wp-content/uploads/2015/04/Shifting-Minds-LEARNING-INDEX-Sept-29.pdf

- *Drawing conclusions*: "Performing the activity suggested on the instruction board" minus "Advancing by jumping over the gate." All chambers ended with a door that the student needed to learn how to open. Jumping over the door indicated wrong decision and cheating)
- *Learning from mistakes*: "Experimenting with the activity" plus "Going back to re-read the instruction board if the door opening failed" minus "Advancing by jumping over the gate."

Each skill was associated with a motif and scoring was done by giving equal weight to each activity within the skill. For example, in Chamber 6, the student had to read the instructions on the board, teleport into a room in Carleton Virtual, read the board in the new room to get a code, open the door and complete the maze. Each of these activities would get 25% to make up 100% for the Communication skill/motif that involves reading the board and performing the activity.

Students' holistic HOTS assessment and year-end academic grades were provided by the course instructor.

Note that the point of our study was not to assess the validity of the skills and how they were defined, but to show that an assessment based on process metrics could be valid compared to what the instructor does. As such, the process metric-based assessment followed what the instructor defined as skills. We realize that our four essential skills could be defined in many other forms.

5.5 Results

During the data analysis phase of the study, R correlation matrix data analysis was performed on all collected data, including both basic and process metrics with the instructor's observation-based skill assessment. The results of the analysis are shown in Table 3. The correlation analysis found no strong relationship between year-end grade or the instructor's observation-based assessment and the VLE-collected basic metrics. The highest and only notable relationship is between year-end academic grade and the number of unique (the number of different types of platform features) features used on the platform. As the objective of the learning activity was learning the VLE platform's features, the results could be interpreted as academically strong students achieving the lesson objective. More research should be done to confirm.

Table 3. Correlation values for VLE metrics vs. year-end grade and instructor skill assessments

	Year End Grade	Communication	Engagement	Learning from Mistake	Drawing Conclusion
Spoke to Others	0.20	-0.13	-0.14	-0.45	-0.17
Entered New Areas	-0.06	-0.33	-0.49	-038	-0.05
Unique Areas Entered	-0.39	-0.13	-0.18	-0.51	0.27
Time Spend	-0.11	0.31	0.14	-0.83	-0.22
Gestured Performed	-0.27	-0.12	-0.36	-0.41	-0.01
Unique Features Used	**0.64**	0.18	0.15	-0.4	0.17
Calculated Communication	0.18	**0.92**	0.87	0.33	0.78
Calculated Engagement	0.33	0.70	**0.90**	0.24	0.65
Calculated Learning from Mistake	0.46	0.30	0.34	0.09	0.35
Calculated Drawing Conclusion	0.41	0.64	0.70	0.18	**0.87**

On the other hand, the motif scores (shown as Calculated skills) showed a high correlation with the instructor's assessments and provided us with insight into the metrics that were more likely to be useful for assessment. The study provided strong evidence that learning process metrics are better indicators in terms of providing insight into students' HOTS. Correlation values r=.92, r=.90, r=.87 were found between the instructor's assessment versus process metric-based calculated assessments for Communication, Engagement, and Drawing conclusions, respectively. No strong correlation was identified for Learning from mistakes (calculated vs. instructor assessment). This was due to a script error that affected the order of activities in scoring and unfortunately, we were not able to correct it later as we lost some of our data. The data for **Calculated Learning from Mistake** should be ignored.

The motifs did not show any significant correlation with year-end grades. This makes sense as the final grade depended on many other factors, such as the language skills, and HOTS

development was only one component of the course objectives. It is worth noting that the highest correlations for year-end grades were between calculated Learning from mistakes (r=.46) and instructor-assessed Draw conclusions (r=.39)

5.6 Discussion

5.6.1 Reflection on Findings

Our findings support the notion that while basic metrics are essential in understanding what the user does, they cannot be used by themselves as they are too specific to have a meaningful relation to HOTS. This helps answer research question 1b about basic metrics. Further, we showed that small groups of basic metrics can be defined as motifs for HOTS assessment by combining basic metrics that are part of a skill (research question 1d), and that scoring these motifs based on the metrics they include can be used for assessment (research question 2b). This approach can make assessment more flexible and easier to explain and understand.

We found that some specific process metrics that are not commonly provided by many 3DVLEs (such as "reading a board") are crucial in HOTS assessment and should be included in basic VLE metrics to allow automated assessment. We have also identified that the duration of the interaction is also important to give meaning to interaction, which also should be captured when basic metrics are logged. For example, "looking at an object in an environment" can be interpreted as a casual passing by when it is just a couple of seconds or can be interpreted as an investigation if it takes around a minute. The ability to group metrics by instructors to define a motif is also an essential feature that is missing in many 3DVLEs —for example, time looking at a board plus performing the suggested action or failing to solve the problem plus going back to a board to re-read it.

Overall, Study 3 provided supporting evidence that 3DVLE-collected metrics can form a combined process metrics over motifs and can be used to assess learners' HOTS through score-based stealth assessment.

5.6.2 Limitations and Further Research

The lack of proper 3DVLE features (such as some data items and grouping) caused a major limitation of this research and resulted in manual scoring. While manual operations during the research are not unusual, further research is needed to develop and use proper automated built-in facilities in 3DVLE (see Study 4). Scoring the elements of a motif, as discussed in Chapter 2, has

the disadvantage of not considering the order of activities, as we saw in the case of **Calculated_Learning_from_Mistake**. Series-based assessment is a preferred (yet more complicated) method in this regard, which was also considered in Study 4. The small number of participants was also another limitation of this study that made it more of an exploratory effort. Choosing a more appropriate course and assessing all required skills could result in a better comparison with the year-end or mid-term grades.

Our Study 4 aimed at addressing some of the above limitations through the development of a 3DVLE prototype that had the required data collection abilities and also the use of series-based assessment after investigating and developing skills and tools for pattern matching and similarity analysis.

6 Study 4: Series-based HOTS Assessment in a 3DVLE

Our Foundation Study (Study 1, Chapter 3) suggested that 3DVLE platforms have the potential to offer curricula effective at fostering HOTS due to their specific affordances. Studies 2 and 3 provided evidence that using learners' full interactions and using combined metrics (aggregated or motifs) can lead to effective assessment and prediction of learning and HOTS. Time series clustering and score-based motif assessment were the methods employed in Studies 2 and 3, respectively. Findings of these two studies suggested that (1) a variety of data items are needed as process metrics that are not commonly available in VLEs, and (2) motifs, while useful, are patterns of behaviour and the order of actions in them matters; meaning that a purely score-based assessment fails to assess the related skills.

In Study 4, our objective was to perform HOTS assessment procedurally over motifs through defining a single HOTS for each motif as in Study 3 but using series-based assessment. We used the similarity index approach as the most common way of comparing time series and expert data as the target action. Our objective is to investigate the approach of identifying separate motifs in a learning curriculum and perform assessment for each to provide more granular feedback to students on their weaknesses and strengths.

Initial findings of our Study 4 were reported in IEEE VR conference (poster session), as listed in Section 1.4, and a more comprehensive paper is being prepared.

6.1 Overview

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 on performance. In Chapter 2, we cited several analysis methods working with series. The most notable ones are:

i) Time series clustering approach for profiling and future prediction (Cavalli-Sforza, L.L., 1965, Peffer, 2019). The downside of this method is that it requires a lot of data before it can be used. Study 2 in this dissertation used this approach.

ii) Series similarity measure-based analysis (Loh and Sheng, 2014, 2015; Sawyer R., Rowe, J., Azevedo, R., & Lester, J. 2018). 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 final study, we contributed to series-based assessments by using motifs and investigated different similarity indices for each motif. We considered this approach to be suitable for HOTS assessment for two reasons: (i) no need to collect enough data to train and create a model, as an expert could record expert path/s to be used for similarity analysis; and (ii) a comparison between learners' performance sections to expert' would provide actionable insight and development points to learners. We hoped to investigate an assessment approach that is flexible and explainable (Fiok et al., 2021). Classroom teachers, when they need to or are challenged to, should have the ability to dive into learners' captured activity sequence and provide explanations to students on specific areas needing improvements and interpretations of how the analysis works. We believed that the method of similarity index-based assessment on motifs would be more explainable than some blind machine learning-based assessments (especially deep learning), as they can offer feedback on small motifs.

Study 4 was performed in the context of a chemistry lab in a 3DVLE. 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.

In this study, we aimed to answer the following research questions:

RQ1d: How can small yet meaningful series of process metrics (motifs) be used for HOTS assessment?

RQ2c: How can similarity analysis between student and expert motifs be used to assess HOTS?

RQ2d: Which similarity indices are more effective to assess HOTS?

Figure 28. A student testing chemical lab on the chemistry app developed for Study 4

Figure 29 above shows a student testing the Study 4 chemistry app. A collaborative research study team was formed under the supervision of Dr. Ali Arya to study the effectiveness of 3DVLE-based chemistry lab safety training, including PhD student Nuket Nowlan (assessment research), PhD student Hossain Sam Qorbani (app development), and PhD student (now graduated) Maryam Abdinejad (chemistry subject matter expert). The roles of each team member were as follows: Sam Qorbani developed a standalone single user 3DVLE using the Unity Game Engine, Maryam Abdinejad acted as the Subject Matter Expert (SME), and Nuket Nowlan performed the HOTS assessment analysis.

Similar to previous studies, we followed our guidelines for the good use of data.

6.2 Participants

The research team invited university students to participate and collect research data in the Winter 2021 semester. The study was approved by Carleton University Research Ethics Board. 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 in 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 as before the pandemic.

Ultimately, we recruited 36 university participants to participate in Study 4. Of the 36 participants, 20 were male and 16 females. 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 (Maryam Abdinejad), 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. Study participants' age range is between 20 to 52 with median range is 22.50.

6.3 Materials

Our ScienceVR 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 environment had three areas for training and testing purposes. 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.

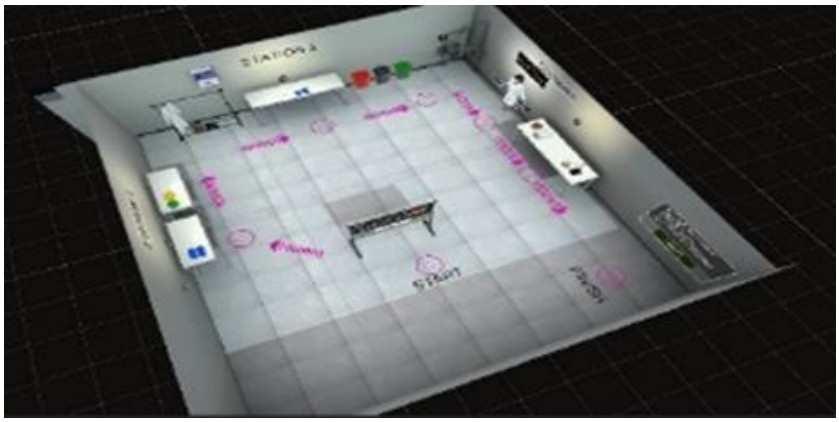


Figure 29. Training Area one

Area one: This area was designed to provide basic training activities such as picking up simple objects (cubes, sphere), which enabled participants to learn how to use touch controllers (Figure 30). 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.

Figure 30. Training Area two & three

Areas two and three: While area one of the lab 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 (Figure 31, 32), 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.

As discussed in the previous chapter, one of our biggest challenges in using Carleton Virtual was collecting activity metrics. Carleton Virtual was using Avaya Engage platform when we were running Study 3. At the time of Study 4, Carleton Virtual was using Virbela platform, which had even more limited data collection abilities. Therefore, a new standalone app was created using the Unity engine that could capture all student's interactions within the VR space, record the order and duration of each, and email the data to the researcher 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.

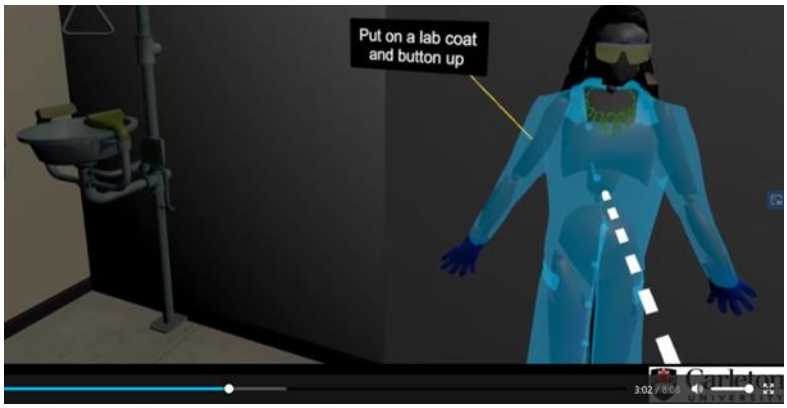


Figure 31. A screenshot from the Area two

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).

The study included expert data provided by the instructor.

6.4 Procedure

Overall, the following procedure was employed for collecting and analyzing data in Study 4:

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.
    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 are 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 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.

A list of all the activities that can be performed in all three areas can be found in Appendix B.

Table 4. 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 • After 2 hours • STOP the reaction with: 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 • WAIT TO LET IT COOL DOWN
Example of an emergency in VR lab and responsive actions	• Put a round bottom flask inside the oil bath WITHOUT clamping it • In the event that the flask falls inside the oil bath and its solution will be poured into the bath and causes fire • The student should unplug the hotplate first • Pull down the hood door/glass and let the oil-bath completely cool down • Pull up the hood door again • Remove the hotplate and pieces of broken glass from the fume hood • Place the hotplate on the bench • Dump the pieces of broken glass into the red bin • Clean up the fume hood using napkins • Use acetone to clean up the oily fume hood

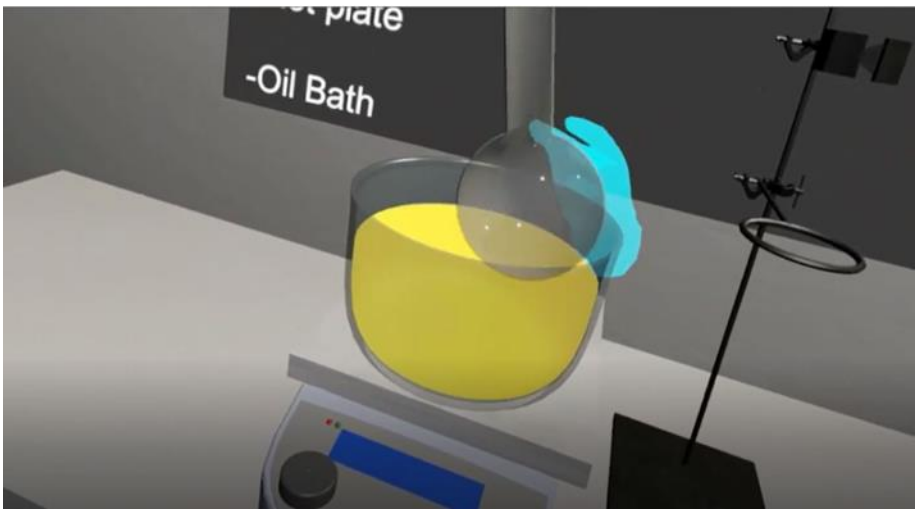


Figure 32. Student performing an experiment in three-dimensional virtual chemistry lab

A screenshot of a student performing an experiment is shown in Figure 33. The system log in Figure 34 shows a sample of the digital activity trace that was collected during students' learning session:

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

Figure 33. Sample of the data captured in the application audit file

6.4.1 Motif Identification for HOTS Assessment

Loh (2015) and colleagues 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 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 Study 3, we defined motifs as an overlapping combination of activities that build four separate skills. These skills existed in 6 chambers that students visited. For Study 4, the instructor was looking for three skills that could be identified within three separate tasks. There was no longer any overlap (shared activities, as in Study 3) 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.

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 and explored the lab equipment's components and practiced using the components. The main HOTS in use in this section was Information Collection.

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 Decision Making.

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

- Information Collection
- Critical Thinking
- Decision Making

Data points (basic metroics) were collected on each of these motif's for each student, 135 for information collection, 104 for critical thinking, and 70 for decisionmaking. 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

6.4.1.1 Identifying Activity Series Gram Number

Action series are orderly data points 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 uni-gram (n= 1), bi-gram (n= 2), tri-gram (n= 3), four-gram (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 unigram and bigram 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

6.4.1.2 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 Study 4, 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.

6.4.1.3 Similarity Index Selection

Maximum Similarity Index (MSI) is a term proposed by Loh and Yanyan (2014), as the similarity index that gives the best match, to study the performance of a player in games. Figure 35 shows how the Jaccard index can be used to evaluate players' performance level.

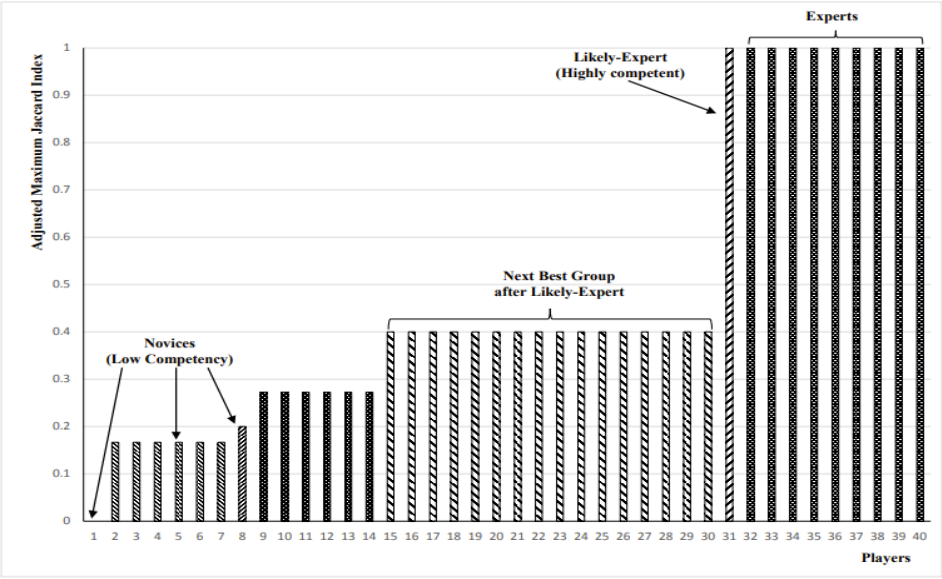


Figure 34. Players ranking as per Jaccard similarity index (Loh, Yanyan, 2014)

In experimenting with different similarity indexes as a performance measure, Loh and Yanyan (2015) used multiple similarity indexes to profile players' playing styles. Figure 36 shows the different indexes they used. They concluded that a combination of multiple index-based similarity measurements provides the best classification in terms of profiling players' playing styles.

Table 1
Formulas for Dice, Jaccard, Overlap, Cosine, and Longest Common Substring coefficients.

	Similarity coefficient	Formula
(a)	Dice coefficient *(Dice)*	$Dice(A,B) = 2\frac{\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$
(b)	Jaccard coefficient *(Jac)*	$Jac(A,B) = \frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$
(c)	Overlap coefficient *(OVL)*	$OVL(A,B) = \frac{\lvert A\cap B\rvert}{\min(\lvert A\rvert,\lvert B\rvert)}$
(d)	Cosine coefficient *(Cos)*	$Cos(A,B) = \frac{A\cdot B}{\lVert A\rVert\cdot\lVert B\rVert}$
(e)	Longest Common Substring coefficient *(LCS)*	$LCS(A,B) = 1 - \frac{d_{LCS}(A,B)}{d_{max}(A,B)}$

Figure 35. Popular similarity index formulas

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, 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 the similarity to master (MSI) by Loh et al. (2015), 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:

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

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 (Foreman, 2014). Figure 37 shows the way two texts are carried over to a vector space.

$$Cos(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Conversion from a string to a vector.

Vector space	S	T	O	P	I	N	G	
A: STOPPING	1	1	1	2	1	1	1	
B: POPPINS	1	0	1	3	1	1	0	
	1 x 1	1 x 0	1 x 1	2 x 3	1 x 1	1 x 1	1 x 0	$A \cdot B = 1+0+1+6+1+1+0 = 10$

$\|A\| = \sqrt{(1^2+1^2+1^2+2^2+1^2+1^2+1^2)} = \sqrt{10}$,

$\|B\| = \sqrt{(1^2+0^2+1^2+3^2+1^2+1^2+0^2)} = \sqrt{13}$, and

$\|A\| \cdot \|B\| = \sqrt{10} \times \sqrt{13} = 11.402$

Figure 36. Moving STOPPING and POPINS series into vector space (LOH, 2016)

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 less operations required to transform one series to another means more similar the series are.

There are different ways of applying an edit base 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 higher cost to some of the operations, (iii) allowing transposition of two adjacent elements (Bard. 2007), (iv) banning substitution from the operation list (Bergroth at.al., 2000), and (v) applying only transposition in adjacent nodes (Jaro, 1989)

We decided to use the basic Levinshtein algorithm, where operations 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 to another through edit operations. Due to this approach, the Levenshtein similarity index is highly sensitive to the order of the series.

6.4.2 Data Analysis Method

As captured in research questions 2c and 2d, we aimed to investigate whether a metric-based similarity analysis could be used for assessing students' HOTS. 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 log:** 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. Figure 38 shows a sample of the audit file received from the app.
   b. *Creating motif-based sections*: Motif creation for this study is already explained in section 6.4.1.

2- **Deciding the similarity index:** As mentioned in 6.4.1.3, 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 bigram and unigram series.

6.5 Results

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

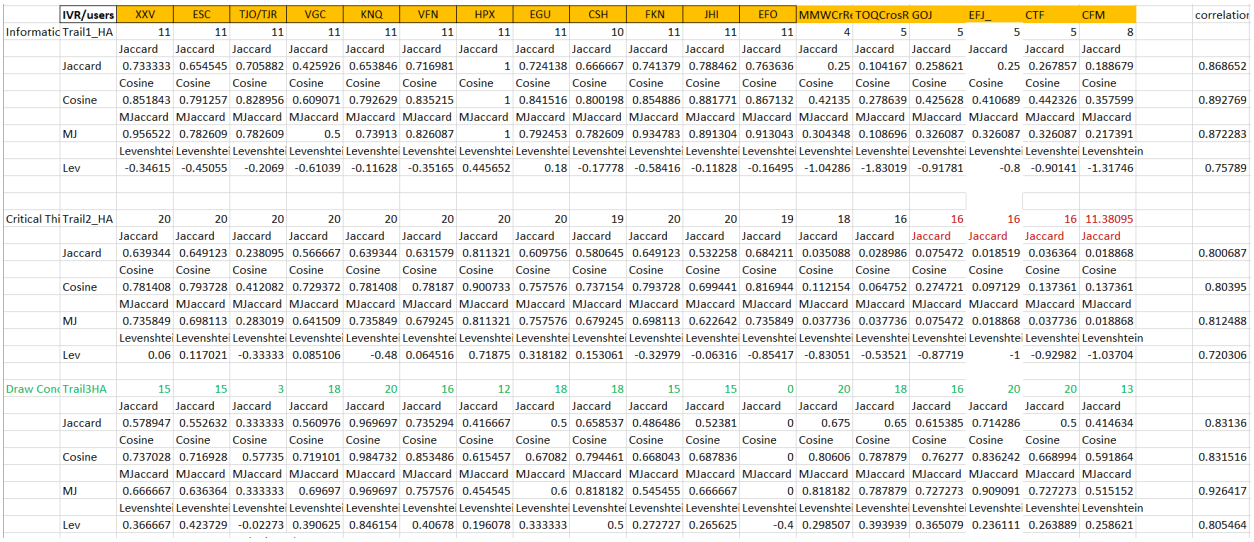
summarized steps were implemented, which were explained in the previous sections and are repeated here for completeness:

- 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 unigram and bigram series.

2:26	Sphere	Select	Left Hand	1.03
2:27	Sphere	Select	Right Hand	0.92
2:28	Sphere	Hover	Right Hand	0.53
2:28	Sphere	Select	Left Hand	7.66
2:42	Quit Button	Hover	UI Pointer	0.18
2:42	Reset Button	Hover	UI Pointer	0.32
2:52	Sphere	Select	Right Ray Interactor	0.54
2:57	Sphere (2)	Select	Right Ray Interactor	0.78
3:03	Cube (5)	Select	Right Ray Interactor	0.72
3:08	gogglefinal4 (3)	Hover	Left Ray Interactor	4.59
3:08	HelpBtnGoggles	Hover	Right Ray Interactor	5.1
3:14	gogglefinal4 (3)	Hover	Right Hand	0.75
3:13	Glove (1)	Hover	Right Hand	1.43
3:14	gogglefinal4 (2)	Select	Right Ray Interactor	3.83
3:18	gogglefinal4 (2)	Hover	Left Hand	0.82
3:18	gogglefinal4 (2)	Hover	Right Hand	0.82
3:18	Glove (2)	Hover	Right Hand	0.68
3:14	gogglefinal4 (3)	Hover	Left Ray Interactor	0.75
3:14	gogglefinal4 (3)	Hover	Left Hand	0.75
3:19	Glove (1)	Select	Right Ray Interactor	1.22
3:20	Glove (1)	Select	Left Hand	1.09
3:21	Glove (1)	Select	Right Hand	0.95
3:21	gogglefinal4 (3)	Hover	Left Hand	0.95
3:22	Glove (1)	Select	Left Hand	0.57
3:23	Glove (1)	Select	Right Hand	0.61
3:23	gogglefinal4 (3)	Hover	Left Hand	0.61

Figure 37. Audit file output from study application

We performed a correlation analysis on three different filtered series from the full series where students demonstrated the following HOTS: *information gathering, critical thinking, decision making* (Figure 39). 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 SME to associate each phase with one specific HOTS that was most relevant.



	IVR/users	XXV	ESC	TJO/TJR	VGC	KNQ	VFN	HPX	EGU	CSH	FKN	JHI	EFO	MMWCrR	TOQCrosR	GOJ	EFJ_	CTF	CFM	correlatior
Informatic	Trail1_HA	11	11	11	11	11	11	11	11	10	11	11	11	4	5	5	5	5	8	
		Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	
	Jaccard	0.733333	0.654545	0.705882	0.425926	0.653846	0.716981	1	0.724138	0.666667	0.741379	0.788462	0.763636	0.25	0.104167	0.258621	0.25	0.267857	0.188679	0.868652
		Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	
	Cosine	0.851843	0.791257	0.828956	0.609071	0.792629	0.835215	1	0.841516	0.800198	0.854886	0.881771	0.867132	0.42135	0.278639	0.425628	0.410689	0.442326	0.357599	0.892769
		MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	
	MJ	0.956522	0.782609	0.782609	0.5	0.73913	0.826087	1	0.792453	0.782609	0.934783	0.891304	0.913043	0.304348	0.108696	0.326087	0.326087	0.326087	0.217391	0.872283
		Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshtein	
	Lev	-0.34615	-0.45055	-0.2069	-0.61039	-0.11628	-0.35165	0.445652	0.18	-0.17778	-0.58416	-0.11828	-0.16495	-1.04286	-1.83019	-0.91781	-0.8	-0.90141	-1.31746	0.75789
Critical Thi	Trail2_HA	20	20	20	20	20	20	20	20	19	20	20	19	18	16	16	16	16	11.38095	
		Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	
	Jaccard	0.639344	0.649123	0.238095	0.566667	0.639344	0.631579	0.811321	0.609756	0.580645	0.649123	0.532258	0.684211	0.035088	0.028986	0.075472	0.018519	0.036364	0.018868	0.800687
		Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	
	Cosine	0.781408	0.793728	0.412082	0.729372	0.781408	0.78187	0.900733	0.757576	0.737154	0.793728	0.699441	0.816944	0.112154	0.064752	0.274721	0.097129	0.137361	0.137361	0.80395
		MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	
	MJ	0.735849	0.698113	0.283019	0.641509	0.735849	0.679245	0.811321	0.757576	0.679245	0.698113	0.622642	0.735849	0.037736	0.037736	0.075472	0.018868	0.037736	0.018868	0.812488
		Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshtein	
	Lev	0.06	0.117021	-0.33333	0.085106	-0.48	0.064516	0.71875	0.318182	0.153061	-0.32979	-0.06316	-0.85417	-0.83051	-0.53521	-0.87719	-1	-0.92982	-1.03704	0.720306
Draw Con	Trail3HA	15	15	3	18	20	16	12	18	18	15	15	0	20	18	16	20	20	13	
		Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	Jaccard	
	Jaccard	0.578947	0.552632	0.333333	0.560976	0.969697	0.735294	0.416667	0.5	0.658537	0.486486	0.52381	0	0.675	0.65	0.615385	0.714286	0.5	0.414634	0.83136
		Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	
	Cosine	0.737028	0.716928	0.57735	0.719101	0.984732	0.853486	0.615457	0.67082	0.794461	0.668043	0.687836	0	0.80606	0.787879	0.76277	0.836242	0.668994	0.591864	0.831516
		MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	MJaccard	
	MJ	0.666667	0.636364	0.333333	0.69697	0.969697	0.757576	0.454545	0.6	0.818182	0.545455	0.666667	0	0.818182	0.787879	0.727273	0.909091	0.727273	0.515152	0.926417
		Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshte	Levenshtein	
	Lev	0.366667	0.423729	-0.02273	0.390625	0.846154	0.40678	0.196078	0.333333	0.5	0.272727	0.265625	-0.4	0.298507	0.393939	0.365079	0.236111	0.263889	0.258621	0.805464

Figure 38. Correlation analysis of SME assessment versus similarity index-based calculation

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 5 captures the correlation coefficient for both gram series.

Table 5. Similarity index-based correlation analysis

Motif/HOTS	**Similarity Measure Used**	**Unigram**	**Bigram**
Information Gathering	Jaccard	0.868652	0.465106
	Modified Jaccard	0.872283	0.61041
	Cosine	0.892769	0.5811
	Levenshtein	0.75789	0.70652
Critical Thinking	Jaccard	0.800687	0.48334
	Modified Jaccard	0.812488	0.59402
	Cosine	0.80395	0.59893
	Levenshtein	0.720306	0.72426
Decision Making	Jaccard	0.83136	0.37737
	Modified Jaccard	0.926417	0.52581
	Cosine	0.831516	0.46942
	Levenshtein	0.805464	0.73738

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 conclusion, the order of the activities were 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.

6.6 Discussion

6.6.1 Reflection on Findings

In this study, we applied learner versus expert series similarity measure-based HOTS assessment to find the answer to our research questions:

RQ1d: How can small yet meaningful series of process metrics (motifs) be used for HOTS assessment?

RQ2c: How can similarity analysis between student and expert motifs be used to assess HOTS?

RQ2d: Which similarity indices are more effective to assess HOTS?

Study 4 answered our research questions and found that the 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. Our results also showed that defining motifs based on dominant skill in a task is a potentially suitable method for assessment and that applying different similarity methods is potentially more effective and possible when using motifs.

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:

- 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 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.

6.6.2 Limitations and Further Research

In performing Study 4, 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 for Study 4's 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 (Qorbani, 2021). 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 more flexible assessment, but the general approach could stay the same.

An additional major challenge surfaced when the research team realized the assessment approach differed between the SME's assessment and the metric-based assessment. When the first correlation analysis was performed, we found a high correlation on the Section 1 skill, information collection). However, we found a weak correlation for critical thinking and decision-making skills. Discussing with the SME, we realized that the SME viewed and assessed Sections 2 and 3 logically together. As such, in cases where a learner performed the experiment and emergency procedure in Section 3, they also received credits for Section 2. If this approach had been agreed upon in advance, the metric-based assessment could have been adjusted. However, our current method design did not support that, as it was designed to perform assessment per section per HOTS. Therefore, we requested the instructor re-assess each section independently and update the marks accordingly. After that adaptation, we found a high correlation in Sections 2 and 3. The results reported in the study findings section were based on the instructor's second assessment. Investigating other combinations, such as those in the instructor's first assessment, can be the subject of future research.

This study confirmed the difficulty of having a variety of expertise that increases the cost and time involved in running metric-based assessments. To collect data and perform the data analysis as reported in Study 4, we had to hire two external experts: (1) an expert to create the logging reporting system of each activity in the virtual learning environment (chemistry lab), and (2) a programmer to create the gram series and calculate the similarity indexes for each index as per the research team's design requirements.

Our research collaborator Sam Qorbani was the research team's 3D designer, and we are grateful for his work. Typically, for educational researchers, this is another expert they would need to hire.

7 Overall Reflections

7.1 Summary of Findings and Contribution

Within the scope of the reported dissertation, we performed four individual studies and reported the findings. Below is the summary and findings of our studies:

Study 1: Integrating Virtual Spaces with Twenty-first Century Learning:

In this study, we argued that changing the educational need of the new century requires developing HOTS for future citizens. Study 1 investigates 3DVLEs as educational platforms facilitating HOTS development.

Findings of this Study 1 suggest that VLEs can implement educational activities based on learning theories such as experiential and situated learning (research question 1a), strategies often recommended for HOTS development.

Study 1 contributes to the literature by identifying the affordances of 3DVLEs for learning activities such as authentic spaces, immersion, interactivity and highly recommending purposeful space design for HOTS development by employing the platform's affordances. Study 1 also suggests the use of analytics in combination with observational data to clarify the role of space and students' progress.

It should be noted that Study 1 was a foundation and motivator for the rest of the studies reported in this document.

Study 2- Learning Skill Assessments on LMS: Study 2 investigates combined process metrics and insights they might be providing as a HOTS/learning assessment tool (research questions 1b, 1c and 2a).

Study 2 findings suggest that combining platform collected metrics into aggregated dimensions such as Attention and Participation groups can provide additional insight into students' future performance.

Study 2 contributes to the literature by offering a new way of analyzing LMS data for understanding students learning skills and providing instructors the ability to offer help in a timely manner.

Study 3 – Score-based HOTS Assessment in 3DVLE: Study 3 investigates HOTS assessment in 3DVLE by forming motifs of learning tasks and scoring basic metrics for granular assessment over motifs (research questions 1d, 2b)

Study 3 findings show a high correlation between the instructor's manual assessment and motif-based controlled scoring assessment over three of the four HOTS assessments investigated.

Study 3 provides supporting evidence that motif-based granular assessment in identifying students' HOTS might be possible by employing a script-based scoring analysis method. Study 3 also highlights the importance of following students' attention and information collection time (reading and investigating objects), data that are not usually captured by 3DVLE educational platforms.

Study 4- Series-based HOTS Assessment in 3DVLE: Study 4 investigates employing time series and similarity analysis in the form of process metrics for the task of assessing students' HOTS skills (research questions 1d, 2c and 2d).

Study 4 findings shows high correlation over motif-based similarity analysis results in comparison to SMEs assessment. Study 4 findings also suggest that the characteristics of the learning tasks at hand should guide the selection of the gram number of the motif series and the similarity index used for similarity analysis.

Study 4 contributes to the field of series-based similarity assessment research by providing a more granular assessment approach and highlighting the compatibility concern in between learning tasks and gram series/similarity index.

7.2 Overall Discussion and Limitations

Our research hypothesized that metric-based HOTS assessment in VLEs is possible, as supported by evidence from the individual studies reported in this dissertation. We initially thought we could identify one unique way of defining process metrics that could be used for HOTS assessment and a method of analyzing them; however, our findings suggested that there are multiple ways of creating process metrics by applying different filtering algorithms along with a compatible analysis method. Raw metrics collected during the same learning unit of a task can be divided into different HOTS dimensions, as shown in Studies 2 and 3, to gain insight on different HOTS components; Attention and Participation or Information Collection and Drawing Conclusion, respectively. Alternatively, different units of the same learning curriculum might provide evidence of different HOTS and be divided accordingly, as was the strategy used in Study 3.

When it comes to analysis methods, we found that different methods can be applied, such as time series cluster analysis in Study 2, score-based stealth analysis in Study 3, and series-based stealth analysis with similarity index in Study 4. All of the methods provided valuable insight into learners' thinking skills and development.

Overall, we favoured studying methods that were relatively easy to apply, adaptable to curriculum change, and did not require many data points to be collected in advance to be applied in machine model creation to ease of use. As such, we focused on similarity-based procedural assessment in Study 4.

Some of the limitations of our studies that can be directions for future research were:

- Due to COVID-19 restrictions during our research, we could not invite participants to the university laboratory. Recording screens at home while participants using the learning app presented challenges due to the size of the video. As a result, we could not use participants' videos as planned. We recommend future investigation using these videos.
- We could not use data collected from desktop app participants due to the differences in data point creations between DVR and IVR. We recommend future studies to investigate these differences beforehand so as not to lose data.
- Our initial correlation analysis did not show alignment between the SME's assessment and the data assessment. Discussion with the SME showed different strategies in two assessment methods, where SME did not evaluate each section according to its own internal objectives. We recommend more detailed discussion with SMEs to coordinate the grading approach in future research.

Through our literature review, we established three principles to guide the good use of data: (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. Our studies followed these principles as we used no proxy data, used the data only to offer feedback to students, and applied clear and explainable algorithms. Despite these efforts, more is needed to be done to achieve an ethical use of learning data. We did not address the issue of anonymity directly. Future research could investigate methods to collect data without identifying the students. Also, under-represented groups and the

issue of power dynamics were considered beyond the scope of our research but are important for future studies.

All our studies provided insight into the technical challenges of creating a 3DVLE curriculum. The challenge of creating such a curriculum increases if the aim is to connect metric-based assessment with the curriculum. Below is a list of the different experts' disciplines involved in our studies:

- Subject matter experts (SMEs)
- 3D content creation expert
- Metric embedding and collection expert
- Data analyst

Despite this expert help, we had issues collecting data from desktop app versions of 3DVLEs and could not use the data collected from desktop participants due to problems in applying our filtering logic. Yet, we still believe that, as the information is becoming more and more accessible, and with the educational focus moving from the knowledge transfer to skill development, assessment approaches should also shift from output-based to process-based. Evidence collection on the process can be too expensive and not practical if performed by the human observer. The above-mentioned challenges in data preparation and analysis as well as the multiple ways of approaching data collection, filtering, and analysis, highlighted the need for an easy-to-use, comprehensive tool to help scholars and educators with the tasks. In response, we propose the assessment framework detailed in the next section to help educators understand students' development needs and provide feedback.

7.3 Stealth Assessment Framework

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. In this section, we propose a standalone assessment framework that can work with different 3DVLE platforms through using different units that perform functions such as process metric identification, process metric collection, and assessment method application. The proposed framework collects data for assessment without interrupting students' learning process, hence *stealth assessment* (Shute. 2011). Unlike traditional methods, in stealth assessment, learners would not be interrupted while

going through the learning task for the purpose of assessment. This is as long as consent was already given. Assessment data collection would be taking place in the background without interrupting the flow of the learning task. As we stated in our ethical guidance, we believe, no data should be collected without the consent of the creator and collected data should only be used in providing service to the creator. Further research is required to investigate the effect of such data collection on the students, compared to when they are not "observed."

While the present dissertation studies were underway, a similar standalone stealth assessment framework was proposed by Georgiadis et al. (2018). As shown in Figure 41, the authors stated that the common method (referred to as *Original* in Figure 41) is for assessments to be embedded into the virtual environment (referred to as *Serious Game*). They proposed that instead, the assessment should be done as a standalone unit to provide a standalone framework.

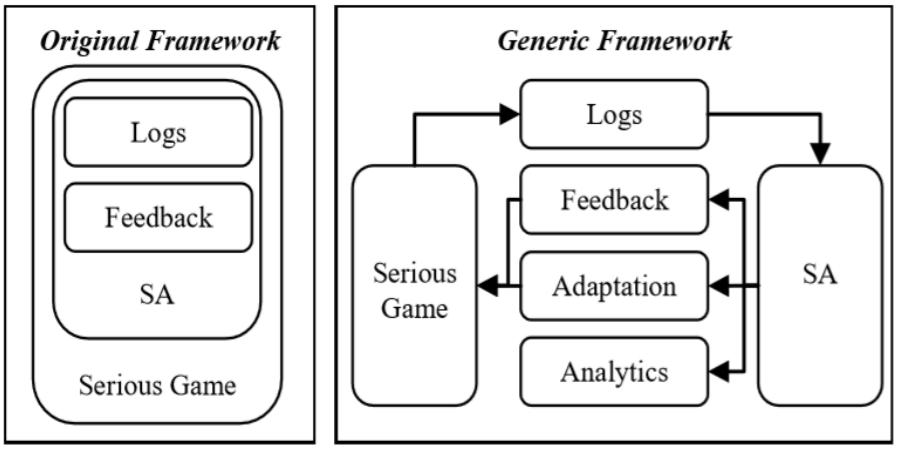


Figure 39. Stealth assessment original and proposed generic framework (Georgiadis et al., 2018)

In recent years, practical difficulties in running a metric-based assessment and the lack of tool supporting researchers and educators have received attention from some researchers. Westera et al. (2020) in the paper called "Artificial Intelligence Moving into Series Gaming", mentioned supporting the approach of Georgiadis et al. (2018) and repeated the need for a stand-alone framework. A study reported by Ahmed et al.(2022) suggested a tool called Chatbot and it's high-level architecture as a way of integrating AI assessment into the academic world. Pointing out the importance of the ethical components being missing in previous proposals, Christopoulos et al.

(2021) proposed ARLEAN: An Augmented Reality Learning Analytics Ethical Framework for augmented reality-based applications having a component checking ethical rules on data.

All these proposals share the common components of i) data collection component; over dedicated network or cloud; ii) a machine learning based assessment component, and iii) display component for the result of the assessment. It is mentioned in previous sections that due to lengthy training data collection phase and the difficulty in changing the course content once the assessment model is trained, we find machine learning-based frame proposals not necessarily a practical solution for classroom teachers. Therefore, in this section, we propose extensions to the existing process metric-based stealth assessment framework proposals to support researchers and classroom teachers. The list below identifies the components of the proposed framework designed to address the needs identified in previous chapters, i.e., an easy-to-use way of defining process metrics, applying assessment methods, and analyzing the results. 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 (Figure 40).

1. **Main control unit:** There should be an interface for the user along with a menu for the framework to access sub-components.
2. **3DVLE and SAF synchronization:** A common understanding and the list of metrics in use should be established for the 3DVLE and SAF to work and for the activity designer to become familiar with basic metric options. It would be ideal for 3DVLEs to keep a master metric file to show all the basic metrics collected for the platform. Once this file has been retrieved by SAF, the user can select the type of basic metric of interest to be collected/reported to make further analysis easier.
3. **Process metrics identification and collection unit:** The purpose of this unit is to help create an activity rubric (such as expert path). Following an activity being created, this component should initially be used to record the teacher's own activity. This recording can be used to define (1) the expert series or (2) the unit of learning tasks and process metrics to be collected. A metric recording function should provide a recording start/stop ability to capture users' activity when needed, including interaction time, username, and specific triggering metrics for filtering capability. The process metric collection unit should only be used with students' consent. An in-built mechanism

should be implemented to acknowledge students in data collection for assessment and feedback purposes.

4. **Assessment method identification and feedback definition:** This component should help educators define the assessment method and customize the selected metrics. With the help of this unit, teachers can define learning task components and process metrics; a meaningful series of activities, identification number of the metric; time; and the task it belongs to. Educators can also define feedback that might be given to learners based on steps missing along the process, such as, "Make sure you get all the information" if any information collection is missing, or "You may like to try boiling it to see the result" if boiling is required within the process. In the future, this unit might be updated to provide real-time content changes and additions based on students' progression in the activity.
5. **Visualization (assessment results display) unit:** This unit provides options to display assessment results and feedback to users. Assessment result visualization was not within the scope of this dissertation.

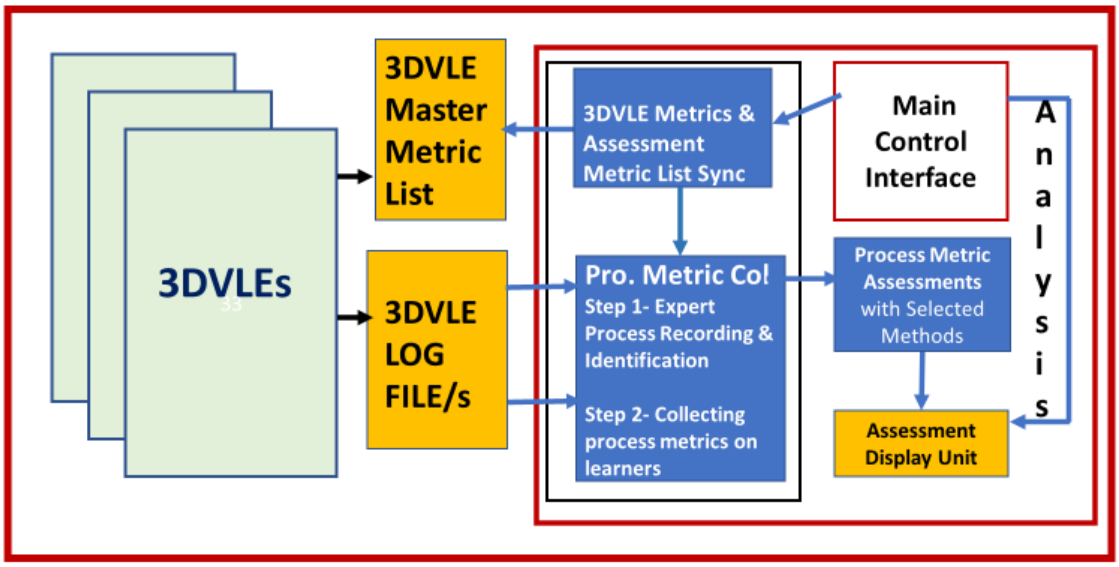


Figure 40. Proposed Stealth Assessment Framework components and interactions

The aim of this framework is to offer a collection of tools that can easily be used by instructors to develop various assessment strategies for different VLEs. We understand that there

will be many modifications and iterations before the proposed SAF is mature. For example, a standard interface for working with various VLEs or the user interface are among the items that we have not discussed at all. Similarly, the AI algorithms for assessment and providing feedback and suggestion are necessary parts of this framework that should be investigated.

We are aware of the concerns that when students know they are "being watched," they may feel additional stress. An opportunity for performing 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.

8 Conclusion

This dissertation investigated HOTS-related metrics and assessment strategies in VLEs. VLEs, with their ability to offer remote virtual classrooms, play an important role when physical classrooms cannot be used. The COVID-19 pandemic forced educational organizations to revisit the importance of remote educational technologies; as such, VLEs are getting increasing attention as a viable space for learning.

We started this research with these two observations:

- HOTS are important to build a healthy society and a strong economy; we need innovative curricula and assessment methods to foster them for our future citizens.
- HOTS assessment requires process-based metrics, not output-based ones.

Knowing that VLEs, especially 3DVLEs, have unique affordances and data collection abilities throughout the process, we hypothesized that they can be used for HOTS assessment. Through four studies, we investigated and demonstrated that various combination of VLE metrics and methods could in fact be used for HOTS assessment.

As such, our biggest contribution is confirming the possibility of automated assessment of HOTS, that traditionally only is done by human observer. In today' education system, assigning a human observer to each student is almost impossible to while they are going through a process-based learning task. Another contribution of this study is suggesting those specific metrics and methods and showing that different combinations and analyses can suit different scenarios. In particular, we demonstrated the use of motifs with their ability to focus on parts of a learning task. Finally, we proposed a framework that is built around the idea of being easy-to-use for instructors.

We envision that in future with the help of proposed assessment tool 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 individualised feedback on the process. We encourage 3DVLE platforms to not only provide detail basic process metrics, but also provide recording and comparison ability. It would be very beneficial for students to reflect on their performance watching themselves going through the process and compare themselves with an expert on the same task in a split screen where deviations can be automatically identified and be pointed out. Classroom teachers, after designing a situated experiential task, can record their process on the

learning task and offer to students for comparison and self reflection with embedded additional text suggestions and feedback.

There are several limitations in our research (listed in Section 7.2) that can be addressed in future research. Our work in this research only investigated individual learner process and assessment. However, future research might investigate team-based learning task and the assessment of each members skills. Our proposed framework is an initial idea that was only partially investigated through the studies reported in this dissertation. There are components of a comprehensive framework that we could not address due to scope, such as:

- Application Programming Interface (API) design for proper interfacing of components and VLE platforms
- Visualization recommendations of the assessment results and visualization component design such as creatively showing the activity patterns and mistakes on them.

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

Achterbosch, L., Pierce, R., & Simmons. G. (2007). Massively multiplayer online role-playing games: The past, present, and future. Computers In Entertainment, 5(4, no. 9).

Aghabozorgi S, Shirkhorshidi AS, Wah TY. (2015). Time-series clustering–A decade review. Information Systems. 1;53:16-38.

Ahadi, A., Lister, R., Haapala, H. and Vihavainen, A., (2015). Exploring machine learning methods to automatically identify students in need of assistance. In Proceedings of the eleventh annual International Conference on International Computing Education Research (pp. 121-130). ACM.

Ahmad, S. F., Alam, M. M., Rahmat, M. K., Mubarik, M. S., & Hyder, S. I. (2022). Academic and Administrative Role of Artificial Intelligence in Education. *Sustainability*, *14*(3), 1101.

Aleksieva-Petrova, A., Chenchev, I. and Petrov, M., (2019). LMS Data Collection, Processing and Compliance with EU GDPR. In EDULEARN19.

Ally, M. (2004). Foundations of educational theory for online learning. Theory and practice of online learning, 2, 15-44.

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).

Alonso-Fernandez, C., Calvo-Morata, A., Freire, M., Martinez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, *141*, 103612.

Alotaibi, F E, AlZhrani, G A, Sabbagh, A J, Azarnoush, H, Winkler-Schwartz, A, Del Maestro, R F (2015). Surgical innovation, 2015-12, Vol.22 (6), p.636-642

Amershi, S., & Conati, C. (2006). Automatic recognition of learner groups in exploratory learning environments. In International Conference on Intelligent Tutoring Systems (pp. 463-472). Springer, Berlin, Heidelberg.

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching and assessing: a revision of Bloom's Taxonomy of educational objectives. New York: Longman.

Annetta, L. A. (2008). Video games in education: Why they should be used and how they are being used. Theory into practice, 47(3), 229-239.

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. (2011). Virtual space as a learning environment: Two case studies. In International Education Technology, Istanbul, Turkey.

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/

Arya, A., Nowlan, N., & Sauriol, N. (2010). Data-driven framework for an online 3D immersive environment for educational applications. In Proceedings of the International Conference on Education and New Learning Technologies (pp. 4726-4736). Barcelona, Spain.

Attaran, M., Stark, J., & Stotler, D. (2018). Opportunities and challenges for big data analytics in US higher education: A conceptual model for implementation. Industry and Higher Education, 32(3), 169-182.

Azarnoush, H., Alzhrani, G., Winkler-Schwartz, A., Alotaibi, F., Gelinas-Phaneuf, N., Pazos, V., Choudhury, N., Fares, J., DiRaddo, R. and 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), pp.603-618.

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. Educational psychologist, 40(4), 199-209.

Baker, R. & Clarke-Midura, J. (2013). Predicting successful inquiry learning in a virtual performance assessment for science." International conference on user modeling, adaptation, and personalization. Berlin, Heidelberg: Springer.

Baker, R. S. J. D. (2010). Data mining for education. International encyclopedia of education, 7(3), 112-118.

Beal, C. R., Qu, L., & Lee, H. (2006). Classifying learner engagement through integration of multiple data sources. In AAAI (pp. 151-156).

Bellanca, J. A., (2010). 21st century skills: Rethinking how students learn. Solution Tree Press.

Belotti, F., Kapralos, B., Lee, K., and Ger, P.M. (2013). Assessment in and of Serious Games, Adv. Human-Comput. Interaction, vol. 2013, Article 1.

Bennett, H. (2003). Successful K-12 technology planning: Ten essential elements." Teacher Librarian, 31(1), 22.

Bienkowski, M. (2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics. US Department of Education. https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Biggs, J. B. (1989). Approaches to the enhancement of tertiary teaching. Higher education research and development, 8(1), 7-25.

Biocca, F. (2014). Connected to my avatar. In International Conference on Social Computing and Social Media (pp. 421-429). Springer, Cham.

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.

Brooks, C. A., Thompson, C., & Teasley, S. D. (2014). Towards A General Method for Building Predictive Models of Learner Success using Educational Time Series Data. In LAK Workshops.

Brown, J. S. (2005). New learning environments for the 21st century. Retrieved from http://www.cais.ca/uploaded/Temp/21st_century_school/newlearning.pdf

Brown, J. S., Collins, A., & Duguid, P. (1989). Commentary: Debating the Situation: A Rejoinder to Palincsar and Wineburg. Educational Researcher, 18(4), 10-62.

Brown, M. (2011). Learning analytics: The coming third wave. Washington, DC: Learning Initiative. http://net. educause.edu/ir/library/pdf/ELIB1101.pdf

Bruner, J. S. (1960). The Process of education. Cambridge, MA: Harvard University Press.

Bruner, J. S. (1961). The act of discovery. Harvard Educational Review, 31, 21-32.

Bruner, J. S. (1966). Toward a theory of instruction. Cambridge, MA: Belkapp Press.

Bryman, A. & Bell, E. (2011) “Business Research Methods” 3rd edition, Oxford University Press

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.

Casey, K. and Azcona, D., (2017). Utilizing student activity patterns to predict performance. International Journal of Educational Technology in Higher Education, 14(1), p.4.

Catherine D'Ignazio, Lauren F. Klein, (2020). Data Feminism

Cerezo, R., Esteban, M., Sánchez-Santillán, M. and Núñez, J.C., (2017). Procrastinating behavior in computer-based learning environments to predict performance: A case study in Moodle. Frontiers in psychology, 8, p.1403.

Cerezo, R., Esteban, M., Sánchez-Santillán, M. and Núñez, J.C.. (2017). Procrastinating behavior in computer-based learning environments to predict performance: A case study in Moodle. Frontiers in psychology, 8, p.1403.

Christopoulos, A., Mystakidis, S., Pellas, N., & Laakso, M. J. (2021). ARLEAN: An Augmented Reality Learning Analytics Ethical Framework. *Computers*, *10*(8), 92.

Chung, H. M., & Gray, P. (1999). Data mining. Journal of management information systems, 16(1), 11-16.

Ciftci, S. (2018). Trends of Serious Game Research from 2007 to 2017: A Bibliometric Analysis, J. Educ. Train. Stud., vol. 6, no. 2, pp. 18–27.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. Journal of Research on Technology in Education, 42(3), 309-328.

Code, J., & Zap, N. (2013). Assessments for Learning, of Learning, and as Learning in 3D Immersive Virtual Environments. EdMedia: World Conference on Educational Media and Technology. Association for the Advancement of Computing in Education (AACE).

Code, J., Clarke-Midura, J., Zap, N., & Dede, C. (2012). Virtual performance assessment in immersive virtual environments. In Interactivity in e-learning: Case studies and frameworks (pp. 230-252). IGI Global.

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.

Cowan, J. (2005) In: Designing assessment to enhance student learning. http://www.heacademy.ac.uk/assets/ps/documents/practice_guides/practice_guides/

ps0069_designing_assessment_to_improve_physical_sciences_learning_march_2009.pdf [7th February 2012].

Csikszentmihalyi, M., Montijo, M. N., & Mouton, A. R. (2018). Flow theory: Optimizing elite performance in the creative realm.

Dalgarno, B., & Lee, M. J. (2010). What are the learning affordances of 3D virtual environments? British Journal of Educational Technology, 41(1), 10–32. doi:10.1111/j.1467-8535.2009.01038.x

Daniel, B. K. (2019). Big Data and data science: A critical review of issues for educational research. British Journal of Educational Technology, 50(1), 101-113.

Dede, C. (2007). Reinventing the role of information and communications technologies in education. 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., Grotzer, T. A., Kamarainen, A., & Metcalf, S. (2017). EcoXPT: Designing for deeper learning through experimentation in an immersive virtual ecosystem. Journal of Educational Technology & Society, 20(4), 166-178.

Driscoll, M. (2000). Psychology of learning for instruction. Needham Heights, MA: Allyn & Bacon.

Duffy, T. M., & Savery, J. R. (1994). Problem-based learning: An instructional model and its constructivist framework. In B. G. Wilson (Ed.), Constructivist learning environments: Case studies in instructional design. Englewood Cliffs, NJ: Educational Technology Publications.

DuFour, R. & DuFour, R. (2010). The role of professional learning communities in advancing 21st-century skills. In J. Bellanca (Ed.), 21st-century skills: Rethinking how students learn (pp. 77-95). Solution Tree Press.

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. and Cavalli-Sforza, L.L., (1965). A method for cluster analysis. Biometrics, pp.362-375.

European Commission Staff. (2012). Assessment of Key Competences in Initial Education and Training: Policy Guidance. Retrieved from http://ec.europa.eu/education/news/rethinking/sw371_en.pdf

Facione, P. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Executive summary: The Delphi report. Millbrae, CA: Academic Press.

Fatahi, S., Shabanali-Fami, F. and Moradi, H., (2017). An empirical study of using sequential behavior pattern mining approach to predict learning styles. Education and Information Technologies, pp.1-19.

Fiok K, Farahani FV, Karwowski W, Ahram T. (2021). Explainable artificial intelligence for education and training. The Journal of Defense Modeling and Simulation. doi:10.1177/15485129211028651

Floryan, M., Dragon, T., Basit, N., Dragon, S., & Woolf, B. (2015). Who Needs Help? Automating Student Assessment Within Exploratory Learning Environments. In proceedings from International Conference on Artificial Intelligence in Education (pp. 125-134). Springer.

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). Springer, Cham.

Foreman, J. (2014). Data Smart, Wiley

Fosnot, C. T. (2013). Constructivism: Theory, perspectives, and practice. Teachers College Press.

Freedman, R., Ali, S. S., & McRoy, S. (2000). Links: what is an intelligent tutoring system?. intelligence, 11(3), 15-16.

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

Furnham, A., (1996). The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. Personality and Individual Differences, 21(2), pp.303-307.

Galarneau, L.L. (2005), Spontaneous Communities of Learning: Learning Ecosystems in Massively Multiplayer Online Gaming Environments, SSRN, June.

Galindo, I. (2014). What is constructivism? Retrieved from https://www.ctsnet.edu/what-is-constructivism/

Georgiadis, K., van Lankveld, G., Bahreini, K., & Westera, W. (2018). Accommodating Stealth Assessment in Serious Games: Towards Developing a Generic Tool. In 2018 10th

International Conference on Virtual Worlds and Games for Serious Applications (VS-Games) (pp. 1-4). IEEE.

Gibson, D., & de Freitas, S. (2016). Exploratory analysis in learning analytics. Technology, Knowledge and Learning, 21(1), 5-19.

Gibson, J.J., (1979). The Ecological Approach to Visual Perception. Houghton Mifflin Harcourt (HMH), Boston (especially Ch. 8, pp. 127-137: 'The Theory of Affordances').

Gibson, J.J., (1982). Reasons for Realism: Selected Essays of James J. Gibson. Resources for Ecological Psychology. L. Erlbaum, New Jersey. p.411

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Gratch, J. & Marsella, S. (2005). Some Lessons for emotion psychology for the design of lifelike characters. Journal of Applied Artificial Intelligence (Special issue on Educational Agents—Beyond Virtual Tutors), 19(3–4), 215–33

Green, C., Pouget, A., and Bavelier, D., (2010). Improved Probabilistic Inference as a General Learning Mechanism with Action Video Games, Current Biol., vol. 20, no. 17, pp. 1573–1579.

Greene, D. and Cunningham, P., (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In Proceedings of the 23rd international conference on Machine learning (pp. 377-384).

Greeno, J. G. (1994). Gibson's affordances. Psychological Review, 101(2), 336–342. doi:10.1037/0033- 295X.101.2.336 PMID:8022965

Greiff, S. and Funke, J., (2009). Measuring complex problem solving: The MicroDYN approach.

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.

Guardiola, E. & Natkin, S. (2015). A Game Design Methodology for Generating A Psychological Profile Of Players. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), Serious games Analytics (pp. 363-380). Springer International Publishing. .

Gudivada, V.N., Rao, D.L. and Ding, J., (2018). Evolution and Facets of Data Analytics for Educational Data Mining and Learning Analytics. Responsible Analytics and Data Mining in Education: Global Perspectives on Quality, Support, and Decision Making.

Hanushek, E. A. (2003). The failure of input-based schooling policies. The Economic Journal, 113, 64-98.

Harasim, L. (2017). Learning theory and online technologies. Routledge.

Harford, T. (2014). Big data: A big mistake? Significance, 11(5), 14-19.

Hartigan, J.A. and Wong, M.A., (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), pp.100-108.

Hartigan, J.A. and Wong, M.A., (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), pp.100-108.

Hartwick, P. & Nowlan, N. S. (2018). Integrating Virtual Spaces: Connecting Affordances of 3D Virtual Learning Environments to Design for Twenty-First Century Learning. In Q. Yufeng (Ed.), Integrating Multi-User Virtual Environments in Modern Classrooms (pp. 111-136). IGI Global. Web. 1 Jan. 2018. doi:10.4018/978-1-5225-3719-9

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

Holland, J. L. (1966). The psychology of vocational choice. A theory of personality types and model environments. Waltham, MA: Blaisdel.

Holt, D. G. & Willard-Holt, C. (2000). Let's get real – students solving authentic corporate problems. Phi Delta Kappan. 82(3).

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. and 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.

Ibáñez, M. B., Garcia, J. J., Galan, S., Maroto, D., Morillo, D., & Kloos, C. D. (2011). Design and implementation of a 3D multi-user virtual world for language learning. Journal of Educational Technology & Society, 14(4), 2–10.

Ice, P., Diaz, S., Swan, K., Burgess, M., Sharkey, M., Sherrill, J., Huston, D. R., & Okimoto, H. (2012). The PAR Framework Proof of Concept: Initial Findings from a Multi-Institutional Analysis of Federated Postsecondary Data. Journal of Asynchronous Learning Networks, 16(3), 63-86.

Ifenthaler, D., and C. Widanapathirana.(2014). Development and validation of a learning analytics framework: Two case studies using support vector machines. Technology, Knowledge and Learning, vol. 19, no. 1–2, pp. 221–240, 2014, https://doi.org/10.1007/s10758-014-92 26-4.

Johnson, L. F. & Levine, A. H. (2008). Virtual Worlds: Inherently Immersive, Highly Social Learning Spaces. Theory Into Practice, 47(2), 161–170.

Johnson, L., Levine, A., Smith, R., & Stone, S. (2010). The 2010 Horizon report. Austin, TX: New Media Consortium.

Jonassen, D. H. (1999). Designing constructivist learning environments. Instructional design theories and models: A new paradigm of instructional theory, 2, 215-239.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kaptelinin, V., & Nardi, B. A. (2006). Acting with technology: Activity theory and interaction design. Cambridge, MA: MIT Press.

Kelman, P. (1989). Alternatives to integrated instructional systems. Paper presented at the National Educational Computing Conference, Nashville, TN.

Kevin, M. (2017). The ethics of surveillance: An introduction. Routledge.

Kiili, K. (2004). Learning with technology: cognitive tools in multimedia learning materials. EdMedia: World Conference on Educational Media and Technology. Association for the Advancement of Computing in Education (AACE).

Kilpatrick, J., Swafford, J., & Findell, B. (2001). Adding It Up: Helping Children Learn Mathematics. Washington, DC: National Academies Press.

Kim, S. Y. S., Prestopnik, N., & Biocca, F. A. (2014). Body in the interactive game: How interface embodiment affects physical activity and health behavior change. Computers in Human Behavior, 36, 376-384.

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). Retrieved from http://www.cala.fsu.edu/files/higher_order_thinking_skills.pdf

Kolb, D. A. (2014). Experiential learning: Experience as the source of learning and development. FT press.

Kozlova I. (2019). Factors affecting learner collaboration in 3D virtual worlds. in Assessing the effectiveness of virtual technologies in foreign and second language instruction (pp. 26-60). IGI Global.

Krasnow, K. W. & Bruening, P. J. (2014). Big Data analytics: risks and responsibilities. International Data Privacy Law, 4(2), 89-95.

Kristine L. Nowak, Frank Biocca;(2003). The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. Presence: Teleoperators and Virtual Environments; 12 (5): 481–494. doi: https://doi.org/10.1162/105474603322761289

Kuhn, D. (2001). How do people know? Psychological Science, 12, 1–8.

Kuhn, D. (2005). Education for thinking. Cambridge, MA: Harvard University Press. Kuhn, D. & Udell, W. (2001). The path to wisdom. Educational Psychologist, 36(4), 261-264.

Langer, J. A. (1986). A sociocognitive perspective on literacy. Viewpoints, 120, 1-38.

Lefor, A. K., Harada, K., Kawahira, H., & Mitsuishi, M. (2020). The effect of simulator fidelity on procedure skill training: a literature review. International journal of medical education, 11, 97. Aldrich, C. (2009). The complete guide to simulations and serious games: How the most valuable content will be created in the age beyond Gutenberg to Google. John Wiley & Sons

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. (2013). Performance metrics for serious games: Will the (real) expert please step forward? Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games (CGAMES) 18th International Conference on IEEE.

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. (2015). Measuring the (dis-) similarity between expert and novice behaviors as serious games analytics. Education and Information Technologies, 20(1), 5-19.

Loh, C. S. (2012). Information trails: In-process assessment of game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), Assessment in game-based learning (pp. 123-144). New York: Springer.

Loh, Christian Sebastian, and Yanyan Sheng. (2015). Measuring expert performance for serious games analytics: From data to insights. In Serious Games Analytics, pp. 101-134. Springer, Cham.

Lokse, M., Låg, T., Solberg, M., Andreassen, H. N., & Stenersen, M. (2017). Teaching information literacy in higher education: Effective teaching and active learning. Chandos Publishing.

M. Halkidi, Y. Batistakis, M. Vazirgiannis,(2001). On clustering validation techniques, Journal of Intelligent Information Systems 17 (2001) 107–145.

Martín-Gutiérrez, J., Mora, C. E., Añorbe-Díaz, B., & González-Marrero, A. (2017). Virtual technologies trends in education. EURASIA Journal of Mathematics Science and Technology Education, 13(2), 469-486.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? American psychologist, 59(1), 14.

McCall, M. & Clarke-Midura, J. (2013). Analysis of gaming for assessment.

McEachen, J., Fullan, M., & Quinn, J. (2018). 2018 NPDL Global Report. New Pedagogies for Deep Learning: A global partnership. Deep learning series, 5.

McLeod, S. A. (2008). Bruner. Retrieved from www.simplypsychology.org/bruner.html

Mehdi, T., Bashardoost, N. and Ahmadi, M., (2011). Kernel smoothing for ROC curve and estimation for thyroid stimulating hormone. Int J Public Health Res, 1, pp.239-242.

Mendes, R., & Vilela, J. P. (2017). Privacy-preserving data mining: methods, metrics, and applications. IEEE Access, 5, 10562-10582.

Młynarska, E., Greene, D. and Cunningham, P., (2016). Indicators of good student performance in moodle activity data. arXiv preprint arXiv:1601.02975.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to time series analysis and forecasting. John Wiley & Sons.

Moon, Sumyung, Joel R. Reidenberg, and N. Cameron Russell. (2017). Privacy in Gaming and Virtual Reality Technologies: Review of Academic Literature.

Moon, T.K., (1996). The expectation-maximization algorithm. IEEE Signal processing magazine, 13(6), pp.47-60.

Mooney, C.H. and Roddick, J.F., (2013). Sequential pattern mining--approaches and algorithms. ACM Computing Surveys (CSUR), 45(2), pp.1-39.

Moos, D. C. & Azevedo, R. (2009). Learning with computer-based learning environments: A literature review of computer self-efficacy. Review of Educational Research, 79(2), 576-600.

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). Distance measures for time series in R: The TSdist package. R journal, 8(2), p. 451-459.

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. International journal of modern education & computer science, 8(11).

Munoz, A. (2014). Machine Learning and Optimization. Retrieved from https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf

Mwalumbwe, I. and Mtebe, J.S., (2017). Using learning analytics to predict students' performance in moodle learning management system: a case of mbeya university of science and technology. The Electronic Journal of Information Systems in Developing Countries, 79(1), pp.1-13.

N.R. Pal, J. Biswas, (1917). Cluster validation using graph theoretic concepts, Pattern Recognition 30 , 847–857.

Narayanan, S. A., Prasanth, M., Mohan, P., Kaimal, M. R., & Bijlani, K. (2012). Attention analysis in e-learning environment using a simple web camera. In 2012 IEEE International Conference on Technology Enhanced Education (ICTEE) (pp. 1-4). IEEE.

Nowlan, N. S., Hartwick, P., & Arya, A. (2018). Skill assessment in a virtual environment. Paper presented at the IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Ottawa, Canada.

Nowlan, NS, Riesen, E, Morley, M, Arya, A, Sauriol, N. (2011). A framework for an immersive learning environment with telemetries and simulation, Ubiquitous Learning: An International Journal, 2011

O'Neil, C. (2017) *Weapons of Math Destruction*, Crown

Pan, Z., Cheok, A. D., Yang, H., Zhu, J., & Shi, J. (2006). Virtual reality and mixed reality for virtual learning environments. Computers & Graphics, 30(1), 20-28.

Parry, M. (2011). Colleges mine data to tailor students' experience. Chronicle of Higher Education. Retrieved from https://chronicle.com/ article/A-Moneyball-Approach-to/130062/

Parry, M. (2012). Big data on campus. New York Times. Retrieved from http://www.nytimes.com/2012/07/22/education/edlife/col leges-awakening-to-the-opportunities-of-data-mining.html

Patel, K. M. A., and Thakral, P. (2016). "The best clustering algorithms in data mining," 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 2042-2046, doi: 10.1109/ICCSP.2016.7754534.

Pearlman, B. (2010). Designing new learning environments to support 21st-century skills. In J. Bellanca & R. S. Brandt (Eds.), 21st-century skills: Rethinking how students learn (pp. 116-147). Solution Tree Press.

Peffer, M., Quigley, D. and 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).

Peterson, M. (2006). Learner interaction management in an avatar and chat-based virtual world. Computer Assisted Language Learning, 19(1), 79–103. doi:10.1080/09588220600804087

Philips, R., McNaught, C., & Kennedy, G. (2010). Towards a generalized conceptual framework for learning: the learning environment, learning processes and learning outcomes (LEPO) framework. Proceedings of Edmedia: World conference on educational Media and Technology, Association of Advancement of Computing in Education.

Pickup M. (2014). Introduction to time series analysis. Sage Publications

Pirnay-Dummer, P., Ifenthaler, D., & Seel, N. M. (2012). Designing Model-Based Learning Environments to Support Mental Models for Learning. In Jonassen, D. H. and Land, S.

(Eds.), Theoretical Foundations of Learning Environments (2nd ed.) (pp. 66 – 94), New York, NY: Routledge.

Prensky, M. (2003). Digital game-based learning. Computers in Entertainment (CIE), 1(1), 21-21.

Qian, M. & Clark, K. R. (2016). Game-based Learning and 21st century skills: A review of recent research. Computers in Human Behavior, 63, 50-58.

Qian, Meihua, and Karen R. Clark, (2016). "Game-based Learning and 21st century skills: A review of recent research." Computers in human behavior 63 50-58.

Qorbani, H. S., Arya, A., Nowlan, N., & Abdinejad, M. (2021). Simulation and Assessment of Safety Procedure in an Immersive Virtual Reality (IVR) Laboratory. In 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) (pp. 589-590). IEEE.

Quesnel, D., & Riecke, B. E. (2018). Are you awed yet? How virtual reality gives us awe and goose bumps. Frontiers in psychology, 9, 2158.

Reilly JM, 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).

Reisoğlu, I., Topu, B., Yılmaz, R., Yılmaz, T. K., & Göktaş, Y. (2017). 3D virtual learning environments in education: A meta-review. Asia Pacific Education Review, 18(1), 81-100.

Resnick, L. B. (1987). Education and learning to think. National Academies.

Robinson, D. H. & Schraw, G. J. (2011). Assessment of Higher Order Thinking Skills. Charlotte, N.C.: Information Age Publishing.

Romero, C., & Ventura, S. (2013). *Data mining in education*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

Romero, C., & Ventura, S. (2017). *Educational data science in massive open online courses*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(1), e1187.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. Computers & Education, 68, 458-472.

Roschelle, J. M. (2000). Changing how and what children learn in school with computer-based technologies. Future Child, 10(2), 76-101.

Rowe, J., Mott, B. W., McQuiggan, S. W., Robison, J. L., Lee, S., & Lester, J., C. (2009). Crystal island: A narrative-centered learning environment for eighth grade microbiology.

Proceedings from Workshop on intelligent educational games at the 14th international conference on artificial intelligence in education, Brighton, UK.

Rubel, A. & Jones, K. M. (2016). Student privacy in learning analytics: An information ethics perspective. The Information Society, 32(2), 143-159.

Sabourin, J. L. (2013). Stealth assessment of self-regulated learning in game-based learning environments. North Carolina State University.

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.

Sabourin, J. L., Shores, L. R., Mott, B. W., & Lester, J. C. (2012). Predicting student self-regulation strategies in game-based learning environments. International Conference on Intelligent Tutoring Systems. Berlin, Heidelberg: Springer.

Savin-Baden, M. (2008). Learning spaces: Creating opportunities for knowledge creation in academic life. New York, NY: McGraw Hill.

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.

Scheffer, T. (2001). "Finding association rules that trade support optimally against confidence," in Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery (Berlin: Springer), 424–435. doi: 10.1007/3-540-44794-6_35

Schmidt, B, Stewart S. (2009). Implementing the virtual reality learning environment: Second Life. Nurse Educ. 2009 Jul-Aug;34(4):152-5. doi: 10.1097/NNE.0b013e3181aabbe8. PMID: 19574850.

Shaffer, D. W. (2005). Video games and the future of learning. Phi delta kappan, 87(2), 105-111.

Shewaga, R., Uribe-Quevedo, A., Kapralos, B., Lee, K. & Alam, F. (2018). A Serious Game for Anesthesia-Based Crisis Resource Management Training. Computers in Entertainment (CIE), 16(2), p. 6.

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

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.

Siemens, G. & Long, P. (2011, February–March). Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE (2011). Review from the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, 46(5, September/October 2011).

Sisovic, S., Matetic, M. and Bakaric, M.B., (2016). January. Clustering of imbalanced moodle data for early alert of student failure. In 2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI) (pp. 165-170). IEEE.

Smith, L. (1985). Making Educational Sense of Piaget's Psychology. Oxford Review of Education, 11(2), 181–191.

Smith, S. P., Blackmore, K., & Nesbitt, K. (2015). A meta-analysis of data collection in serious games research. Serious games analytics (pp. 31–55). . https://doi.org/10. 1007/978-3-319-05834-4_

Snow, E. L. (2015). Promoting Self-Regulation and Metacognition through the Use of Online Trace Data within a Game-Based Environment (Doctoral Dissertation). Arizona State University.

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.

Sousa, C. (2014). 2014 Ontario budget: Building opportunity, securing our future. Retrieved from: www.fin.gov.on.ca/en/budget/ontariobudgets/2014/papers_all.pdf

Spires, H A. (2008)."21st century skills and serious games: Preparing the N generation." Serious educational games: 13-23.

Squire, K., Giovanetto, L., Devane, B., & Durga, S. (2005). From users to designers: Building a self-organizing game-based learning environment. TechTrends, 49(5), 34-42.

Steffen, J. H., Gaskin, J. E., Meservy, T. O., Jenkins, J. L., & Wolman, I. (2019). Framework of affordances for virtual reality and augmented reality. Journal of Management Information Systems, 36(3), 683-729.

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. Journal of Big data, 2(1), 1-32.

Uğurlu, Y. (2014). User attention analysis for e-learning systems using gaze and speech information. In 2014 International Conference on Information Science, Electronics and Electrical Engineering (Vol. 1, pp. 1-5). IEEE.

Van Laar, E., Van Deursen, A. J., Van Dijk, J. A., & De Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. Computers in human behavior, 72, 577-588.

van Limpt-Broers, H., Louwerse, M. M., & Postma, M. (2020). Awe yields learning: A virtual reality study. In CogSci.

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

Velev, D. & Zlateva, P. (2017). Virtual reality challenges in education and training. International Journal of Learning and Teaching, 3(1), 33-37.

Vince, R. (1998). Behind and beyond Kolb's learning cycle. Journal of Management Education, 22(3), 304-319.

Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. Journal of curriculum studies, 44(3), 299-321.

Voogt, J., Erstad, O., Dede, C., & Mishra, P. (2013). Challenges to learning and schooling in the digital networked world of the 21st century. Journal of computer assisted learning, 29(5), 403-413.

Vygotsky, L. S., (1967). Play and its role in the mental development of the child. Soviet psychology, 5(3), 6-18.

Wang, D., Liu, H. and Hau, K.T., (2021). Automated and interactive game-based assessment of critical thinking. Education and Information Technologies, pp.1-23.

Wang, S. H. (2012). Applying a 3D situational virtual learning environment to the real world business—an extended research in marketing. British Journal of Educational Technology, 43(3), 411-427.

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.

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. doi:10.1111/j.1467-8535.2009.00952.x

Weisz, G., Smilowitz, N. R., Parise, H., Devaud, J., Moussa, I., Ramee, S., ... & Gray, W. A. (2013). Objective simulator-based evaluation of carotid artery stenting proficiency (from Assessment of Operator Performance by the Carotid Stenting Simulator Study [ASSESS]). The American journal of cardiology, 112(2), 299-306.

Westera, W., Prada, R., Mascarenhas, S., Santos, P. A., Dias, J., Guimarães, M., ... & Ruseti, S. (2020). Artificial intelligence moving serious gaming: Presenting reusable game AI components. *Education and Information Technologies*, *25*(1), 351-380.

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

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognising and responding to student affect. International Journal of Learning Technology, 4(3-4), 129-164

Zuiker, S. J., (2012). Educational virtual environments as a lens for understanding both precise repeatability and specific variation in learning ecologies. British Journal of Educational Technology, 43(6), 981-992.

Zurkowski, Paul G. (1974). The Information Service Environment Relationships and Priorities. Related Paper No. 5.

Appendix A

Towards Defining 21st Century Competencies for Ontario

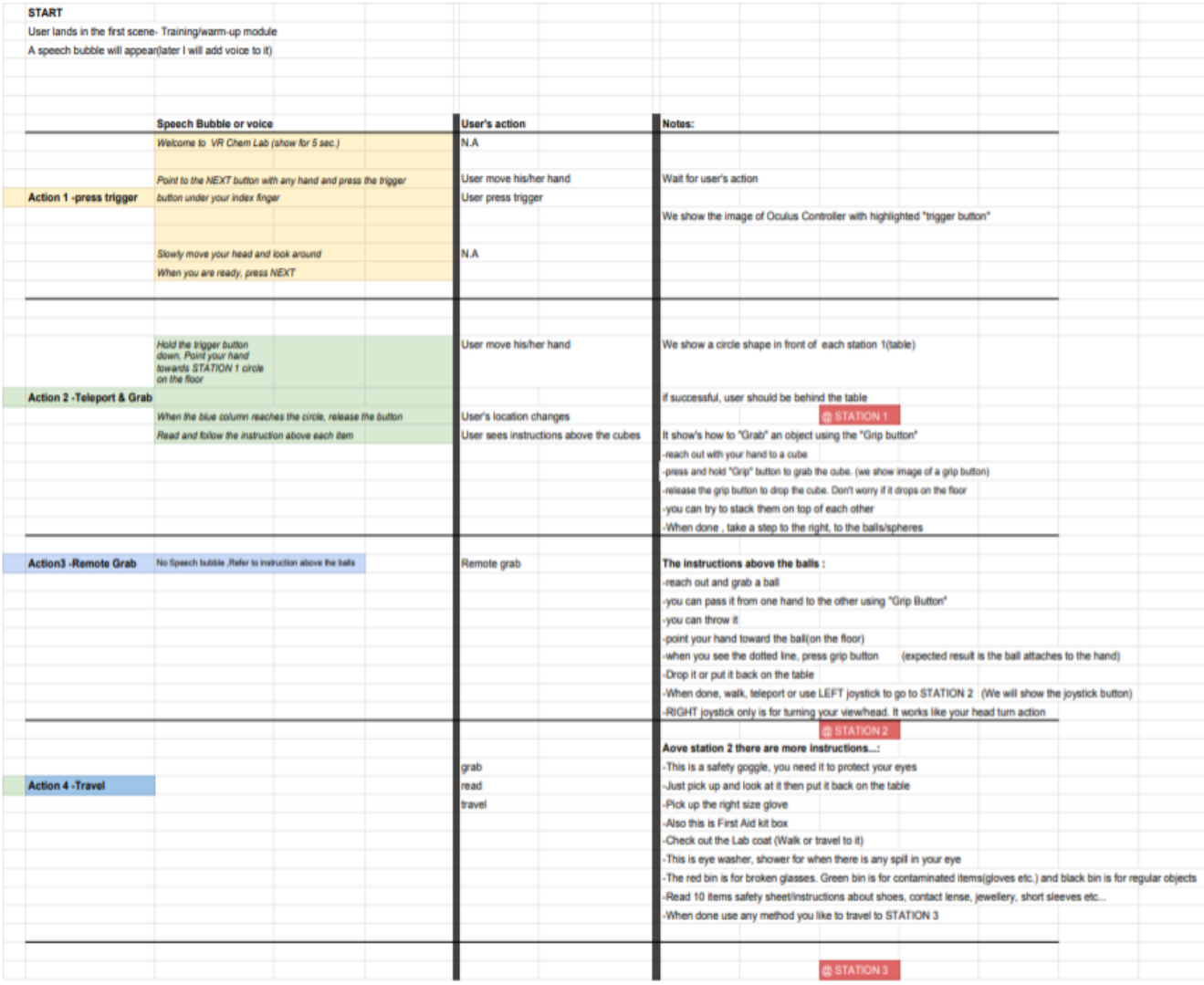
"This document will provide a focus for discussions among ministry and external education, policy, and research experts about how best to shape provincial policy to help students develop the 21st century competencies they need to succeed. These discussions will build on the consultations to renew Ontario's vision for education that took place in the autumn of 2013"[14]

Ontario Vision/Budget Based on Results of Public Consultations (2014)	ATC21S (2012) (Summary of International Frameworks)	Fullan and Scott (2014) The Six Cs
"Achievement also means raising expectations for valuable, higher-order skills like **critical thinking, communication, innovation, creativity, collaboration,** and **entrepreneurship**." (*Achieving Excellence*, p. 3) "[O]ur learners will also need to develop characteristics such as **perseverance, resilience**, and **imaginative thinking** to overcome challenges. Combined with a deep sense of **compassion** and **empathy** for others, our learners will develop the skills and knowledge they need to become actively **engaged citizens**." (*Achieving Excellence*, p. 5) "To achieve success, Ontario will: . . . Foster more young entrepreneurs in Ontario schools by increasing training in **innovation, creativity,** and **entrepreneurship**. . . ." (*Achieving Excellence*, p. 6) "By 2025 . . . Ontario will be a world leader in higher-order skills, such as **critical thinking** and **problem solving**, which will allow Ontario to thrive in the increasingly competitive global marketplace." (*2014 Ontario Budget* [Sousa, 2014], p. 9)	***Ways of Thinking*** 1. Creativity and innovation 2. Critical thinking, problem solving, decision making 3. Learning to learn, metacognition ***Ways of Working*** 4. Communication 5. Collaboration (teamwork) ***Tools for Working*** 6. Information literacy 7. Information and communication technology literacy ***Living in the World*** 8. Citizenship – local and global 9. Life and career (including adapting to change; managing goals and time; being a self-directed learner; managing projects; working effectively in diverse teams; being flexible; producing results; guiding and leading others) 10. Personal and social responsibility (including cultural awareness and competence)	1. **Character** – "qualities of the individual essential for being personally effective in a complex world including: grit, tenacity, perseverance, resilience, reliability, and honesty." (Fullan & Scott, 2014, p. 6) 2. **Citizenship** – "thinking like global citizens, considering global issues based on a deep understanding of diverse values with genuine interest in engaging with others to solve complex problems that impact human and environmental sustainability." (Fullan & Scott, 2014, p. 6) 3. **Communication** – the "mastery of three fluencies: digital, writing, and speaking tailored for a range of audiences." (Fullan & Scott, 2014, p. 6) 4. **Critical Thinking** – "critically evaluating information and arguments, seeing patterns and connections, constructing meaningful knowledge and applying it in the real world." (Fullan & Scott, 2014, p. 7) 5. **Collaboration** – "the capacity to work interdependently and synergistically in teams with strong interpersonal and team-related skills including effective management of team dynamics, making substantive decisions together, and learning from and contributing to the learning of others." (Fullan & Scott, 2014, p. 6) 6. **Creativity** – *"having an 'entrepreneurial eye' for economic and social opportunities, asking the right questions to generate novel ideas, and demonstrating leadership to pursue those ideas into practice." (Fullan & Scott, 2014, p. 7)*

[14] http://www.edugains.ca/resources21CL/About21stCentury/21CL_21stCenturyCompetencies.pdf

Appendix B

A story board created by Study 4 application designer and SME.



START			
User lands in the first scene- Training/warm-up module			
A speech bubble will appear(later I will add voice to it)			
	Speech Bubble or voice	**User's action**	**Notes:**
	Welcome to VR Chem Lab (show for 5 sec.)	N.A	
	Point to the NEXT button with any hand and press the trigger	User move his/her hand	Wait for user's action
Action 1 -press trigger	*button under your index finger*	User press trigger	
			We show the image of Oculus Controller with highlighted "trigger button"
	Slowly move your head and look around	N.A	
	When you are ready, press NEXT		
	Hold the trigger button down, Point your hand towards STATION 1 circle on the floor	User move his/her hand	We show a circle shape in front of each station 1(table)
Action 2 -Teleport & Grab			if successful, user should be behind the table
	When the blue column reaches the circle, release the button	User's location changes	@ STATION 1
	Read and follow the instruction above each item	User sees instructions above the cubes	It show's how to "Grab" an object using the "Grip button"
			-reach out with your hand to a cube
			-press and hold "Grip" button to grab the cube. (we show image of a grip button)
			-release the grip button to drop the cube. Don't worry if it drops on the floor
			-you can try to stack them on top of each other
			-When done , take a step to the right, to the balls/spheres
Action3 -Remote Grab	No Speech bubble ,Refer to instruction above the balls	Remote grab	**The instructions above the balls :**
			-reach out and grab a ball
			-you can pass it from one hand to the other using "Grip Button"
			-you can throw it
			-point your hand toward the ball(on the floor)
			-when you see the dotted line, press grip button (expected result is the ball attaches to the hand)
			-Drop it or put it back on the table
			-When done, walk, teleport or use LEFT joystick to go to STATION 2 (We will show the joystick button)
			-RIGHT joystick only is for turning your view/head. It works like your head turn action
			@ STATION 2
			Aove station 2 there are more instructions...:
Action 4 -Travel		grab	-This is a safety goggle, you need it to protect your eyes
		read	-Just pick up and look at it then put it back on the table
		travel	-Pick up the right size glove
			-Also this is First Aid kit box
			-Check out the Lab coat (Walk or travel to it)
			-This is eye washer, shower for when there is any spill in your eye
			-The red bin is for broken glasses. Green bin is for contaminated items(gloves etc.) and black bin is for regular objects
			-Read 10 items safety sheet/instructions about shoes, contact lense, jewellery, short sleeves etc...
			-When done use any method you like to travel to STATION 3
			@ STATION 3

User gets to this scene after completing the training in scene 1

User actions/Tasks		Required programming
	Similar to station 2 actions	After each item is touched/ picked, we need to show an inventory OR
Travel	1-Put on a lab coat-	an image /avatar on the wall showing those items are used/applied
Grab/touch	2- Button up the lab coat completely.	
	3-Put on a safety goggle.	Users may also answer a few questions about whether they have a contact lense or jewellery on
	4-Put on the right size glove.	
	5-Check shoes (No sandals, front open shoes	-Once all conditions are met, they are allowed to open the door, walk into the lab
	allowed).	
	6-Check pants (Should be long, no skin around	When Done,
	the ankle should be exposed.	**Go to next scene**
	7- If long hair, it should be tied.	
	8-Check Contact lense. No contact lens allowed	
	9-	
	10-	

User gets to this scene after doing the locker room

TASKS

on the entrance user are asked to : - 1- find and go to the location of the eye washer, shower, bins and collect points/coins.(or something).

2- Go to the fume hood and conduct tasks as explaned below

User action/Tasks		Required programming	
	-User should know the location of eye washer, shower,	A system of collecting /recording where the player goes to collect the objects(coins)	
	bins, first aid kit and go to each of those place and collect a coin or	-record the timing	
Travel	symbol object to prove they have been tere...		
Grab/touch			
Manipulate	User should go to the fume hood and conduct the following		
	FUME HOOD EXPERIMENT		
	1- Pull up the hood door/glass up (1/3rd) 2- Grab the stand (from the bench), put it inside the fume hood. 3- Grab/put the hot plate beside the stand 4- Plugin the hotplate 5- Grab/put the oil bath on the hotplate 6- Turn on the hotplate and increase the temperature (140 C) 7- Put a round bottom flask inside the oil bath WITHOUT clamping it 8- The flask falls inside the oil bath and its solution will be poured into the bath and causes fire	-Adding animation, interactions to the fume hood door, -adding temperature increasing animation to the hot plate -adding breaking glass animation, - adding fire	
	The expected behaviour from the student		
	9- The student should unplug the hotplate first 10- Pull down the hood door/glass and let the oil-bath completely cool down 11- Pull up the hood door again 12- Remove the hotplate and broken glasses form fume hood 13- Place the hotplate on the bench 14- Dump the broken glasses into the red bin 11- Clean up the fume hood using napkins 12- Next, acetone should be used to clean up the oily fume hood.	while the fire is on?	timer animation?
			End

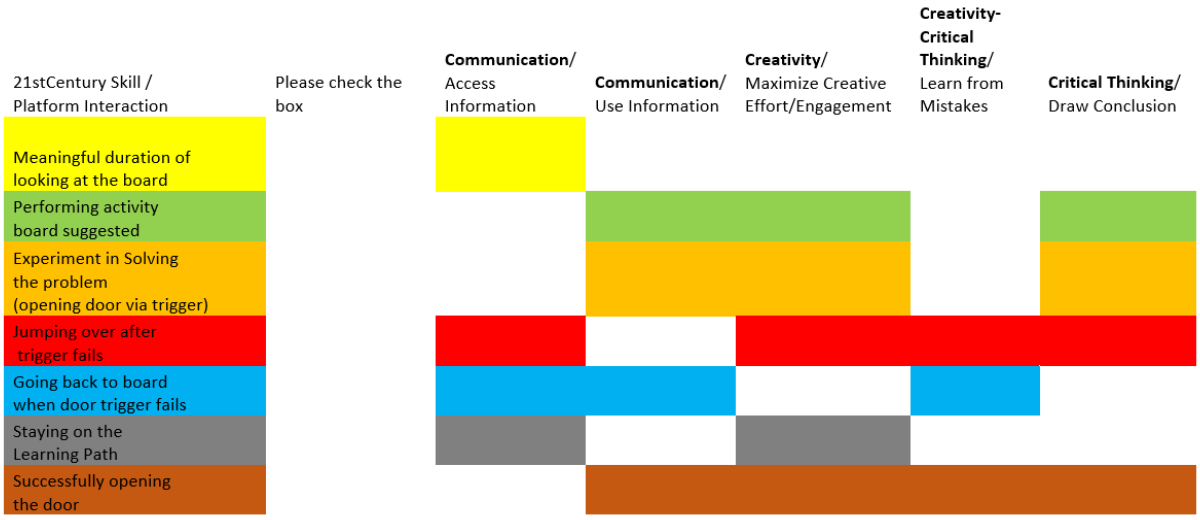
Appendix C

Sample of rubrics used in Study 3:

Study #3 Interaction based rubric:

3dInteraction Based Skill Rubric 1.1 **Participant #_____**

1. **Welcome, walking, jumping**



Appendix D

Study #3 Holistic Rubric Example:

Holistic Assessment Rubric for Activity 4.1 **Participant #__110_____**

1. **Welcome, walking, jumping**

Skill	Sub component of skill	Descriptor of skill	Observation	Comments
Critical thinking	Draw conclusion	Successfully completes task	1 (unobservable) ------2 (low) ------3--------4 (high)	
Critical thinking and creativity	Learn from mistake	Go back and read	1 (unobservable) ------2 (low) ------3--------4 (high)	
Creativity	Engagement	Staying on task while experimenting in the environment	1 (unobservable) ------2 (low) ------3--------4 (high)	
Communication	Information Collection	Reads instructions on board	1 (unobservable) ------2 (low) ------3--------4 (high)	