A CAUSAL-COMPARATIVE STUDY OF NOVICES AND EXPERTS TESTING NIELSEN'S FIVE USER ASSUMPTION WITHIN A LEARNING TECHNOLOGY

Jessica Danielle Gatewood
A CAUSAL-COMPARATIVE STUDY OF NOVICES AND EXPERTS TESTING
NIELSEN’S FIVE USER ASSUMPTION WITHIN A LEARNING TECHNOLOGY

by

Jessica Gatewood

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Education

Major: Instruction and Curriculum Leadership

The University of Memphis

May 2024
Abstract

This non-experimental causal-comparative study aims to explore the possible effect of expertise on learning experience design (LXD) deviation identification and the classification of these deviations in alignment with provided learning experience design constructs within a learning technology. Additionally, this study challenges Nielsen’s (1993) Five User Assumption regarding how many novices or experts are needed to identify 80% of LXD deviations within the learning technology. According to Nielsen’s (1993) Five User Assumption, only five participants are required to identify 80% percent of usability problems; however, this assumption has yet to be tested within a learning technology (Nielsen, 1993). A convenience sample of 10 participants (five novices and five experts) were recruited from a business corporation in the Mid-South region of the United States. Participants were presented with a Gooru module and asked to identify LXD deviations present within the module and rate their severity. Before this, two outside LXD experts evaluated the learning technology and comprised a list of LXD deviations and classifications. Descriptive statistics were used to calculate the total average LXD deviations, average severity ratings, and average for the number of interaction within the learning environment and interaction within the learning space LXD problems that novices and experts identified. Results suggest that experience may impact the LXD deviation identification and classification, but there are no significant differences between groups on severity rating.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>2</td>
</tr>
<tr>
<td>Problem of Practice Statement</td>
<td>2</td>
</tr>
<tr>
<td>Conceptual Frameworks</td>
<td>5</td>
</tr>
<tr>
<td>Learning Experience Design (Tawfik et al., 2022a)</td>
<td>5</td>
</tr>
<tr>
<td>Purpose Statement</td>
<td>7</td>
</tr>
<tr>
<td>Questions</td>
<td>10</td>
</tr>
<tr>
<td>2. Review of the Literature</td>
<td>14</td>
</tr>
<tr>
<td>Introduction</td>
<td>14</td>
</tr>
<tr>
<td>Importance of Evaluation Designs of Learning Technologies</td>
<td>14</td>
</tr>
<tr>
<td>Efficient Strategies to Conduct Usability</td>
<td>16</td>
</tr>
<tr>
<td>Heuristic Evaluation</td>
<td>17</td>
</tr>
<tr>
<td>Cognitive Walkthrough</td>
<td>20</td>
</tr>
<tr>
<td>Conclusion of Heuristic Evaluations and Cognitive Walkthroughs</td>
<td>23</td>
</tr>
<tr>
<td>Unique Considerations of Usability for Learning Environments</td>
<td>23</td>
</tr>
<tr>
<td>Technological and Pedagogical Usability</td>
<td>25</td>
</tr>
<tr>
<td>Learning Experience Design to Address Learning Environment Interactions</td>
<td>26</td>
</tr>
<tr>
<td>Summary</td>
<td>33</td>
</tr>
<tr>
<td>3. Methodology</td>
<td>35</td>
</tr>
<tr>
<td>Introduction</td>
<td>35</td>
</tr>
<tr>
<td>Participants/ Learner Characteristics</td>
<td>36</td>
</tr>
<tr>
<td>Setting</td>
<td>37</td>
</tr>
<tr>
<td>Instrumentation/ Data Collection Methods</td>
<td>37</td>
</tr>
<tr>
<td>Analysis</td>
<td>44</td>
</tr>
<tr>
<td>Limitations</td>
<td>47</td>
</tr>
<tr>
<td>Ethical Considerations</td>
<td>48</td>
</tr>
<tr>
<td>Bias/Subjectivities</td>
<td>49</td>
</tr>
<tr>
<td>4. Results</td>
<td>50</td>
</tr>
<tr>
<td>Introduction</td>
<td>50</td>
</tr>
<tr>
<td>Participant Demographics</td>
<td>50</td>
</tr>
<tr>
<td>Instrument Validity with LXD Experts</td>
<td>53</td>
</tr>
</tbody>
</table>
Research Question 1: What is the difference, if any, between novices and experts on the percentage of LXD deviations (problems) identified? 54
Research Question 2: What is the difference, if any, between novices and experts on the average severity ranking of LXD deviations? 56
Research Question 3: What is the difference, if any, between novices and experts on the number of interaction within the learning environment problems vs. interaction within the learning space problems they identify? 58
Summary 63

5. Discussions and Conclusions 64
   Introduction 64
   Summary of Findings 64
   Discussion 71
   Conclusion 79

References 82

Appendices 99
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Tawfik et al., (2022a) Interaction within the Learning Environment Codes and Definitions</td>
<td>27</td>
</tr>
<tr>
<td>2. Tawfik et al. (2022a) Interaction within the Learning Space Codes and Definitions</td>
<td>28</td>
</tr>
<tr>
<td>3. Research Questions Alignment Statistics with Instrumentation</td>
<td>45</td>
</tr>
<tr>
<td>4. Participants’ Demographic Survey Data</td>
<td>52</td>
</tr>
<tr>
<td>5. Novice v. Expert Percentage of LXD Deviations Identified</td>
<td>55</td>
</tr>
<tr>
<td>8. Novice v. Expert within the Learning Space Identification</td>
<td>60</td>
</tr>
<tr>
<td>9. Inter-rater Reliability for PI and LXD Expert for Test Items</td>
<td>60</td>
</tr>
<tr>
<td>10. Inter-rater Reliability for PI and LXD Expert for Dummy Items</td>
<td>62</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figures</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Learning Journey Overview</td>
<td>39</td>
</tr>
<tr>
<td>2. Skills for Employability Dashboard</td>
<td>40</td>
</tr>
<tr>
<td>3. Research Question Alignment with Data Analysis</td>
<td>47</td>
</tr>
</tbody>
</table>
List of Abbreviations

Analyze, Design, Develop, Implement, and Evaluate (ADDIE)

Human Computer Interaction (HCI)

Learning Design and Instructional Design and Technology (LIDT)

Learning Experience (LX)

Learning Experience Design (LXD)

Interaction within the Learning Environment (ILE)

Interaction within the Learning Space (ILS)

Usability (UX)
CHAPTER ONE: INTRODUCTION

Practitioners in the field of learning design and instructional design and technology (LIDT) are often called upon to create learning materials, develop training courses, and provide insight into business training designs such as job-specific training and professional development (DeVaughn, 2022). One frequently used method by instructional designers is the Analyze, Design, Develop, Implement, and Evaluate (ADDIE) model (Branson et al., 1975), which is comprised of five phases: analysis, design, development, implementation, and evaluation (Branson et al., 1975; Peterson, 2003). Designing course content and learning systems has become an iterative process, allowing designers to gain insights to create pleasing and highly usable learning environments (Cennamo & Kalk, 2018; Schmidt et al., 2020). Due to the iterative nature of design, formative evaluations are essential to obtain feedback during the design process rather than at the end (DeVaughn & Stefaniak, 2021). The feedback received from these evaluations is crucial because it provides insight into intended learning outcomes, learner experience, and design usability (Schmidt et al., 2020).

As the field has evolved, studies find that practitioners receive user feedback beyond the final stage of the ADDIE process. Indeed, research suggests instructional designers use formative evaluation (continuous) more than summative evaluation (end-stage) during their design projects (DeVaughn & Stefaniak, 2021). Formative evaluation occurs continuously throughout the design process and gives the designer feedback on the design and crucial human-computer interaction elements and usability issues (Scriven, 1967; DeVaughn & Stefaniak, 2020). Alternatively, summative evaluation – such as usability testing – occurs at the end of the design process once
the design has been implemented (Scriven, 1967; DeVaughn & Stefaniak, 2020). Despite research suggesting designers prefer formative evaluation, instructional designers may only conduct evaluations once or twice a year or when requested (DeVaughn & Stefaniak, 2021). There are multiple reasons for this finding. When research examined the challenges prohibiting designers from performing the evaluation set of the design process, many cited high costs, lack of participants, and lack of support as hindering the evaluation process (DeVaughn & Stefaniak, 2021; Cennamo & Kalk, 2018).

Moreover, the design process is costly and often consumes time, resulting in limited resources being available for evaluation (Cennamo & Kalk, 2018). These challenges highlight a need to increase the efficiency of usability evaluations because current methods take too much time and expend too many resources (DeVaughn & Stefaniak, 2021; Cennamo & Kalk, 2018). Therefore, it is crucial to find the optimal type and number of participants needed to identify problems within the system for each design cycle to reduce the time spent on these evaluations and other resources, such as time and available evaluators (participants).

**Problem of Practice Statement**

Current literature is limited regarding the best evaluation practices for instructional designers (Armstrong, 2004). Griffin (2011) highlighted the imperative need for a theoretical framework for the field of evaluation; moreover, a single concept model for evaluation is needed so that instructional designers may develop “measures for learning acquisition and transfer of learning” (DeVaughn, 2020, p. 446). Due to this lack of theoretical framework and conceptual evaluation model, researchers in LIDT have begun borrowing evaluation methods from other
fields (Lu et al., 2022). One related field that has garnered traction for its application to the evaluation of learning technologies is the field of usability (Lu et al., 2022; Albert & Tullis, 2022). Usability testing methods are “a robust and effective user-centered evaluation method that has been broadly used across various disciplines to uncover design flaws related to ease-of-use, utility, and accessibility” (Schmidt & Earnshaw, 2023, p. 14). Although definitions vary, usability can be defined as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” (ISO 9241-210, para. 1). Usability evaluation, specifically technological usability, comprises a multi-faceted methodology that examines users’ experiences with the learnability, intuitiveness, satisfaction with a system or product (Moradian et al., 2018; Nielsen, 1994a, 1994b).

Many researchers contend that solely focusing on technological usability is inadequate for thoroughly evaluating learning technologies because technological usability was created initially to be broadly applicable to any system or user (Lu et al., 2022; Schmidt & Huang, 2021; Jahnke et al., 2020). Despite these challenges, increasing popularity of usability evaluations in the field of LIDT has led to many using Nielsen’s (1993) Five User Assumption to justify using small sample sizes to evaluate learning technologies (Tawfik et al., 2022; Faulkner, 2003). The Nielsen Five User Assumption posits that five participants will find 80% of the usability problems (Nielsen, 1993). The basis of this assumption is related to Virzi’s (1992) Law of Diminishing Returns, which suggests that once a certain number of participants is achieved, participants will no longer identify new problems; instead, their problem identification will overlap increasingly (Virzi, 1992). Given its purported efficiencies and acceptance within the
field, many practitioners employ Nielsen’s (1993) assumptions during formative and summative usability evaluations (Tawfik et al., 2022; Faulkner, 2003). Despite its application in practice, researchers consider Nielsen's (1993) assumptions controversial, and researchers in the field of Human-Computer Interaction (HCI) and usability continue investigating the necessary number of users needed to evaluate a program (Alroobae & Mayhew, 2014). The optimal number of participants necessary for a usability evaluation thus remains an ongoing debate in the field. As such, no consensus exists among researchers (AlRoobaea & Mayhew, 2014), especially in evaluating the usability of learning technologies. This debate has led to two camps emerging in the field of usability: one that argues five users is adequate to isolate most usability issues and another that argues that five is insufficient (Woolrych & Cockton, 2001).

Applying Nielsen's (1993) Five User Assumption to learning technologies is problematic in several ways. Nielsen’s (1993) Five User Assumption is a dated paradigm focused on technological usability and older technologies and does not consider advances such as multi-user collaboration, mixed realities, simulations, and others. Prior views on technological usability were broadly applicable to any user and system; current learning technologies are instead intended for a particular user (the learner) who completes specific learning tasks (Schmidt & Huang, 2021). Moreover, these learning interactions contain two parts – how the learner interacts with the content and how the learner interacts with the technology (Lu et al., 2022; Schmidt & Huang, 2021; Tawfik et al., 2022). However, technological usability was not intended to evaluate these complex interactions that occur when a learner interacts with a learning technology (Lu et al., 2022). Moreover, this assumption supposes that randomly selecting five users from the population will represent the overall population. The assumption does not consider the users’
experiences and how these experiences impact their evaluation and identification of usability issues.

**Conceptual Frameworks**

The five user assumption is derived from Nielsen's (1993) research; the assumption proposes that five testers are sufficient to identify 80% of usability problems. The purpose of this assumption was to exhibit that less rigor is necessary for usability testing (Faulkner, 2003). Research supporting the idea of less rigor is needed for usability studies, providing usability researchers the ability to reduce the sample sizes required for statistical power (Faulkner, 2003). This assumption is widely accepted in usability, mainly where the subject pool is limited (Faulkner, 2003). However, this assumption is now used for learning technologies, which is problematic because the assumption was not originally intended to evaluate the unique complex interactions that occur when interfacing with a learning technology.

**Learning Experience Design (Tawfik et al., 2022)**

Research has suggested that learning technologies contain unique complex interactions that go beyond the scope of technological usability testing (Lu et al., 2022; Schmidt & Huang, 2021; Tawfik et al., 2022). Investigating the interactions between the user and the interface and the user and the learning materials led to the development of the field of learning experience design (LXD) (Schmidt & Huang, 2021). However, learning experience design is still a developmental field; therefore, researchers have yet to agree upon a singular definition of Learning Experience Design. One proposed definition of LXD defines it as “a human-centric, theoretically-grounded, and socio-culturally sensitive approach to learning design, intended to
propel learners towards identified learning goals, and informed by UXD [user experience design] methods” (Schmidt & Huang, 2021, p. 141). Moreover, LXD is primarily concerned with the learner and their sociocultural context (Abramenka-Lachheb, 2023). LXD goes beyond the scope of existing instructional design theories by accounting for the end-users' actual experiences as they steer their learning, particularly in learning technologies (Abramenka-Lachheb, 2023). When designing for the optimal learning experience, LXD advises designers to consider three unique dimensions: the pedagogical dimension, the technological dimension, and the sociocultural dimension (Jahnke et al., 2020).

To gain further clarity on the aspects that comprise LXD and its dimensions, Tawfik et al. (2022), in a grounded-theory study, describe these LXD interactions as interaction with the learning environment and interaction with the learning space. Interaction with the learning environment addresses the technological usability side of the learning technology. In contrast, the interaction with the learning space goes beyond technological usability and addresses the pedagogical usability of the learning technology. Pedagogical usability focuses on the overall user experience with the entire learning environment (Jahnke et al., 2020; Marell-Olsson & Jahnke, 2019; Moore et al., 2014; Nakamura et al., 2018; Quinones et al., 2018). The two categories of interaction with learning environment/space are further broken down into nine sub-constructs. Interaction with the learning environment contains the following sub-constructs:

"customization of the interface, expectation of content placement, functionality of component parts, interface terms aligned with existing mental models, and navigation" (Tawfik et al., 2022, p. 9). In comparison, interaction with the learning space contains four sub-constructs:

"engagement with the modality of content, dynamic interaction, the perceived value of
technology feature to support learning, and scaffolding" (Tawfik et al., 2022, p. 9). These two interactions within learning technologies may result in Nielsen’s (1993) Five User Assumption not applying to learning technologies because these interactions go beyond solely technological usability.

**Purpose Statement**

As noted earlier, finding efficient ways to collect user data in formative and summative evaluations is imperative. Despite current literature suggesting that learning technologies contain usability and learning experience interactions, Nielsen's (1993) Five User Assumption is still used to justify testing educational technologies with small sample sizes (Tawfik et al., 2022; Schmidt & Huang, 2022). However, the rigor of Nielsen's (1993) Five User Assumption has yet to be tested for learning technologies. Therefore, concerning learning technologies, we are still determining if five users will provide the statistical rigor to identify most of the LXD problems present within a learning technology.

Another issue relates to the unique considerations that are inherent as users employ learning technologies, such as prior knowledge, ease of use, and others. Considering prior knowledge, one might argue that novices may identify more interaction with the learning environment errors, and experts identify more interaction with the learning space errors. Given their advanced knowledge, one might propose that fewer experts are needed to identify the same number of problems as a larger group of novices. Another might argue that both are needed to identify the problems that can arise within a learning technology. More research is necessary to test the use of Nielsen’s (1993) Assumption for testing learning technologies. This information is
essential to improve inefficient, time-consuming, expensive testing practices that diminish crucial resources such as the participant pool.

This causal-comparative study aims to identify the average number of LXD deviations that LXD novices and LXD experts identify within a learning technology. Moreover, the researcher seeks to examine the possible cause-and-effect relationship between usability evaluation experience/expertise and the average number of usability deviations identified; additionally, an exploration concerning the LXD construct classification and severity rating of these usability deviations (problems) was conducted at business location in the Mid-South region of the United States. The independent variable of usability experience will divide participants into two groups: LXD novices and LXD experts.

Regarding LXD expert classification, many cite the Ericsson et al. (1993) criterion for expertise when applying expert classification to an individual. Ericsson et al. (1993) posit that a person must have 10,000 hours of deliberate practice in the area they wish to be considered an expert. However, researchers have called this criterion for expert classification into contention (Harris & Eccles, 2021). Many argue that while practice is necessary to reach expert performance, other factors hinder or expedite attaining this level of performance (Harris & Eccles, 2021; Ackerman, 2014). Moreover, despite extensive discussion to operationally define expertise, the operationalization of expertise and what constitutes deliberate practice remain areas of debate in expertise research (Harris & Eccles, 2021; Ericsson et al., 1993; Ericsson & Charness, 1994). Considering these areas of contention, for this paper, a modified version of the Hmelo-Silver et al. (2007) classification for experts will be used. These researchers classified experts as those with 10 or more years of experience and an advanced degree in the subject
matter. Specifically, those classified as experts for this study will have 10 or more years of experience conducting usability evaluations or an advanced degree (master’s or doctorate) in IDT. The reasoning for this modification is to include those that meet the experience criteria but do not have an advanced degree in IDT. Although these potential experts do not have an advanced degree, their many years of experience allow for their inclusion into this category. Alternatively, possible experts could also include those with an advanced degree. These individuals would be included because throughout obtaining their degree they gain knowledge regarding evaluating learning technologies; moreover, those with an advanced degree that work as teachers are constantly evaluating materials and technologies to use in the classroom. These experiences provide them with a bases to conduct LXD evaluations. Novices will be classified as those with little or no experience with usability evaluations such associates working in a Talent Acquisition department.

Regarding dependent variables, three dependent variables will be examined in this study. The first dependent variable is user identification of perceived learning experience design deviations (problems). Learning experience design deviations refer to a component of the learning technology that hinders user progress or learning. The second dependent variable is the severity rating of each deviation using Nielsen's (1992) severity rating system. The third dependent variable is user classification of LXD deviations. These classifications will be derived from Tawfik et al.’s (2022) Learning Experience Design constructs. This study used a convenience sample of 10 participants (five novices and five experts) to examine the differences between novices and experts in usability deviation identification and classification.
Question(s)

Research Question 1. What is the difference, if any, between novices and experts on the percentage of learning experience design deviations (problems) identified?

Research Question 2. What is the difference, if any, between novices and experts on the severity ranking of learning experience design deviations?

Research Question 3. What is the difference, if any, between novices and experts on the number of interaction within the learning environment problems vs. interaction within the learning space problems they identify?
Definitions

ADDIE. Analyze, design, develop, implement, and evaluate. A method instructional designers use to create and evaluate design (Branson et al., 1975; Peterson, 2003).

Constructivism. A theoretical framework posits that learners’ previous experiences influence how users construct knowledge (Jonassen, 1991).

Expert. Those classified as experts for this study will have 10 or more years of experience conducting usability evaluations or an advanced degree (master’s or doctorate) in IDT (Hmelo-Silver et al., 2007).

Formative Evaluation. Formative evaluation occurs continuously throughout the design process and provides the designer with feedback that they can implement during the design process (Scriven, 1967; DeVaughn & Stefaniak, 2020).

Interaction within the Learning Environment. An overarching construct of Learning Experience Design that contains five codes: “customization of the interface, functionality of component parts, expectation of content placement, navigation, and interface terms aligned with existing mental models” (Tawfik et al., 2022, p. 9). These codes address the functionality of the interface and user experience (Tawfik et al., 2022). This construct addresses the technological usability of the learning technology.

Interaction within the Learning Space. An overarching construct of Learning Experience Design that contains four codes: “engagement with modality of content, dynamic interaction, perceived value of technology feature to support learning, and scaffolding” (Tawfik et al., 2022, p. 9). These codes address the incidence of learning and non-learning that is supported by the product or system (Tawfik et al., 2022). This
construct focuses on the content, user supports, and assessments present within the learning technology.

**Learning Experience Design.** For this paper, it will be broadly defined using two broad constructs: "interaction with the learning environment" and "interaction with the learning space" (Tawfik et al., 2022, p. 1).

**Learning Technology.** A technology designed with the purpose to enhance and improve learning (Liu et al., 2020).

**Nielsen’s Five User Assumption.** The Nielsen Five User Assumption posits that five participants will find 80% of the usability problems (Nielsen, 1993). This assumption is often used to justify using small sample sizes in usability testing (Faulkner, 2003).

**Novice.** Novices will be classified as those with little or no experience with usability evaluations such as students in an introduction to Technology and Education course (Hmelo-Silver et al., 2007).

**Pedagogical Usability.** A type of usability that focuses on the overall user experience in relation to the full learning environment (Jahnke et al., 2020).

**Summative Evaluation.** A method of evaluation used after the design has been implemented to gather feedback to evaluate the product or system (Scriven, 1967; DeVaughan & Stefaniak, 2020).

**Usability.** Usability is outlined as the “extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified content of use” (ISO 9241-210, para. 1).
Technological Usability. A multi-faceted methodology that examines users’ experiences with learnability, intuitiveness, and satisfaction with a system or product (Moradian et al., 2018; Nielsen, 1994a, 1994b).
CHAPTER TWO: REVIEW OF THE LITERATURE

Introduction

Researchers agree that evaluation is an integral part of the design process (DeVaughn & Stefaniak, 2021; Schmidt et al., 2020); however, contentions exist regarding the best practices to effectively and efficiently evaluate learning technologies (Lu et al., 2022; Janke et al., 2020). These concerns have led to the development of various learning experience design frameworks (Jahnke et al., 2020; Marell-Olsson & Jahnke, 2019; Moore et al., 2014; Nakamura et al., 2018; Quinones et al., 2018). Moreover, to address this evaluation gap, researchers have borrowed usability principles and methods to apply to learning technologies (Lu et al., 2022). However, applying these UX principles and methods to learning technologies can be problematic because many applied usability paradigms are dated and may not be appropriate for learning technologies. Hence, the prevailing assumptions about how many users to test may be inadvertently applied differently than originally intended, especially as they are used for learning technologies. This chapter will discuss various usability evaluations and their applications to learning technologies. Moreover, the fallacies of the broad acceptance of these principles regarding learning technologies and the relationship between LXD and UX will be covered.

Importance of Evaluation Designs of Learning Technologies

To account for the rapid rise of technology, designers developed an iterative process containing multiple design and development cycles (Cennamo & Kalk, 2018). These design and development cycles depend on continuous feedback to isolate problems within the learning technologies and to examine whether the learner perceives the designer’s learning goals for the program as usable (DeVaughn & Stefaniak, 2021). These include issues with content delivery
but also usability challenges. As the connection between technology and education continues to grow, change, and develop rapidly, demands have increased on instructional designers to expedite the development process of learning technologies (Cennamo & Kalk, 2018). Along with demands for a more rapid design process, the field of educational technology is calling on designers to create more robust methods to assess the learning outcomes of students (Stefaniak et al., 2020). Beyond just conceptual knowledge, an essential aspect of this design process is the evaluation of the design as it can “identify best practices and opportunities for improvement and thus optimize scarce resources” (Marshall & Rossett, 2014, p.2). In educational and training settings, evaluation often examines whether learning outcomes and instructional objectives have been met (Devaughn & Stefaniak, 2021). In multiple studies, instructional designers highlight the importance of continuous feedback during the design process, especially as they are being called on to rapidly design various technologies and programs (Devaughn & Stefaniak, 2021; Devaughn & Stefaniak, 2020). Continuous evaluation thus allows for improvements to be made to the design, to facilitate successful learning (Tang et al., 2021).

Despite instructional designers acknowledging the importance of evaluations, research suggests a disconnect in the degree to which it is implemented in everyday practice (Devaughn & Stefaniak, 2020; William et al., 2011). Literature exploring the disconnect between recognition of the importance of evaluation and the performance of these formal evaluations cites several challenges instructional designers experience when attempting to evaluate (Devaughn & Stefaniak, 2021; Devaughn & Stefaniak, 2020; Stefaniak et al., 2020; Paz & Pow-Sang, 2014; Marshall & Rossett, 2014). For example, designers cite challenges such as formal evaluation being costly, lack of participants, little stakeholder interest, and the failure of
organizations to recognize the importance of evaluation (Devaughn & Stefaniak, 2021; Devaughn & Stefaniak, 2020; Marshall & Rossett, 2014; William et al., 2011). Moreover, designers have been met with resistance from organizations because the organizations felt that evaluations were personal critiques. Organizations assumed these evaluations were unnecessary if the course had been designed correctly and thus a waste of resources (e.g., time and money) (DeVaughn & Stefaniak, 2020). This lack of resources and support has led designers to move away from formal evaluations, which rely on extensive data collection, and instead gravitate to informal and efficient evaluation (Devaughn & Stefaniak, 2020; Marshall, 2014; Willams et al., 2011). This shift to ad-hoc and efficient evaluation has led to evaluation methods lacking standards and consistency (Devaughn & Stefainak, 2020). This lack of research means the effectiveness of instructional designers' programs and products is not consistently evaluated, and little information about the lack of best practices (Richey & Klein, 2005; Armstrong, 2004). Moreover, the lack of consistent evaluation methods has led to designers borrowing methods from other fields, which is problematic because these methods were not designed for these purposes (Schmidt et al., 2020).

**Efficient Strategies to Conduct Usability**

Instructional designers are responsible for curricular decisions, along with how to apply strategies within a learning environment. Therefore, a critical method for evaluating learning technologies is a usability assessment. A usability evaluation aims to assess whether the intended user can use the technology to meet set goals effectively within a specific context of use (ISO 9241-210). For example, these evaluations highlight the necessity of interfaces to provide a positive experience by being easy for users to use to achieve their goals (Paz & Pow-Sang,
Usability evaluation methods aid in discovering human-computer interaction challenges, which can hinder the user experience and the achievement of goals (Alao et al., 2022). Additionally, usability evaluation methods provide insight into real-time user behavior, highlight areas that need improvements to make the interface more efficient (Alao et al., 2022), and provide valuable feedback regarding whether users perceive design intentions/goals. In doing so, usability evaluations give the designers essential feedback, such as the user experience interacting with the design, user engagement, and opportunities for improvement within the interface and the learning content.

Various approaches have emerged among practitioners, given the importance of usability in evaluating technology. During a systematic review by Paz and Pow-Sang (2014), the researchers identified the most common usability evaluation methods: heuristic evaluation, questionnaire, cognitive walkthrough, and cognitive think-aloud. The paper noted how heuristic and cognitive walkthroughs are incredibly efficient and often apply the Nielsen Five User Assumption. Given its assumption application, the heuristic evaluation and cognitive walkthrough are discussed in more detail below.

**Heuristic Evaluation**

Heuristic evaluations are considered an efficient method for evaluating technologies (Nielsen & Mack, 1994; Molich & Nielsen, 1990). This classification is primarily because heuristic evaluations usually rely on a few usability experts to evaluate a technology against an already selected set of usability principles (Schmidt et al., 2020; Gupta, 2015). In many cases, these experts independently evaluate an interface design before meeting as a group to resolve differences in their findings (Schmidt et al., 2020). This evaluation type is often favored because
it is less time-consuming, more cost-efficient, and can be performed during the in-phase of the design cycle (Schmidt et al., 2020; Alonso-Rios et al., 2018; Geng & Tian, 2016). For instance, Jahnke (2022) conducted a case study to showcase learning experience design and its evaluation methods. The research team was tasked with developing a digital learning system that educated learners on food safety statutes and risk categorization. As part of the evaluation process, the research team performed a heuristic evaluation of the prototype of the digital learning system. This evaluation revealed challenges with accessibility features, unbalanced text and video distribution, inadequate error support, and illegible text. These results guided the research team to improve the digital learning system in the subsequent design phase (Jahnke, 2022). While it afforded an efficient approach to evaluation, the perspective of potential end-users is missing due to the heuristic evaluation being conducted by only the research team.

Heuristic evaluations are applied to various learning technologies, such as educational games (Verkuyl et al., 2016). Researchers developing a game focused on teaching and practicing pediatric nursing skills used an expert heuristic evaluation along with a user usability test to evaluate the usability of the technology (Verkuyl et al., 2016). The user usability testing provided a different perspective on the game's usability. The students and nurses who tested the game provided feedback on the dynamic interaction of the simulation, which motivated them to continue to the end of the game. Additionally, students and nurses testing the game provided recommendations to enhance the game, such as diverse difficulty levels to offer higher levels of challenge (Verkuyl et al., 2016). The usability experts focused more on traditional usability components, such as navigation and content placements. The results of this study suggest the need for multiple types of evaluators when evaluating the usability of a learning technology. The
experts can identify problems within the system. However, they are limited in being able to provide recommendations for improvement because they do not have the experiences that the students and nurses have to guide them. It is important to note that this study brings awareness to how user experiences affect their perspectives and interactions with a system. A diverse sample of evaluators may be necessary to evaluate learning technologies effectively.

Although heuristic evaluations are mainly conducted with experts, researchers have begun comparing these expert evaluations with end-user (novice) evaluations (Chang & Johnson, 2021). For example, Chang and Johnson performed a study to evaluate the usability of a game created to teach students about carbon footprints and CO2 emissions. One part of the usability evaluation contained a heuristic evaluation. Four experts with backgrounds in game design, multimedia design, and e-learning independently completed heuristic evaluation questionnaires on the learning technology focused on material use and design. In addition, forty-five end-users (novices) conducted a heuristic evaluation; however, the end-user evaluation focused on user experiences and perspectives regarding the learning technology. These evaluations revealed inconsistencies in interface styles/game scenes, insufficient guidance and operational efficiency, inconsistent support and error guidance, improper error feedback, and poor interaction in instructional videos (Chang & Johnson, 2021). This study contributes to the literature by highlighting that usability evaluators should go beyond the interface's usability and focus on the learning technology’s effectiveness on learning. Despite these contributions, questions remain about usability evaluations. For instance, because the authors provided the experts and the end-users with heuristic evaluation questionnaires with different focuses, they could not compare the performance differences between the two groups. It remains unclear if experts and end-users...
would provide similar evaluations or if the differences in their level of expertise would affect their evaluations.

**Cognitive Walkthrough**

A cognitive walkthrough is another popular method for evaluating usability (Lu et al., 2022; Paz & Pow-Sang, 2014). This inspection method relies on evaluators rather than end-users to perform realistic tasks and provide their judgment (Lewis & Wharton, 1997; Rieman et al., 1995; Wharton et al., 1992; Lewis et al., 1990). Therefore, the data gained from this type of evaluation is not based on actual intended user data; instead, the data is gathered from evaluator judgments (Schmidt et al., 2020). A cognitive walkthrough consists of evaluators performing a pre-selected set of tasks within a technology and identifying the correct paths to accomplish the task; however, the evaluators’ actions may not reflect what a real user will do when interacting with the system (Lewis et al., 1990; Wharton et al., 1992; Reiman et al., 1995). As evaluators complete these tasks, they note and consider the ease of potential users to perform the given tasks (Gupta, 2015; Beer et al., 1997). Literature suggests that cognitive walkthroughs provide insight into various aspects of the interface, such as first impressions, ease of use, interface functions alignment with user mental models, user expectations, and familiarity with terminology (Schmidt et al., 2020).

In addition to the aforementioned heuristic evaluations, cognitive walkthroughs also have a literature base associated with usability of learning technologies. Doumanis and Economou (2019) used a modified cognitive walkthrough method with six evaluators to assess the usability of a virtual reality learning technology. The learning technology provided an immersive world for students to interact with teachers and other students as an avatar. The modified cognitive
walkthrough begins with analyzing two pre-selected tasks: a virtual debate and a learning module on historical artifacts. The evaluators completed the tasks and answered four follow-up questions for each task. Two evaluators acted as students, and the other acted as a teacher for each scenario. Results indicated 47 usability problems were identified in the VR learning technology, with most problems in the more complex online debate scenario. Evaluators acting as teachers assigned different priority levels to the usability issues; however, evaluators acting as students rated all usability problems equally important. These results highlight how different perspectives can influence how usability problems may be perceived by other users (Doumanis & Economou, 2019). However, due to the low number of evaluators, it remains unclear if these differences would have been more pronounced in a larger group of evaluators with diverse backgrounds and roles. The use of small samples in types of evaluations could cost designers and researchers valuable data.

Additional cognitive walkthrough studies also use small sample sizes to evaluate learning technologies. For instance, Floreste et al. (2023) recruited three Human-Computer Interaction (HCI) experts to assess their learning game prototype that focuses on teaching learners about mental health and the professions associated with this topic. Learners enter a virtual world and become the character Bryan. The learners complete various tasks, such as going to a friend’s house and conversing with that friend to learn more about mental health. Evaluators were provided with a task protocol to navigate through the interface and answer four questions developed by the research team. This evaluation revealed 11 problems when evaluators tried to accomplish the tasks on the provided protocol. These usability problems pertained to issues with interface functionality and lack of proper support to aid users in understanding the scenarios and
how to use the interface (Floрестe et al., 2023). While this provides further evidence of cognitive walkthroughs applied to different modalities, the results of this study have limited applicability; due to the sample only including three expert evaluators with HCI backgrounds, their perspectives could be biased when evaluating the interface. The evaluators’ expertise may cause their perspectives to not align with the intended end-users' perspectives and lead to usability issues being overlooked.

Throughout the cognitive walkthrough and learning technology literature, the trend of small sample sizes continues, especially in cognitive walkthroughs of educational video games. For example, Hutagalung and Riwinoto (2023) analyzed user experience on the educational video game *Indonesian Heritage*. In the game, users play as a prince who must travel through several kingdoms and defeat enemies to save a goddess. Throughout these adventures, users learn about Indonesian culture and heritage. To evaluate the user experience of the games, researchers recruited six evaluators to perform a cognitive walkthrough of the game. Evaluators were asked to use a trial-and-error method to complete five tasks: Start the game, complete the learning tutorial, change game settings, play the game, and quit the game. Evaluators were given five follow-up questions on function analysis and five questions on operations analysis for each task. Results from these evaluations indicated that the game did not provide enough clues regarding gameplay functions, such as mission details and weapon characteristics. However, researchers concluded that the lack of clues did not affect the primary function of the game and that the game was easy for both experienced and new players (Hutagalung & Riwinoto, 2023). Similar to previously mentioned studies, only a handful of evaluators were used to evaluate the usability of this learning technology. Therefore, caution must be exercised when determining if
this limited sample could identify the majority of problems users will experience within the learning technology.

**Conclusion of Heuristic Evaluations and Cognitive Walkthroughs**

Learning technologies, especially game-based learning, are becoming more frequently used to educate learners (Floreste et al., 2023; Hutagalung & Riwinoto, 2023; Doumanis & Economou, 2019). Heuristic and cognitive walkthrough evaluations are popular for evaluating these learning technologies (Lu et al., 2022; Paz & Pow-Sang, 2014). However, these evaluation methods may be problematic due to their association with small sample sizes. This association comes from these methods being borrowed from the field of usability, which often cites Nielsen’s Assumption (1993) to justify the use of small sample sizes (Faulkner, 2003). Researchers must explore whether this assumption, developed with complex systems did not exist, is applicable to learning technologies.

**Unique Considerations of Usability for Learning Environments**

*Nielsen’s Assumption Debate*

An essential distinction in the aforementioned approaches includes the type and number of users to employ for evaluation. As previously discussed, instructional designers borrowed evaluation methods from the field of usability in an attempt to effectively evaluate learning technologies (Lu et al., 2022). Along with borrowing these evaluation methods, many in the field of instructional design began using Nielsen’s Five User Assumption to justify the use of small sample sizes (Faulkner, 2003). The assumption argues that five users can identify 80% of usability issues. In recent years, the validity of Nielsen’s Assumption has become an object of debate. Scholars caution practitioners to reconsider this widely and readily used assumption (Faulkner,
Despite being frequently and widely used to justify small samples when completing usability evaluations, Nielsen’s Assumption has several fallacies, especially when applied to learning technologies. The first premise that must be considered is that this assumption is dated (Faulkner, 2003). Nielsen developed this assumption when usability concepts were still in their infancy (Faulkner, 2003). The technologies evaluated during this period were less complex than current learning technologies. These usability evaluations were not initially intended for the technologies they are being applied to presently. Therefore, it stands to reason that this assumption may not be applicable when evaluating complex technologies.

Another flaw within Nielsen’s (1993) assumption is the idea that five randomly selected evaluators can effectively evaluate a learning technology. Not only does the type of evaluation being performed affect the sample size, but research suggests that user experience and user expertise affect the number of evaluators needed to effectively perform a usability evaluation (Nasir et al., 2021; Abulfaraj & Steele, 2020; Othman et al., 2020; Verkuyl et al., 2016). Moreover, these studies have suggested that the type of evaluator, such as expert or novice, impacts not only the type of usability problems identified but also the feedback these evaluators provide (Nasir et al., 2021; Abulfaraj & Steele, 2020; Othman et al., 2020; Verkuyl et al., 2016). Considering these limitations, it is necessary to test whether Nielsen’s (1993) Assumption remains true when evaluating a learning technology and how having evaluators of different experience levels (expert or novice) impacts problem identification, severity ratings, and classification.

**Technological and Pedagogical Usability**
Although the aforementioned evaluations provide designers with feedback on learning technologies, designers must consider if these evaluations are sufficient to evaluate the unique complexity that learning technologies pose. Initially, traditional usability evaluations focused mainly on the technological aspects of an interface (Lu et al., 2022; Schmidt & Huang, 2021). Technological usability was initially designed for ergonomics rather than for learning, which makes its sole application to learning technologies inadequate (Schmidt & Huang, 2021). Over time, traditional or technological usability elements included efficiency, effectiveness, and user satisfaction (Schmidt & Huang, 2021). Many scholars argue that using solely technological usability to evaluate learning technologies is insufficient (Jahnke et al., 2020; Schmidt & Huang, 2021). To address the limitations of technological usability, scholars have argued for the development of pedagogical usability to address the additional unique interactions that occur within learning technologies (Costabile et al., 2005; Nokelanin, 2006; Reeves, 1994; Silus et al., 2003; Squires & Preece, 1996, 1999). In contrast to the feature-centric approach to technical usability, pedagogical usability focuses on the overall user experience within a learning environment (Jahnke et al., 2020; Marell-Olsson & Jahnke, 2019; Moore et al., 2014; Nakamura et al., 2018; Quinones et al., 2018).

As learning technologies acknowledge the importance of usability and HCI, several frameworks for pedagogical usability exist (Tang et al., 2021). One fundamental framework widely used in the field is Nokelainen’s ten criteria of pedagogical usability framework (Lu et al., 2022; Tang et al., 2021). This framework includes the following criteria: learner control, learner activity, cooperative/collaborative learning, goal orientation, applicability, added value,
motivation, valuation of previous knowledge, flexibility, and feedback (Nokelainen, 2006). Some researchers have called for expanding this framework (Vuorio et al., 2017).

In a study, Vuorio et al. (2017) expanded Nokelainen’s framework by combining the Pedagogical Usability Framework, the Capability Maturity Model (CMM), and the User Experience. The Pedagogical Usability Framework is based on Nokelainen’s (2006) work, and the Capability Maturity Model examines the “organizational maturity in order to successfully implement and develop educational technology in organization” (Vuorio et al., 2017, p.2). The learner experience (UX) evaluation uses the SUXES evaluation model, which measures user experiences and expectations to support their effect (Vuorio et al., 2017). Results from Vuorio et al. (2017) research indicated that participants felt that there should be criteria to capture the teacher’s use of the learning application to support learning outcomes and the teacher’s role in the learning process. Moreover, this research indicated that Nokelainen’s ten criteria were valid; however, the participant results from this study suggest that Nokelainen’s (2006) ten criteria did not capture all the criteria that comprise the overall experience (Vuorio et al., 2017).

**Learning Experience Design to Address Learning Environment Interactions**

As previously discussed, learning environments contain unique interactions that go beyond the limits of technological usability (Lu et al., 2021; Schmidt & Huang, 2021). To address these limitations, scholars have called for further development of pedagogical usability principles (Costabile et al., 2005; Nokelainen, 2006; Reeves, 1994; Silus et al., 2003; Squires & Preece, 1996, 1999). However, recent research studies have suggested that current pedagogical usability frameworks do not fully capture the overall learning experience and do not fully address the interactions that occur when a learner interfaces with a learning environment (Lu et
al., 2022; Tang et al., 2022; Vurorio et al., 2017). In 2021, Tawfik et al. (2022) conducted a grounded theory study to examine the constructs that comprise learning experience design (LXD). Participants in the study completed tasks within a learning technology that ranged from navigation to learning tasks such as problem-solving (Tawfik et al., 2022). Participants were asked to think aloud as they completed the tasks, and these responses were recorded and later coded by researchers. Results from the study suggested that learning experience design comprises two broad overarching constructs based on interactions within learning technology:

(a) *interaction with the learning environment* and (b) *interaction with the learning space*. In this context, interaction with the learning environment (the technology) is the user interaction with the technology itself through means such as navigation, content placement, and functionality of interface components (Tawfik et al., 2022). In doing so, the conceptual framework applies technical and pedagogical usability aspects.

**Table 1**

*Tawfik et al., (2022) Interaction within Learning Environment Codes and Definitions*

<table>
<thead>
<tr>
<th>Interaction within the learning environment (UX)</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Customization of the interface</td>
<td>User autonomy to control elements of the interface as needed (e.g., closed captions, tutorial, search bar, other learning tools)</td>
</tr>
<tr>
<td>2. Expectation of content placement</td>
<td>Where the user expects items to be; suggestions for new sections</td>
</tr>
<tr>
<td>3. Functionality of component parts</td>
<td>Functionality of items that are present</td>
</tr>
</tbody>
</table>
4. Interface terms aligned with existing mental models Labels are consistent with previous learning interactions (e.g., ebook, learning websites)

5. Navigation How users move from one location to another on the site.

Table 2

*Tawfik et al. (2022) Interaction within Learning Space Codes and Definitions

<table>
<thead>
<tr>
<th>Interaction within the learning space (LX)</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Engagement with modality of content</td>
<td>Thoughts on learning design format, aesthetics, and users’ desire to engage with elements used on learning interfaceIL</td>
</tr>
<tr>
<td>2. Dynamic interaction</td>
<td>Interaction that engenders learners’ desire to continue or discontinue in their self-directed learning</td>
</tr>
<tr>
<td>3. Perceived value of technology feature to support learning</td>
<td>Perceived value of a specific technology feature impacted one's learning</td>
</tr>
<tr>
<td>4. Scaffolding</td>
<td>Cues, hints, etc. that expanded the learners’ prior knowledge.</td>
</tr>
</tbody>
</table>

Although LXD represents a new approach to evaluating learning technology, the literature suggests that Nielsen's (1993) assumptions are still used to justify small sample sizes in usability evaluations. For instance, Carey and Stefaniak (2018) conducted a study to explore and evaluate the use of badges in higher education using open-ended structured interviews. A purposeful convenience sample of ten experts was used to gather insights. Results from these
interviews suggest that participation badges were not as valued as skill-based badges, while top focuses for badge implementation included learning objectives, badge purpose, and transferability (Carey & Stefaniak, 2018). Despite these results, the small sample size restricts the ability of these results to be generalized; moreover, the sample size limits the ability of the researchers to gain perspectives from a more diverse sample. Regarding previously mentioned studies, restricting the sample results makes it difficult to evaluate the technology being discussed comprehensively. Although this study does not directly cite Nielsen’s (1993) Five User Assumption as justification for its small sample size, it still provides evidence that these small sample size studies are present throughout the LXD and UX literature. To best design for diverse learning populations, researchers must examine whether these small sample sizes for evaluation provide designers with the most valuable insights needed to best design for the intended end-user.

Additional studies in the LXD field concerning UX evaluations and learning techniques provide further evidence of the frequency of small sample sizes used. For instance, Tawfik et al. (2024) examined how the learning experience design of computational thinking novices compared to that of more advanced computational thinkers when interacting with a block coding learning tool explicitly designed to aid computational thinking. Ten participants (five novices and five advanced learners) completed set tasks within the learning tool Blockly while following a think-aloud protocol provided by the researcher. After completing the set tasks, participants completed a post-hoc interview, answering questions about their experience using the tool. The data gathered from these interviews was then coded by researchers using Tawfik et al. (2022) LXD constructs. Although there were similarities in some subconstructs (such as engagement
with modality of content and perceived value of technology features), novices wanted more scaffolding regarding how to perform actions within the tool, such as using the play button to run the code; advanced users did not report this need because their prior knowledge of coding allowed them to use the tool more effectively. Concerning the *perceived dynamic interaction* construct, more advanced users participated in iterative testing within the interface without prompting, while novices wanted the system to initiate the next interaction (Tawfik et al., 2024). The results from the study provide valuable insights regarding how prior knowledge impacts the learners’ experiences while engaging with a learning tool. However, similar to Carey and Stefaniak (2018), the results of this study are limited due to the small sample size. Tawfik et al. (2024) use Nielsen (2000) to justify using a small sample to evaluate the learning tool. However, as previously discussed, this paradigm is dated and was not intended for use on advanced learning technologies. In addition to the small sample size being used in the aforementioned higher education LXD studies, the five user assumption is also applied to other user groups, such as neurodiverse learners. Schmidt et al. (2023) performed a LXD study to gain insight from autistic users to develop a virtual reality technology that is accessible, usable, and sensitive to the needs of these end-users. As part of the study, a usability test was performed to evaluate the usability of the technology. Five participants were selected to participate in this evaluation demonstrating that the practices of UX (i.e., a small sample) are present within an LXD study. Results from this usability test reported an above-average perceived usability and a high satisfaction, acceptability, and relevance among users (Schmidt et al., 2023). Because of the small sample size, it cannot be said if the perspectives of these five evaluators will accurately represent the potential perspectives of the diverse population of intended end-users. A more
extensive and diverse sample could provide invaluable insight into the learner experience that this small sample could not offer.

One final example of the frequent use of small sample sizes when evaluating learning technologies within the field of LXD is Huang (2023). This research aimed to design and evaluate an English language learning technology that used a virtual world game to aid in learning English. The researcher used various LXD methods to identify usability problems and assess the usability of the learning technologies. The LXD methods used for this study included a heuristic evaluation (four participants), empathy interview (four participants), concurrent think-aloud (five participants), and cognitive walkthrough (four participants). As in the case of earlier studies, convenience sampling and small sample sizes were used during the LXD study despite ongoing debate suggesting that more users are needed to provide a comprehensive evaluation of learning technologies (AlRoobaea & Mayhew, 2014; Faulkner, 2003; Woolrych & Cockton, 2001).

In conclusion, the previously discussed research studies highlight the pervasiveness of using small sample sizes within LXD studies when evaluating learning technologies. LXD developed out of a need to assess the complex interactions that occur within learning technologies; however, LXD has become embedded within UX methodologies and led to acceptance of assumptions, such as Nielsen’s (1993) Five User Assumption, without testing if this assumption is still valid with current learning technologies. Lu et al. (2022) conducted a literature review regarding usability research in educational technology. Despite LXD emerging as a new approach to the evaluation of learning technology, results from this study revealed several limitations regarding usability research and educational technologies. These limitations
included procedural and conceptual flaws, misunderstandings regarding usability evaluation methodology, unsuitable use of usability methods, and repeated use of small sample sizes. These findings suggest that current usability and learning technology literature results may be unreliable and problematic (Lu et al., 2022). Several limitations regarding usability research and educational technologies exist. These limitations included procedural and conceptual flaws, misunderstandings regarding usability, evaluation methodology, unsuitable use of usability methods, and repeated use of small sample sizes. These findings suggest that results from current usability and learning technology literature may be unreliable and problematic (Lu et al., 2022). LXD and UX have become highly correlated with one another, but researchers must exercise caution when applying UX principles and methods to LXD evaluations.
Summary

Throughout the literature, researchers have highlighted the importance of evaluation in the design process (Lu et al., 2022; Schmidt et al., 2020). Instructional designers agree that evaluation is a vital part of the design process; however, a lack of an evaluation framework for the evaluation of learning technologies has led to evaluation methods and practices being borrowed from the UX field (Lu et al., 2022; Albert & Tullis, 2022). Despite acknowledging that learning technologies contain interactions that go beyond the intentions of usability, researchers throughout the literature are using UX principles and theories such as Nielsen’s (1993) Five User Assumption to justify their evaluation practices (Carey & Stefaniak, 2018; Huang, 2023; Schmidt et al., 2023; Tawfik et al., 2022; Tawfik et al., 2024). Specifically, as a result of Nielsen’s (1993) Assumption and the connection between LXD and UX, the use of small sample sizes in evaluation practices has become the norm (Carey & Stefaniak, 2018; Huang, 2023; Schmidt et al., 2023; Tawfik et al., 2022; Tawfik et al., 2024). This normalization of small sample sizes is problematic because Nielsen’s (1993) Five Assumption has not been validated for application to learning technologies.

Testing the validity of this assumption necessary to evaluate learning technologies efficiently and effectively is an ongoing debate. Researchers suggest that learning technologies pose unique complexities beyond the scope of traditional UX evaluations (Jahnke et al., 2020; Schmidt & Huang, 2021). Research indicates that not only is the usability of technology essential, but designers must consider the user experience that occurs when interacting with a learning technology (Lu et al., 2022; Schmidt & Huang, 2021). Researchers have tried to address this gap by combining multiple usability methods to gain a more comprehensive evaluation and
applying various frameworks to learning technology evaluation (Huang, 2023; Jahnke et al., 2020). However, throughout the LXD literature, studies attempting to address learning technology evaluations continue using small sample sizes without validating if small samples provide the rigor necessary to evaluate these technologies properly. Future studies should examine whether variations of evaluator type and number of evaluators significantly impact the results of learning technology evaluation. (Carey & Stefaniak, 2018; Huang 2023; Schmidt et al., 2023; Tawfik et al., 2022).
CHAPTER THREE: METHODOLOGY

Introduction

Literature concerning the best evaluation practices for instructional designers to evaluate educational technology is sparse (Armstrong, 2004). Although LXD has evolved to advance evaluation practices for learning technologies, fallacies, and misconceptions from the field of UX are present throughout the literature (Carey & Stefaniak, 2018; Huang, 2023; Schmidt et al., 2023; Tawfik et al., 2022; Tawfik et al., 2024). The strong association between LXD and usability evaluations has led to misconceptions about UX practices becoming prevalent as they are applied indiscriminately within LXD (Lu et al., 2022; Schmidt et al., 2020; Schmidt & Huang, 2021; Tang et al., 2022; Vurorio et al., 2017). One of the misconceptions present throughout the LXD evaluation literature is the validity of small sample sizes to evaluate learning technologies (Carey & Stefaniak, 2018; Huang, 2023; Schmidt et al., 2023; Tawfik et al., 2022; Tawfik et al., 2024). This misconception stems from Nielsen’s (1993) early work in usability, which suggests using small sample sizes for evaluations. It is thus imperative to challenge whether five users is enough to identify the 80% of usability deviations in a learning technology. Identifying the number of testers needed for the usability evaluation of learning technologies will aid in speeding up iterative cycles and protect limited resources such as time and participants (Tawfik et al, 2024).

This non-experimental causal-comparative study aimed to explore the relationship between expertise/experience on usability deviation identification, severity rating, and classification in a learning technology. This exploration occurred at one location in the Mid-South region of the United States that consisted of a business corporation. Ten participants (five
novices and five experts) were recruited to participate in this study. The independent variable of experience divided the participants into novice and expert groups. Novices were classified as those with five years or less of experience and experts were classified as those with 10 or more years of experience.

In the chapter that follows, the methodology is discussed in-depth regarding the following research questions:

**Research Question 1.** What is the difference, if any, between novices and experts on the percentage of learning experience design deviations (problems) identified?

**Research Question 2.** What is the difference, if any, between novices and experts on the average severity ranking of learning expert design deviations?

**Research Question 3.** What is the difference, if any, between novices and experts on the number of interaction within the learning environment problems vs. interaction within the learning space problems they identify?

**Participants/ Learner Characteristics**

The population for this non-experimental causal-comparative study included one sample population: business employees from a large corporation whose headquarters are in the Mid South region. A non-probability convenience sample was used for this study. A non-probability convenience sample is appropriate for this study because participants were recruited voluntarily from a sample readily accessible to the researchers (Gall et al., 2015). In terms of justification of the sample size, Nielsen's (1993) assumption argues that five users are needed to identify 80% of
usability problems within a platform; however, this assumption has not been tested within a learning technology that includes learning tasks and usability features.

**Setting**

This experiment was conducted at a business corporation with headquarters located in the Mid South region of the United States. This corporation included 7,500 associates across 417 locations; however, for this study, 31 associates working for the talent and training department were recruited. Regarding ethnicity, 58.06% of associates are White, and 41.93% are Black or African American. Concerning gender, 77.42% of associates are female, and 33.34% are male. Most potential participants are White females, thus creating a skewed homogenous sample that may threaten external validity. However, the similarity of the participants across demographic variables may limit the influence of possible extraneous variables.

**Instrumentation/ Data Collection Methods**

**Recruitment and Demographics Data Collection**

Before the study, a recruitment e-mail was sent to training associates (Appendix A). Those agreeing to participate in the study received a second e-mail containing a link to the informed consent and demographics survey (Appendix C & Appendix D). The demographics survey was given using Google Forms to gather information regarding age, gender, race, education level, and usability experience. The UX experience demographic were used for classifying participants as novices or experts. Furthermore, this information was used to compare how similar the groups are across demographic variables. Groups similarity across demographic variables will aid in limiting the potential influence of extraneous variables (Schenker & Rumrill, 2004). Even though Nielsen (1993) suggests a sample size of five users is enough to identify
80% of usability problems, more recent literature has suggested a sample size of 10 +/- 2 is more effective for identifying 80% percent of usability problems (Bagheri et al., 2023; Borsci et al., 2013). However, this study aimed to directly challenge this assumption in relation to learning technologies. Therefore, a sample size of five novices and five experts was chosen to align with Nielsen (1993)’s methodology. The final sample of ten participants received a third e-mail asking the participants to schedule a time to meet with the primary investigator to complete the example. The participants recorded their results in Microsoft Forms.

**Description of Learning Environment (Gooru)**

Gooru is an open-source online learning platform that collaborates with research and technology experts to develop learning technologies to promote learning outcomes and aid in lifelong learning globally. Gooru’s goal is to provide equal access to educational opportunities worldwide. Gooru provides classes on various topics, from 8th Grade Math to Skills for Employability. The technology uses the learners’ real-time data, such as learner context and performance, to provide recommendations for additional learning.

The specific Gooru module employed for this study — *Skills for Employability* — focuses on soft skills that are important for being a successful worker in a business environment. The module addressed nine overarching themes: teamwork, leadership, communication, professionalism, customer solutions, change management, applying technology, sustainable practices, and safety and welfare. Each theme contained multiple lessons for the learner to complete. These lessons contained videos, readings, and quizzes to check for understanding (Figure 1).
Participants were asked to complete seven tasks from the protocol (Appendix G) using the dashboard of the learning module (Figure 2). The dashboard contained four main sections: competency gains, learning journey, performance overview, and pending activities. Learners are able to see in real-time their performance, progress in their learning journey, and find opportunities to improve their learning. The tasks outlined in the protocol had the participants explore the interface and assess the usability and learner experience of the learning technology.
Prior to the full study, two LXD expert researchers were recruited to complete a pilot study of the Gooru *Skills for Employability* learning module. During a heuristic evaluation, participants for the pilot study were given principles or construct guidelines and then asked to evaluate an interface per those guidelines (see description below in Pilot Study; Nielsen & Molich, 1990). The participants identified LXD deviations and then rated the severity of those LXD using Nielsen (1992)’s Severity Rating scales (Nielsen & Molich, 1990; Khajoui et al., 2018; Bagheri et al., 2023). As it relates to this study, the LXD expert researchers were provided with the protocol located in Appendix F. Initially, usability expert researchers were instructed to
freely explore the interface and note any LXD problems they encounter. Moreover, the researchers were asked to describe the problem and rate its severity on a 1-4 scale (Nielsen, 1992). This initial exploration acted as an informal heuristic evaluation for the learning module. Next, the expert researchers completed the seven tasks that acted as a cognitive walkthrough of the learning module (Appendix G). During this portion of the pilot study, the experts noted any LXD deviations they encounter while completing the outlined tasks. Moreover, these experts rated the severity of each issue identified. The primary investigator recorded the results of the pilot LXD study in Google Sheets.

**Instrumentation (Full study): Checklist Based on Expert Heuristic Evaluation**

To answer the proposed research questions, participants used a checklist developed from an expert heuristic evaluation derived from the pilot study to evaluate the Gooru module (Bagheri et al., 2023). This checklist was comprised of 32 LXD deviations identified in the pilot study, along with 30 “dummy” items. Using Microsoft Forms, participants were instructed to check off the LXD deviations they believe are present as they completed the study protocol (Appendix G). Moreover, participants were to provide a severity rating for each LXD deviation (Bagheri et al., 2023; Khajouei et al., 2018).

**Data Collection/ Procedures**

**IRB Approval**

The researcher gathered the necessary information to complete IRB forms and submitted the study for IRB approval. After IRB review, the IRB committee determined this study does not meet the Office of Human Subjects Research Protections definition of human subjects research.
Therefore, the IRB committee has decided this study does not require IRB approval or review. A letter of this determination is located in Appendix B.

**Pilot Study**

Two LXD expert researchers with experience with usability evaluations and learning technologies were recruited to perform a pilot study on the Gooru module. The lead researcher for this study contacted potential participants that meet the inclusion criteria via e-mail. Those who agreed received a second e-mail with instructions to schedule a time to meet synchronously with the primary investigator to complete the study. These two LXD researchers were selected for not only their experience with LXD evaluations but also their extensive experience with the Tawfik et al.’s (2022) framework. Moreover, one of these LXD experts was essential in the development of Tawfik et al’s (2022) LXD framework.

Expert researchers conducted a preliminary heuristic evaluation of the selected learning module using a live focus group method with the primary investigator present. The live focus group allowed the researchers to discuss their findings and solve any discrepancies. In a collaborative setting, these researchers were asked to identify LXD deviations within the learning module and provide a severity rating for each problem. According to recent literature, collaborative heuristic evaluation with evaluators from different backgrounds find more UX issues (Celik et al., 2023). Following the completion of this study, the primary investigator classified those LXD deviations per Tawfik et al.’s (2022) two overarching Learning Experience Design constructs (e.g. *interaction with the learning environment* or *interaction with the learning space*) (Appendix G). One of the usability experts then independently classified these interactions and this was compared to the primary investigators’ classifications (in-depth
discussion in Chapter 4). Primary investigator compiled the LXD experts findings in a Google Sheets document and included the following information: problem name, description, location, and a severity rating. The primary investigator used these results to create a checklist for the full study participants.

**Recruitment of Participants for Primary Study**

Participants were recruited via e-mail. An initial recruitment e-mail was sent to associates working in the talent department at a business corporate (Appendix A). The e-mail contained a general overview of the research study and asked those who are interested to respond to the e-mail. The interested parties received a second e-mail containing a link to the demographic survey (Appendix D). The demographics survey collected data on race, gender, age, education level, and usability experience.

The results of this demographics survey were used to classify participants as novices or experts. The first five participants to meet the requirements for a novice and the first five participants to meet the requirements for an expert were selected to participate in the study. The 10 total participants received a third e-mail containing instructions to schedule a experiment time slot with the primary investigator.

**Data Collection for Primary Study**

Each participant completed the heuristic checklist of the Gooru module independently. During these evaluations, ten participants (Novices = 5; Experts = 5) completed a checklist via Microsoft Forms; participants were asked to

1. select any usability deviations they found within the learning module
2. rate the severity of the problem
Participants were asked for identifying information; however, this information was only used to pair demographic survey results with the study results. After pairing, the data were assigned a random participant number to protect the identity of each participant. All identifying information was deleted.

Deviation classification refers to which of the given heuristics the participants believe the LXD deviation violated. As such, the first requirement for the evaluation asked participants to examine the learning module using the provided checklist to identify LXD problems and submit their checklist via Microsoft Forms. On the form, participants were asked to provide the severity rating information for the items they checked (Bagheri et al., 2023; Khajoui et al., 2018). The second requirement asked participants to use Nielsen's (1992) severity rating system to rate each LXD deviation identified. Descriptive statistics were used to assess the differences between novices and experts on severity ratings. These evaluations and descriptive statistics were used to address research questions one, two, and three.

**Analysis**

For research question one, descriptive statistics were used to explore if novices and experts differed on the dependent variable of LXD deviation identification. Descriptive statistics were used to examine these means. Descriptive statistics were used to examine these means because the sample size does not meet the requirements for an independent samples t-test (Urdan, 2016). Moreover, in order to conduct an independent samples t-test, the independent variables must be nominal or categorical, and the dependent variable must be on an interval or continuous scale (Urdan, 2016). Due to the nature of this data, tests of significant cannot be performed.
For research question two, initial descriptive statistics were used to calculate the
differences in means among novices and experts on the dependent variable of severity ratings.

**Table 3**

*Research Questions Alignment Statistics with Instrumentation*

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Alignment to Instrumentation</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1: What is the difference, if any, between novices and experts on the <strong>percentage of LXD deviations</strong> (problems) identified?</td>
<td><strong>Heuristic Checklist Part 1:</strong> Problem identification. Participants will use a checklist to identify LXD deviations present within the learning module.</td>
<td><strong>Descriptive Statistics</strong></td>
</tr>
<tr>
<td>RQ2: What is the difference, if any, between novices and experts on the <strong>average severity ranking of LXD deviations</strong>?</td>
<td><strong>Nielsen (1992)’s severity rating.</strong> Participants will provide a rating for each LXD deviations identified.</td>
<td><strong>Descriptive Statistics</strong></td>
</tr>
<tr>
<td>RQ3: What is the difference, if any, between novices and experts on the number of</td>
<td><strong>Primary investigator and LXD expert will classify each</strong></td>
<td><strong>Descriptive Statistics</strong></td>
</tr>
</tbody>
</table>
interaction within the learning environment vs. interaction within the learning space problems they identify?

Descriptive statistics were used to assess the data set to address research question three. The LXD problems the participant identified were coded by the PI and one LXD expert from the pilot study. These codes were used to assess the number of total interaction within the learning environment and interaction within the learning space problems identified by novices and experts.
There are several limitations present in this study. The non-experimental, causal-comparative, descriptive design of this study places multiple limitations on the results. First, the limitations to internal validity and external validity will be discussed. A causal-comparative study examines the differences in pre-existing groups; therefore, internal validity cannot be ensured due to the lack of manipulation of the independent variable (Schenker & Rumrill, 2004). The lack of manipulation of the independent variable means that causation cannot be inferred from the results. The results may suggest a relationship exists between two variables; however,
the researchers cannot discount the possibility of potential extraneous influencing the results (Schenker & Rumrill, 2004).

Regarding external validity, using a purposeful, convenience sample threatens external validity. The lack of random selection of participants may cause the selected sample not to be representative of the overall population (Schenker & Rumrill, 2004). The use of a purposeful, convenience sample was chosen because the literature suggests that heuristic evaluations are complex, and these types of samples are needed to address this complexity (Lu et al., 2022). To reduce the potential for extraneous variables, the researcher will match participant demographics to the general population at the study locations.

**Ethical Considerations**

The researcher’s ethical goal is to ensure private transparency throughout the study while protecting the participants’ welfare and privacy. Before the study, participants received informed consent, which provided a brief overview of the study, the purpose of the research, procedures, potential risks, and confidentiality. Moreover, the participants were informed that their participation in this study is voluntary, and they may withdraw their participation at any time without penalty. Participants saw this information twice. The first time was in the informed consent, and the second time was before starting the study.

Concerning data storage, all data will be stored with participant numbers to protect identities. The researcher matched the participants’ demographic and evaluation data and then deleted any identifying information before storing the data. Data was protected using passwords that will be changed periodically through the completion of the study. Only the primary investigator had access to the data. Data was stored following all IRB requirements.
Bias/Subjectivities

Due to personal experience, the driving motivation for this research is to reduce the inefficient evaluation practices that many instructional designers actively working in the field experience when trying to evaluate a technology. Current practices are time-consuming, inefficient, and deplete the limited participant/tester pool. These evaluation practices create difficulty for instructional designers working in the business industry and academia. The evaluation practices hinder the rapid development that many stakeholders are now asking of designers. Stakeholders want the design and evaluation process to be quicker and more efficient so that their consumers receive the product or system sooner. However, current practices do not allow for this rapid development.

The researcher acknowledges as an instructional designer working in the business industry that they are interested in the current study. Study results could improve industry practices and change the expectations placed on instructional designers like the researcher. Additionally, the researcher acknowledges a modern alignment with the Constructivist philosophical paradigm. Therefore, the researcher aligns with the belief that people’s past experiences influence how they construct knowledge (Jonasson, 1991). The researcher is aware that these philosophical beliefs could influence the researcher’s perspective of this study. The researcher recognizes these biases and their potential implications and will evaluate these biases throughout the study to limit their potential influence.
CHAPTER FOUR: RESULTS

Introduction

The study was performed to explore the validity of the application of Nielsen’s Five User Assumption when evaluating learning technology. Although not originally intended for application to learning technology, Nielsen (1993)’s Five User Assumption is present throughout the literature to justify the use of small sample sizes when evaluating learning technologies (Carey & Stefaniak, 2018; Huang, 2023; Schmidt et al., 2023; Tawfik et al., 2022; Tawfik et al., 2024). The study aimed to extend this approach beyond traditional technological usability and test this methodological assumption when applied to the realm of learning technologies. More specifically, this study explored whether usability or learning experience design experience impacted evaluation results by using novices and experts within the study. The purpose of the results chapter is to present the data analysis or evaluation. This chapter contains the results of that exploration.

Results

Participant Demographics

Before completing the study, participants completed a demographics survey and informed consent. Participants were given a nine-question survey on various demographic items such as age, gender, and education. 12 participants completed the survey; however, the data from two participants were excluded from the study because these participants did not meet the novice or expert criteria set for this study. Additionally, survey questions regarding industry, job title, and country of residence are excluded from the table below because all the participants worked in the same industry (business) and country (United States).
Upon completing the demographics survey, the participants were split into novice and expert groups based on prior LXD experience. Five participants met the novice inclusion criteria of five years or less of LXD experience, and five met the expert inclusion criteria of ten years or more of LXD experience (Hmelo et al., 2007). This expert/novice criteria was selected to align with previous research, such as Hmelo et al. (2007), that classified experts as those with 10 or more years of experience and an advanced degree in the subject matter. Specifically, those classified as experts for this study had 10 or more years of experience conducting usability evaluations for learning technologies or an advanced degree (master’s or doctorate) in IDT. The reasoning for this modification is to include those who meet the experience criteria but do not have an advanced degree in IDT. Although these potential experts do not have an advanced degree, the literature suggests that their 10 plus years of experience allow their inclusion into this category. Concerning novices, though Hmelo et al. (2007) does not provide a specific range for novices, the researchers do indicate that novices should have limited or no experience; therefore, the criteria of less than five years was selected for novices. Regarding the age of these ten participants, two participants were in the 25-34 age range, three participants were in the 45-56 age range, and five participants were in the over 55 age range (Table 4). Concerning gender, eight participants identified as female, and two participants identified as male.

In terms of education, one participant had a high school degree, one had an associate’s degree, four had a bachelor’s degree, and four had a master’s degree. The fields of these degrees varied across participants. These fields included Business Administration, Education, Finance, Human Resources Management, and Sports Administration. Four participants did not provide a degree field, or the field did not apply to the participant. Moreover, these participants have a
variety of job titles, including Instructional Designer, various HR positions, Talent Acquisition, and Talent Development. Concerning years spent in these positions, two participants have been in their position for 1-3 years. One participant has 6-9 years in their position, and seven participants have over 10 years in their position. Table 4 provides a numerical representation of this data.

**Table 4**

*Participants’ Demographic Survey Data*

<table>
<thead>
<tr>
<th></th>
<th>Novice</th>
<th>Expert</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>25-34</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>45-54</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Over 55</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td><strong>Highest Degree</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Associate’s</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Bachelor’s</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Master’s</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><strong>Years in the Field</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-3</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>4-5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

52
In line with other studies (Bagheri et al., 2023), the checklist used for this study was developed from the pilot study results. Two LXD experts with extensive experience with Tawfik’s et al. (2022) LXD framework collaboratively completed seven tasks within the learning technology (see Appendix F). Examples of these tasks included navigating from the dashboard to the Professional Behavior lesson and exploring the lesson content to find where to record notes that learners might want to write to support their learning. While completing the tasks, the LXD experts identified task-specific deviations within the learning technology and rated the severity of each problem on a scale of 1-4 using Nielsen’s (1992) Severity Rating scale.

For an hour and a half, the two LXD experts discussed the problem and rating among themselves until they reached an agreement. After the pilot experts completed their tasks, the primary investigator used their results to develop a checklist for the full experiment, which included 32 LXD deviations from experts and an additional 30 dummy items. The primary
investigator justified the use of using two LXD experts in the pilot study through consulting recent literature that provides support that collaborative heuristic evaluations with experts provides a higher problem identification rate than individual ratings (Celik et al., 2023). Whereas Nielsen’s (1993) Five User Assumption criteria pertains to five people individually completing tasks in an isolated setting. The lack of collaboration between the evaluators for Nielsen’s (1993) Five User Assumption could explain the need for more users; additionally, Nielsen (1993) is not specifically referring to experts in his assumption. Rather, Nielsen (1993) poses if a researcher selects five random participants for a UX evaluation, these individuals will identify 80% of the problems within the technology.

**Research Question 1: What is the difference, if any, between novices and experts on the percentage of LXD deviations (problems) identified?**

In order to test the Nielsen (1992) assumption of N=5 will find 80% of users, the first question sought to examine if a difference between novices and experts was present regarding the percentage of usability problems identified within the experiment. Participants were asked to complete seven tasks within the learning technology and were provided a checklist that contained LXD problems identified prior to the experiment by LXD experts. As noted above, the LXD experts used in the pilot identified 32 total LXD deviations present within the learning technology. Additionally, this checklist contained 30 dummy LXD deviations that was generated by the primary researcher.

After completing each task, participants checked off from the checklist all the LXD problems they agreed were present within the learning technology. Descriptive data revealed that novices identified 27 LXD deviations present within the learning technology or 84.38% of the
LXD deviations (Table 5). Alternatively, experts identified 22 LXD deviations within the technology or 68.75% of the LXD deviations (Table 6). The novice identification percentage exceeds the error identification percentage (80%) Nielsen (1993) proposed in his Five User Assumption; however, the expert identification percentage does not meet or exceed this percentage.

**Dummy Items Severity: Overall for Correct/Incorrect**

Concerning the dummy items, the primary investigator generated 30 LXD dummy items to add to the checklist. These dummy items were included to reduce the likelihood of participants guessing the answers correctly without having the needed UX knowledge (Shin et al., 2019). Participants were provided with 32 LXD deviations and the 30 LXD dummy items intermixed on the checklist, which was divided by task. Using descriptive statistics, data revealed that novices identified 6 dummy (incorrect) items as correct, or 20% of the LXD dummy items; whereas experts identified 18 dummy (incorrect) items as correct or 60% of the 30 LXD dummy items (Table 5).

**Table 5**

<table>
<thead>
<tr>
<th>Novice v. Expert Percentage of LXD Deviations Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice Correct</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Task 1</td>
</tr>
<tr>
<td>Task 2</td>
</tr>
<tr>
<td>Task 3</td>
</tr>
</tbody>
</table>
Research Question 2: What is the difference, if any, between novices and experts on the average severity ranking of LXD deviations?

To better understand how Nielsen’s (1992) user assumption might apply to an LXD context, the second research question examined is descriptive differences between novices and experts on severity ratings as users evaluated a learning technology. In addition to selecting LXD deviations they agreed with, participants were asked to rate the severity of each LXD problem they identified. Participants were provided with Nielsen’s (1992) Severity Rating Scale (Appendix E) to rate the severity of each identified problem. Descriptive statistics were used to analyze these results. These results indicated that novices and experts were similar across severity ratings (Table 6)

Table 6

<table>
<thead>
<tr>
<th></th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>N4</th>
<th>N5</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 4</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 5</td>
<td>100%</td>
<td></td>
<td>33.33%</td>
<td>100%</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 6</td>
<td>100%</td>
<td>0%</td>
<td></td>
<td>100%</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 7</td>
<td>80%</td>
<td>0%</td>
<td></td>
<td>100%</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>84.38%</td>
<td>20%</td>
<td>68.75%</td>
<td></td>
<td></td>
<td>60%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
</tr>
</tbody>
</table>

(Note: All data is rounded to two significant digits).
Research Question 3: What is the difference, if any, between novices and experts on the number of interaction within the learning environment vs. interaction within the learning space problems they identify?

The final research question sought to examine if distinctions emerged between novices and experiences regarding the number of interaction within the learning environment vs. interaction within the learning space problems they identified. Prior to the full experiment, the primary investigator used Tawfik et al.’s (2022) framework to classify the LXD deviations articulated by the pilot LXD experts during their evaluation of the learning technology as either interaction within the learning environment or interaction within the learning space. After the primary investigator classified these LXD deviations, one pilot LXD expert reviewed the
classifications for accuracy. Cohen’s κ was executed to determine inter-rater reliability on agreement between the two raters on whether the 32 LXD problems were classified as interaction within the learning environment or interaction within the learning space. There was very good agreement between the two raters on classifications, κ = 1.00, p < .001 (Table 8). The primary investigator and the LXD expert classified 17 UX deviations as interaction within the learning environment and 15 UX deviations as interaction within the learning space.

Table 8

<table>
<thead>
<tr>
<th>Measurement of Agreement</th>
<th>Value</th>
<th>Asymptotic Standard Error</th>
<th>Approximate T</th>
<th>Approximate Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td>1.000</td>
<td>.000</td>
<td>5.657</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Analysis of the results revealed that novices identified 14 (82.35%) interaction within the learning environment UX deviations from the original 17 that were identified by the LXD experts from the pilot study. Alternatively, novices identified 13 (86.67%) of the original 15 deviations derived in the pilot as interaction within the learning space LXD deviations. Experts identified 11 (64.71%) interaction within the learning environment LXD deviations and 11 (73.34%) interaction with the learning space LXD deviations.

Dummy Items: ILE v. ILS Classification

As mentioned in a prior section, the primary investigator generated 30 dummy items to add to the checklist before the full study. The primary investigator and one LXD pilot expert independently coded these dummy items as interaction within the learning environment or interaction within the learning space in accordance with the Learning Experience Design.
framework proposed by Tawfik et al. (2022). The two raters coded 5 dummy items as *interaction within the learning environment* and 25 dummy items as *interaction within the learning space*. Cohen’s κ was run to assess the agreement between the two raters on whether the 30 dummy items fell into the category of *interaction within the learning environment* or *interaction within the learning space*. There was a very good agreement between the two raters, κ = 1.00, p < .001 (Table 9).

**Table 9**

*Inter-rater Reliability for PI and LXD Expert for Dummy Items*

<table>
<thead>
<tr>
<th>Measurement of Agreement</th>
<th>Value</th>
<th>Asymptotic Standard Error</th>
<th>Approximate T&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Approximate Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td>1.000</td>
<td>.000</td>
<td>5.477</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 10**

*Novice v. Expert Interaction within the Learning Environment Identification*

<table>
<thead>
<tr>
<th></th>
<th>Novice Correct (N= 17)</th>
<th>Novice Incorrect (N= 5)</th>
<th>Expert Correct (N= 17)</th>
<th>Expert Incorrect (N= 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>100% (N= 5)</td>
<td>0% (N= 0)</td>
<td>60% (N= 3)</td>
<td>0% (N= 0)</td>
</tr>
<tr>
<td>Task 2</td>
<td>45.45% (N= 5)</td>
<td>0% (N= 0)</td>
<td>50% (N= 4)</td>
<td>0% (N= 0)</td>
</tr>
<tr>
<td>Task 3</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Task 4</td>
<td>100% (N= 2)</td>
<td>NA</td>
<td>100% (N= 2)</td>
<td>NA</td>
</tr>
<tr>
<td>Task 5</td>
<td>100% (N= 1)</td>
<td>33.33% (N= 1)</td>
<td>100% (N= 1)</td>
<td>33.33% (N= 1)</td>
</tr>
<tr>
<td>Task 6</td>
<td>NA</td>
<td>0% (N= 0)</td>
<td>NA</td>
<td>33.33% (N= 1)</td>
</tr>
<tr>
<td>Task 7</td>
<td>100% (N= 1)</td>
<td>NA</td>
<td>100% (N= 1)</td>
<td>NA</td>
</tr>
<tr>
<td>Category</td>
<td>Percentage 1</td>
<td>(N=) 1</td>
<td>Percentage 2</td>
<td>(N=) 2</td>
</tr>
<tr>
<td>----------</td>
<td>--------------</td>
<td>--------</td>
<td>--------------</td>
<td>--------</td>
</tr>
<tr>
<td>Overall</td>
<td>82.35%</td>
<td>14</td>
<td>20%</td>
<td>1</td>
</tr>
</tbody>
</table>
### Table 11

**Novice v. Expert Interaction within the Learning Space Identification**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>100% (N=2)</td>
<td>40% (N=2)</td>
<td>50% (N=1)</td>
<td>60% (N=3)</td>
</tr>
<tr>
<td>Task 2</td>
<td>66.67% (N=2)</td>
<td>0% (N=0)</td>
<td>0% (N=0)</td>
<td>25% (N=2)</td>
</tr>
<tr>
<td>Task 3</td>
<td>100% (N=2)</td>
<td>50% (N=1)</td>
<td>100% (N=2)</td>
<td>100% (N=2)</td>
</tr>
<tr>
<td>Task 4</td>
<td>NA</td>
<td>100% (N=2)</td>
<td>NA</td>
<td>100% (N=2)</td>
</tr>
<tr>
<td>Task 5</td>
<td>100% (N=2)</td>
<td>0% (N=0)</td>
<td>100% (N=2)</td>
<td>100% (N=2)</td>
</tr>
<tr>
<td>Task 6</td>
<td>100% (N=2)</td>
<td>0% (N=0)</td>
<td>100% (N=2)</td>
<td>100% (N=2)</td>
</tr>
<tr>
<td>Task 7</td>
<td>75% (N=3)</td>
<td>0% (N=0)</td>
<td>100% (N=4)</td>
<td>75% (N=3)</td>
</tr>
<tr>
<td>Overall</td>
<td><strong>86.67% (N=13)</strong></td>
<td><strong>20% (N=5)</strong></td>
<td><strong>73.33% (N=11)</strong></td>
<td><strong>68% (N=16)</strong></td>
</tr>
</tbody>
</table>

**Average**

---

**Summary**

This chapter provided data analysis to explore the three proposed research questions challenging the Nielsen’s Five User Assumption methodology. Specifically, the study examined if five users are enough to reach or exceed the 80% problem identification rate Nielsen proposes (1993). Moreover, this study sought to examine if differences were present between novices and experts on LXD deviation identification rate (RQ1), severity ranking (RQ2), and classification (RQ3). Results reveal that novices identified 84.38% of LXD deviations, and experts identified 68.75% of LXD deviations. Moreover, analysis of the dummy items revealed that, while novices
identified 20% (6) of 30 incorrect dummy LXD problems as correct, experts identified 60% (18) of the 30 incorrect dummy LXD problems as correct. The severity rating data analysis revealed no significant differences between novices and experts.

Regarding UX classification, the analysis revealed novices identified 82.35% of interaction within the learning environment problems and 86.67% of interaction within the learning space problems. Furthermore, analysis of the dummy LXD items revealed that novices identified 20% (1) of the five incorrect interaction within the learning environment items as correct and 20% (5) of the 25 incorrect interaction within the learning space items as correct. Alternatively, experts identified 64.71% of interaction within the learning environment problems and 73.33% of interaction within the learning space problems. Regarding the dummy items, experts identified 40% (2) of the five incorrect interaction within the learning environment items as correct and 68% (16) of the 25 incorrect interaction within the learning space items as correct. In Chapter 5, the implications of these analyses are discussed.
CHAPTER FIVE: DISCUSSION AND CONCLUSIONS

Introduction

A prominent methodological assumption within the LXD evaluation literature is the application and validity of small sample sizes to evaluate learning technologies (Carey & Stefaniak, 2018; Huang, 2023; Schmidt et al., 2023; Tawfik et al., 2022). The commonality of small sample sizes to evaluate technology stems from the robust connection between the field of UX and the field of LXD. This assumption derives from Nielsen’s (1993) usability research, which suggests that N=5 will find 80% of usability issues. Therefore, it is imperative to empirically test the validity of this assumption when applied to a learning technology.

To expand on current LXD research regarding the application of small sample sizes to evaluate learning technologies, this study sought to provide an empirical analysis of Nielsen’s (1993) Five User Assumption when applied to learning technologies. Furthermore, this study aimed to expand upon current research by examining the impact of experience on conducting LXD evaluations (RQ1). The differences between novices and experts were explored not only on LXD deviation identification rate but also on LXD problem severity ratings (RQ2) and LXD problem type (classification) identification rates (RQ3).

Summary of Findings

Data Collection Procedures

Data collection for this experiment was conducted at a mid-sized business corporation located in the Midsouth region of the United States. Participation in this study was voluntary for business associates in Talent Development and Acquisition (N= 31, Male= 7, Female= 24). A
recruitment e-mail was sent to all 31 associates within the department. 12 associates responded to the e-mail, and 10 met the inclusion criteria for participation in the study (N=10, Male=2, Female=8). Specifically, five participants met the inclusion criteria for novices, and five met the criteria for experts.

This specific location and department were selected for a variety of reasons. The Talent Development and Acquisition department handles company training, including quality assurance measures and vetting learning content. This department contained a multitude of associates who met the inclusion criteria for both novices and experts. Additionally, this population was easily accessible to the researcher due to their employment at the department.

Research Question 1 Findings: What is the difference, if any, between novices and experts on the percentage of LXD deviations (problems) identified?

The first research question focused on the percentage of learning experience design deviations novices and experts identified while completing seven tasks within a learning technology (Appendix E). Tasks included navigating to the Learning Journey dashboard, starting a lesson, and completing a quiz for the Professional Behavior lesson. Although these may seem like simple tasks, they include a series of nuanced interactions and subtasks that are part of the learner's HCI. For instance, in order to complete the task of starting a lesson, participants use existing mental models to generate contextual cues to guide their LXD interactions as they use learning technologies. Moreover, they must orient themselves to the interface layout and comprehend these interface labels to navigate to the location of the lesson. Once participants located the lesson, they had to locate the play button represented by a sideways triangle and then click the button, which was an important part of learning because this is how users access the
multimedia learning material essential for future interactions. These examples show how a singular task can contain multiple interactions that participants must successfully navigate to complete the assigned task.

The participants were given a checklist via Microsoft Forms that contained 32 LXD deviations generated by the pilot LXD pilots during their evaluation of the learning technology and 30 LXD dummy items generated after the initial pilot study. Participants were asked to refer to the checklist after completing each task outlined in the protocol (Appendix E) and to check off any LXD deviations that they agreed were present during each task. Based on each participant's checklist, this data was coded as found or not found. Following this coding, the deviation identifications for all five participants in each group were combined. The novice and expert group data were compared against the LXD deviation identification data generated by the two LXD experts during their initial evaluation of the learning technology. Novices found 27 of the 32 (84.38%) LXD deviations identified by the LXD experts during the pilot study. Experts found 22 of the 32 (68.75%) identified LXD deviations. These totals were converted into percentages by dividing each group's found LXD deviations by the number of LXD deviations identified by the LXD experts in the pilot study.

Due to the research question's attempt to address the N=5 assumption, significant differences between the groups could not be assessed because the testing did not meet the assumptions of statistically significant comparison tests (Kang, 2021). However, this research question aimed to explore whether five users are enough to identify 80% of LXD deviations within the learning technology. Novices exceeded this metric; however, experts did not meet or exceed this metric. These results suggest that LXD experience/expertise may impact the number
of LXD deviations a user identifies within a learning technology. Differences in experience may cause participants to identify different problems within the learning technology because some participants have knowledge gaps that need substantial scaffolding. In contrast, other participants can apply previous experiences to the novel technology and be successful. Experience may thus impact how a participant experiences and perceives challenges as they engage with LXD interactions within the learning technology. One might conclude thus these different levels of experience, a novice or expert designation, impact aspects of the learning technology evaluation.

Furthermore, the number of incorrect dummy items each group found was assessed. 30 total dummy items were included on the checklist provided to participants. For the group analysis, if any group member checked off a distinct dummy item, that item was calculated into the group total. For example, for Task 7, expert participants 4 and 5 identified the dummy item “The interface does not show time spent on the quiz” as correct when it was incorrect. Though two participants identified this incorrect dummy item as correct, the incorrect LXD deviation total would increase by one. Descriptive statistics revealed that novices identified 20% of the 30 incorrect LXD dummy items as correct, and experts identified 60% of the 30 incorrect LXD dummy items as correct. Again, the small sample size does not allow for significant difference comparison tests (Kang, 2021). However, these results potentially suggest that experience may impact the number of false positives and incorrect LXD deviations that participants identify as correct. Previous usability research suggests that experts focus on efficiency during a technology evaluation (Magbool & Herold, 2024). However, constructivism - which is based on knowledge construction - is not always linear, and often iteration is a key aspect of the learning process (Jonassen, 1997). In line with usability research, it may be that experts misclassified LXD
deviations because they were more focused on efficiency. At the same time, novices were able to accurately select key LXD interactions that were key to their learning process.

Research Question 2 Findings: What is the difference, if any, between novices and experts on the average severity ranking of LXD deviations?

The second question focused on whether differences between novices and experts existed when comparing severity ratings for LXD deviations that were identified within the learning technology. During the study, participants were asked to provide a severity rating for each problem they identified. These ratings were on a scale from 1-4 in accordance with Nieslen’s (1992) Severity Rating scale; problems that participants did not identify within the learning technology were assigned a value of zero during data analysis. These scores were then averaged across participants to provide an average severity score for each participant. Descriptive statistics assessed novices and experts on severity rating. The results suggested that novices and experts are similar across severity ratings. This suggests that although novices and experts differ on the percentage of LXD deviations they identify, they do not differ when rating the severity of these problems. If and when they can accurately identify them, the severity rating is not as problematic to the learning process. These results align with prior research that indicated no significant differences between novices and experts on severity ratings (Othman et al., 2022; Nasir et al., 2022).

Research Question 3 Findings: What is the difference, if any, between novices and experts on the number of interaction within the learning environment problems vs. interaction within the learning space problems they identify?
While the initial Nielsen (1992) assumption provided a holistic rating of usability errors, research question three aimed to examine whether novices and experts differ on the type of LXD deviations they identify. Before the present study, the primary investigator and one pilot LXD expert independently coded the classification of each LXD deviation that the pilot experts identified while evaluating the interface. All the LXD deviations identified in the pilot study were classified as *interaction within the learning environment* (N=17) or *interaction within the learning space* (N=15). Inter-rater reliability was calculated using Cohen’s Kappa. Results indicated a high level of agreement between the two raters on LXD deviation (κ = 1.00, p <.001).

Following the completion of data collection, the researcher used the classification coding to assess the number of *interaction within the learning environment* LXD deviations and *interaction within the learning space* LXD deviations that participants identified. These counts were combined across the participants to create an overall percentage for each LXD classification. Results revealed that, of the LXD deviations novices identified, 14 (82.35%) were classified as *interaction within the learning environment*, and 13 (86.67%) were classified as *interaction within the learning space*. Experts identified 11 (64.71%) LXD deviations classified as interaction within the learning environment, and 11 (73.33%) as *interaction within the learning space*. Due to the nature of this data (count data) and the small sample size, significant differences cannot be assessed.

Additionally, descriptive statistics were conducted to assess the total number of LXD dummy items each group found. Before the full study, all 30 dummy items were independently classified as *interaction within the learning environment* (N= 5) or *interaction within the learning space* (N= 25) by the primary investigator and one LXD expert from the pilot study.
The classifications from the primary investigator and the LXD pilot expert were then used to compare the number of dummy items novices and experts identified. Descriptive statistics revealed that novices identified 1 of the 5 incorrect interaction within the learning environment dummy items as correct. Alternatively, experts identified 2 of the 5 incorrect interaction within the learning environment dummy items as correct. The number of interaction within the learning environment dummy items identified by the groups was divided by the total number of interaction within the learning environment dummy items (N=5) to produce a total percentage of incorrect classifications divided by novice and expert groups. The novice group identified 20% of the incorrect interaction within the learning environment as correct, and the expert group identified 40%.

Regarding the interaction within the learning space dummy items, the novice group identified 5 of the 25 incorrect interaction within the learning space dummy items as correct, and experts identified 16 of the 25 incorrect dummy items as correct. To calculate the total percentage of incorrect novice and expert assignations, the number of dummy items found was divided by the total number of dummy items present for the classification. In this case, novices identified 20% of the ILS dummy items as correct, and experts identified 68% of the dummy items as correct. That is, these results highlight that experts struggle with differentiating what constitutes an LXD deviation especially as it relates to interactions that are focused on pedagogical usability and instead focus on more technical interactions that are often focused on efficient completion of the task. Indeed, recent literature provides support for this explanation, suggesting that limited effort has been devoted to better define this LXD dimension, which
results in practitioners not having clear guidelines for identifying these types of problems (Lu et al., 2022).

**Discussion**

**Theoretical Implications**

Instructional designers are frequently asked to design various training materials, provide insights, and create learning technologies to aid learners in reaching their learning goals (DeVaughn, 2022). An essential piece of the design process is the evaluation stage; however, practitioners often cite being unable to efficiently and effectively perform evaluations due to lack of participants, high costs, and lack of support (DeVaughn & Stefaniak, 2021; Cennamo & Kalk, 2018). These practical restraints have led to the application of Nielsen’s (1993) Five User Assumption to become prevalent throughout the LXD and LIDT literature (Lu et al., 2022; Schmidt et al., 2020; Schmidt & Huang, 2021; Tang et al., 2022; Vurorio et al., 2017). However, this assumption was not initially created to be applied to learning technologies and lacks validation in this arena. The assumption was created to assess the usability of technologies developed 30 years ago that did not possess the complexities of present learning technology (McDonald & Yachar, 2020). Despite the lack of validation and mentioned fallacies, this assumption is still routinely used in LXD literature to justify the use of small sample sizes for learning technology evaluation (Carey & Stefaniak, 2018; Huang, 2023; Schmidt et al., 2023; Tawfik et al., 2022).

This implication is especially important with a recent literature review indicating a spike in the prevalence of LXD evaluations within the literature starting in 2019 (Estrada-Molina et al., 2022). One challenge related to LXD is that evaluations of learning technologies often focus on
the technical aspects and often miss the pedagogical aspects. Indeed, a recent systematic review of the literature found that 85% of the sampled literature only focused on the technological usability dimension of LXD (Lu et al., 2022). However, scholars throughout the literature have argued that technological usability is insufficient when evaluating the nuances and complexities of LXD interactions (McDonald & Yachar, 2020; Schmidt & Huang, 2022; Schmidt & Tawfik, 2022; Schmidt et al., 2020). Due to this insufficiency, scholars have increasingly argued that “that pedagogical and sociocultural considerations are critical aspects of designed learning technologies, and that these factors necessarily play a central role in the overall usability of a learning technology” (Lu et al., 2022, p. 1971). Although these usability dimensions are acknowledged as essential for evaluating learning technologies, the literature does not reflect this observation. For example, a recent systematic review found that only 13.4% of LXD evaluation studies reported pedagogical outcomes, and 1.4% reported sociocultural outcomes (Lu et al., 2022). These results suggest a disconnect between theory and practice within the LXD literature.

One possible explanation for this disconnect is the lack of methodologies, guidelines, and best practices to guide practitioners to evaluate along these dimensions (Lu et al., 2022). The present study builds upon prior theoretical literature by using Tawfik et al.’s (2020a) LXD framework to consider nuances for the HCI of learning technologies. This framework provides guidelines for evaluating both technological usability (i.e., interaction within the learning environment) and pedagogical usability (i.e., interaction within the learning space). Results from the current study suggest that novices and experts differ on the number of LXD deviations that they identify (RQ1). Moreover, these groups differed in both the interaction within the learning environment (i.e., technological usability) deviation category and interaction within the learning environment (i.e., technological usability) deviation category and interaction within the learning environment.
space (i.e., pedagogical usability) category (RQ2). Specifically, novices found 82.35% of the ILE deviations and 86.67% of the ILS deviations. Experts found 64.71% of ILE deviations and 73.33% of ILS deviations (RQ2). These results support the prior calls of scholars for practitioners, researchers, and designers in LIDT to acknowledge the importance of the other usability dimensions when evaluating learning technologies. Without explicit training and guidelines for practitioners along all of the LXD usability dimensions, the adequacy of future evaluation studies to assess learning technologies is called into question (Lu et al., 2022).

Methodological Implications

**Expert-Novice and LXD Evaluations**

Schmidt et al. (2023) suggest that “design flaws are often the result of the software development team failing to consider the user (or, in this case, the learner) sufficiently in the design process” (p. 1). Besides having theoretical implications, the results of the current study also have practical implications for instructional designers in the field. Evaluations are a key part of the iterative design process and provide insight into whether the learner recognizes the program’s design goals as usable (Cennamo & Kalk, 2018; DeVaughn & Stefaniak, 2021; Schmidt et al., 2020). However, high costs, lack of participants, and lack of support impede the evaluation process (DeVaughn & Stefaniak, 2021; Cennamo & Kalk, 2018). To limit these hindrances, many in the field use Nielsen’s (1993) Five User Assumption to justify using small sample sizes for evaluations. To date, the use of Nielsen’s (1993) Five User Assumption is still commonplace, and many researchers still use sample sizes of five for their LXD evaluations (Carey & Stefaniak, 2018; Huang, 2023; Schmidt et al., 2023; Tawfik et al., 2022; Tawfik et al.,
The study results question whether the field of LXD should retain this assumption during evaluations of learning technologies.

The results have important implications as they relate to LXD methodologies that practitioners and researchers might employ, especially the type of user considered during an LXD evaluation. Throughout the literature, LXD experts are often used as the sole evaluators to assess the usability of various technologies (Almeqbaali et al., 2022; Joshi et al., 2015; Arnhold et al., 2014). Moreover, studies have suggested that evaluators with expertise and extensive knowledge of usability problems can find and identify more usability issues when compared to novice evaluators (Othman et al., 2022). However, the results of the present study contradict previous research by suggesting that solely using usability or LXD experts for evaluations may not be the best practice (RQ1). Rather, experts in the study did not identify LXD problems that the novices were able to identify (RQ1). One possible explanation for these present results differing from previous research (Otham et al., 2022) is the phenomenon known as an expert blindspot. For instance, due to their experience, experts possess advanced mental models that contain automated scripts regarding lower-level thought processes. Experts’ advanced experiences regarding the subject matter impact how they approach new problems (Sternberg, 2014). Expert mental models are arranged around core features that guide their thinking and allow them to recognize patterns across the subject matter (Glaser & Chi, 1988; Sternberg, 2014). Through their experience, experts mentally cross-reference and link previously “unconnected knowledge structures into a model mental model” (Lehmann et al., 2019, p. 2). Whereas novices are still in the construction phase of knowledge and do not possess these interlinked mental models to the extent that experts do (Lehman et al., 2019). Moreover, experts
do not see knowledge as context-dependent which allows experts to recognize how features from a current problem align with a past problem. This connection allows experts to use previous problem-solving experience to solve a new problem (Sternberg, 2014).

In contrast, novices do not possess these well-developed advanced mental models and approach solving a new problem differently from experts. The views of novices are often context-dependent, which hinders their ability to align the presented problem with previously learned knowledge; novices often approach problems in isolation, looking for a particular heuristic to guide their thinking rather than drawing on previous experience (Lehman et al., 2019; Sternberg, 2014). Empirical studies have indicated that experts struggle to develop instructional material for novice learners; their perceptual limitation is referred to as the "expert blindspot" (Guo et al., 2020; Rau et al., 2019; Nathan et al., 2001). As experts advance their knowledge within their subject matter area, many problem-solving processes become automatic and intuitive (Sternberg, 2014). This developed intuitiveness can cause experts to overlook nuances essential for novice learning. For instance, experts may skip over lower-level procedural and contextual processes because these processes have become mentally automated. Because novices do not yet possess these automated processes and need more scaffolding to fill their knowledge gaps, they need help transitioning from lower-level processes to more advanced material. Practically, instructional designers should employ end-users to review instructional materials that were developed with subject matter expert (SME) guidance. The end-user review will allow instructors to identify and address obstacles and challenges that novice learners may face when using the instructional material. The results of the current study show that this includes the design and development aspect of ADDIE and evaluations.
More recent research suggests that experts do not recall the process development necessary for novices to expand their knowledge base which may impact the types of LXD interactions considered key to a successful learning technology (Guo et al., 2020; Rau et al., 2019). Experts have developed a mental model concerning common navigational processes and can apply successful experiences concerning past learning technology navigation to the present technology. Because novices have not developed these advanced mental networks, they are hyper-focused on the minute nuances of interface navigation critical for their learning and LXD interactions. Therefore, experts may subconsciously overlook the unique navigational nuances of the current learning technology which results in overlooking navigational deviations that are impactful to novice users. As illustrated by the present study, novices were able to identify LXD problems concerning interface navigation that experts overlooked (RQ1).

Limitations and Future Research

Although the results provide promising data challenging the application of Nielsen’s (1993) Five User Assumption to the evaluation of learning technologies, there are several possible limitations that must be considered. First, this study uses a non-experimental design that relies on descriptive statistics given the nature of the N=5 assumption; therefore, statistically significant differences cannot be made regarding the data. Future research studies should include larger sample sizes so that significant difference comparison tests may be conducted. Researchers could recruit larger samples and then group these samples into groups of five to challenge Nielsen’s (1993) Five User Assumption. Researchers could then use D-Prime to divide their sample size into multiple groups of five allowing future researchers to form more detailed comparison tests and analyze possible significant differences.
Moreover, this study explores the differences between pre-existing groups; therefore, the researchers inability to manipulate the independent variable threatens internal validity (Schenker & Rumrill, 2004). The results may suggest a relationship exists between the two variables, but the researcher cannot disregard the possibility of potential extraneous variables manipulating the results (Schenker & Rumrill; 2004). For instance, half of this study's sample was over age 55. The trend toward a higher age range could skew how participants perceive the present learning technology. Participants may find particular aspects of the technology more challenging than a younger population with higher technological literacy. One possible follow-up study could include a third intermediate experience group to assess LXD deviation identification. Adding a third group would provide a more comprehensive view regarding the potential impact expertise has on LXD deviation identification by assessing a wider range of experiences: novice, intermediate, and expert. It is possible that adding the third intermediate group could allow researchers to assess in depth when the shift from novice to expert begins to occur and what differences in criteria (e.g., deviation identification rate) indicate this change.

An additional limitation is the use of a purposeful, convenience sample. All participants were recruited from the sample business location and worked in the same department. This may bias the results because many of these participants have worked in this line of business for many years and share similar experiences and perspectives. These shared experiences and perspectives may cause the results to not be reflective of a more diverse general population. Using such a constricted sample threatens the external validity study. Not using random selection to recruit participants may cause the selected sample to not be representative of the overall population (Schneker & Rumrill, 2004).
Another potential limitation of this study is using one type of UX evaluation to evaluate the learning interface. Although Nielsen’s (1993) Five User Assumption is often associated with a cognitive walkthrough usability evaluation method, other usability evaluation methods may reveal different results. A future study could investigate how the type of UX evaluation could impact the problem identification rate. Celik et al. (2023) suggested that a collaborative comprehensive UX evaluation may provide a more comprehensive view of the technology being assessed than individual UX experts evaluating the same technology. A collaborative group evaluation could plausibly provide more in-depth insights regarding the technology being evaluated over independent evaluations. A group evaluation would allow multiple diverse perspectives to provide a more holistic LXD evaluation of learning technology. Moreover, fallacies like the expert blindspot may be less likely to occur.

The goal of LXD is to “focus on improving UX and LX of only one type of technology—learning technology—from the perspective of only one type of user—the learner” (Jahnke et al., 2020, p.1). Jahnke et al. (2020) suggests that considering the learner perspective is essential to improving the LXD of a learning technology. In alignment with this suggestion, many researchers in the literature have shifted away from only using expert evaluators to assess learning technologies (Choemprayong et al., 2021; Celik et al., 2023; Pereira et al., 2023; Sanabria et al., 2021; Teles et al., 2021). For example, a study by Celik et al. (2023) recruited diverse reviewers such as online learning experts, quality assurance testers, instructional designers, and end users. Results of the study suggest that expert heuristic evaluation is important; however, collaborative heuristic evaluation with evaluators from various backgrounds find more UX issues, allows for rapid UX evaluations for large scale projects, and is more cost
effective than individual expert heuristic evaluations (Celik et al., 2023). Extending from this research, a follow-up study could explore the type of LXD interactions found across differing groups. Moreover, researchers could use Signal Detection Theory to assess the agreement level of evaluators on true LXD problems and false alarms (i.e. identifying a dummy item as correct).

A final limitation concerning the present study is the use of one type of learning technology for evaluation. In the case of the current study, Gooru Learning was used for the LXD evaluation. Gooru learning is an open-source online learning platform. Other platforms, such as LinkedIn Learning or Coursera, use more complex interfaces and diverse methods of content presentation, which could result in findings that differ from the present study. A logical follow-up study could compare the differences between different groups of evaluators across different learning technologies. The complexity or the diversity of the technology may impact LXD deviation identification, severity rating, and classification.

**Conclusion**

Across the literature, researchers agree that evaluation is a essential element of the design process; however, instructional designers in the field acknowledge the importance of evaluation but indicate they are not able to evaluate as often as they want (Albert & Tullis, 2022; Lu et al., 2022; Schmidt et al., 2020). Reasons for this lack of evaluation include high costs and lack of participants (DeVaughn & Stefaniak, 2021). These hindrances and a lack of a single concept evaluation model for evaluation as led to LIDT practitioners to borrow from the UX field (DeVaughn, 2020; Griffin, 2011; Lu et al., 2022). These borrowing practices have led to the Nielsen’s (1993) Five User Assumption being used to justify the use of small sample sizes for
evaluation of learning technologies despite the validity of this assumption not being tested within learning technology evaluations (Tawfik et al., 2022; Faulkner, 2003).

In addition to providing potential suggestions concerning the type and number of evaluators necessary to complete an effective LXD evaluation, this study proposes that learning technologies contain unique complex interactions that go beyond the scope of technological usability testing, which has been recently argued by other researchers and theorists such as Lu et al. (2022), Schmidt & Huang (2021), and Tawfik et al. (2022). Learning technologies, unlike traditional technologies, contain not only technological usability aspects but also pedagogical usability aspects that focus on the overall user experience (Jahnke et al., 2020; Marell-Olsson & Jahnke, 2019; Moore et al., 2014; Nakamura et al., 2018; Quinones et al., 2018). In previous research, Tawfik et al. (2022) conducted a grounded-theory study to delve deeper into these complex interactions and identified two unique types of interactions that occur within learning technologies: interaction within the learning environment and interaction within the learning space. Building off this previous research, novices and experts in this study were both able to perceive and identify problems within the learning technology that aligned with Tawfik et al.’s (2020a) framework. The participants' ability to recognize not only interaction within the learning environment problems but also interaction within the learning space problems suggest that learning technologies contain multiple UX dimensions that include not only a technological dimension but a pedagogical dimension as well.

This research builds upon previous UX evaluation research by exploring whether Nielsen’s (1993) Five User Assumption holds true when applied to learning technologies (AlRoobaea & Mayhew, 2014). Moreover, this present study explores whether LXD experience
(expertise) impacts the LXD deviation identification rate. The results reveal that novices exceeded Nielsen’s (1993) 80% threshold, whereas experts did not meet this threshold.

Additional descriptive statistics suggest that expertise may impact LXD deviation identification (RQ1) along with LXD deviation classification (RQ3). Analysis of the severity ratings between groups revealed no significant differences between novices and experts on LXD deviation severity ratings (RQ2). The results of this study revealed novices and experts identified different numbers of ILE UX deviations and ILS UX deviations. Novices identified a higher percentage of ILE UX deviations and ILS UX deviations than experts. These results suggest that experience may influence the number of UX deviations and the type of UX deviations that evaluators identify. Finally, this study builds off previous research suggesting that different types of interactions occur when interacting with a learning technology; moreover, these results suggest problems within these types of interactions are perceivable to participants despite experience levels (Tawfik et al., 2022; Tawfik et al., 2024).
REFERENCES


Alroobaea, R., & Mayhew, P. J. (2014). How many participants are really enough for usability studies?. In *2014 Science and Information Conference* (pp. 48-56). IEEE.


https://doi.org/10.1109/HICSS.2005.468


Enrollment by levels and demographics. Enrollment by Levels and Demographics - Office of Institutional Research - The University of Memphis. (n.d.).

https://www.memphis.edu/oir/data/public_student_demographics.php


https://doi.org/10.2196/10932


https://doi.org/10.1080/10494820.2021.1912784


https://doi.org/10.1016/j.csi.2018.03.002


https://doi.org/10.3390/proceedings2019031006

**APPENDIX or APPENDICES**

The Appendix includes a variety of artifacts. The appendix may include the IRB approval letter, instruments, the participant letters, and training materials. If numerous types of artifacts are included as appendices, each type has a section labeled as Appendix A, Appendix B, etc.
Appendix A

Recruitment E-mail

Dear potential participants,

I am contacting you to invite you to be part of my research experiment on usability evaluations and learning technologies. The purpose of this research is to explore the possible relationship between expertise and usability problem identification, severity, and classification. Moreover, this study seeks to understand the optimal number of participants needed to complete a learning technology evaluation. We will be seeking participants with and without experience in evaluating learning technologies. If you choose to participate in this study, you will complete an informed consent, demographics survey, and an experiment. The experiment we created includes using Gooru education to complete eight learning tasks. During these tasks, you will complete a checklist. We anticipate this experiment to take no longer than 30 minutes. If you are interested in participating in this research study, please respond to this e-mail.

Thank you in advance for participating in our research study. If you have any questions, please do not hesitate to let me know!

Jessica

Appendix B.

IRB Letter
Appendix C

Informed Consent

University of Memphis Researcher(s):

Dr. Andrew Tawfik, 901-678-3851, aatawfik@memphis.edu

Jessica Gatewood, 901-451-3723, jgtwood3@memphis.edu

You are being asked to participate in a research study. The section below highlights key information for you to consider when deciding if you want to participate. More detailed
information is provided below the box. Please ask the researcher(s) any questions about the study before you make your decision. If you volunteer, you will be one of about 30 people to do so.

Key Information for You to Consider Voluntary Consent: You are being asked to volunteer for a research study. It is up to you whether you choose to participate or not. There will be no penalty or loss of benefit to which you are otherwise entitled if you choose not to participate or discontinue participation.

Purpose: The purpose of this research is to examine the amount of participants needed when completing a learning task.

Duration: It is anticipated that your participation will last no longer than 60 minutes.

Procedures and Activities:

A member of the research team a demographics survey about yourself. The member will then guide you through learning tasks to be completed using Gooru education.

Risk: Some of the risks that may be expected include discomfort or embarrassment by answering some questions. However, the risk or discomfort that you experience should be no greater than what you might experience from completing daily, routine work.

Benefits: There are no direct benefits to you; however, your responses may also help other schools determine how to better train and support teachers, students, and families regarding evaluation of learning technologies.
Who is conducting this research? Jessica Gatewood and Dr. Andrew Tawfik at the University of Memphis. There may be other research team members assisting during the study.

What happens if I agree to participate in this research? If you agree to participate, you will be asked to complete a demographics survey. Following the demographics survey, you will be asked to complete eight learning tasks in Gooru. After completing the eight learning tasks, participants will be asked to complete a checklist determining if a usability error was present, the severity of the error, and the classification. Participation is voluntary. You may skip any question that makes you uncomfortable and terminate your involvement at any time and for any reason without consequence.

What happens to the information collected for this research? Personally identifiable information gathered in this study will only be used to link data between the demographics survey and the Q sort data. Your name will be replaced with a pseudonym for subsequent data analysis and reporting. All data will be aggregated with others’ responses and analyzed. Identifying information (e.g., names, demographic information) will be de-identified in published reports, conference presentations, and meetings to better ensure confidentiality.

How will my privacy and data confidentiality be protected? We promise to protect your privacy and security of your personal information as best we can. Although you need to know about some limits to this promise.

Measures we will take include:
• Using only pseudonyms on analyses, reports, presentations, and publications

• Storing and analyzing all data on secure, university-controlled computer systems that are only accessible to the research team using secure, two-factor authentication.

• Storing a copy of the dataset for no longer than three years.

Individuals and organizations that monitor this research may be permitted access to inspect the research records. This monitoring may include access to your individual responses. These individual and organization include:

• Institutional Review Board at the University of Memphis

What if I want to stop participating in this research? It is up to you to decide whether you want to volunteer for this study. It is okay to decide to end your participation at any time by mentioning this to the interviewer. Additionally, you may ask at any time prior to presentations/publications that your data be removed from the study. There is no penalty or loss of benefits to which you are otherwise entitled if you decided to withdraw your participation. Your decision about participating will not affect your relationship with the researcher(s) or your workplace.

Will it cost me money to take part in this research? There are no costs associated with participation in this research study

Will I receive any compensation or reward for participating in this research? No
Who can answer my question(s) about this research? Before you decide to volunteer for this study, please ask any questions that might come to mind. Later, if you have questions, suggestions, concerns, or complaints about the study, you can contact the investigator, Dr Andrew Tawfik (901-678-3451; aatawfik@memphis.edu) If you have any questions about your rights as a volunteer in this research, contact the Institutional Review Board staff at the University of Memphis at 901-678-2705 or email irb@memphis.edu.

STATEMENT OF CONSENT I have had the opportunity to consider the information in this document. I have asked any questions needed for me to decide about my participation. I understand that I can ask additional questions through the study.

By signing below, I volunteer to participate in this research. I understand that I am not waiving any legal rights. I have been given a copy of this consent document. I understand that if my ability to consent for myself changes, my legal representative or I may be asked to consent again prior to my continued participation.
Appendix D
Demographics Survey

1. Name (Please note. Individual responses will not be shared. Only the research team has access to names. We will then use to aggregate and present as averages with no identifiable information)

2. Which of the following best describes the industry in which you work? Higher Education, K-12, Museum, Corporate, Healthcare, Library, Military/Government, Other; If Other, please type the name of your industry here.

3. With what gender do you identify? Female, Male, Non-Binary, Prefer Not to Answer

4. What is your country of residence?

5. What is your age? 18-24, 25-34, 35-44, 45-54, over 55, Prefer Not to Answer

6. What is the highest degree you hold? In what field?

7. What is your current job title?

8. How many years have you worked in the field? 1-3, 4-5, 6-9, 10+

9. Do you have experience with usability studies (evaluations)?" Usability refers to extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified content of use. If, so how many, years? 1-3, 4-5, 6-9, 10+
Appendix E

Nielsen Severity Rating

The following 0 to 4 rating scale can be used to rate the severity of usability problems:

0 = I don't agree that this is a usability problem at all
1 = Cosmetic problem only: need not be fixed unless extra time is available on project
2 = Minor usability problem: fixing this should be given low priority
3 = Major usability problem: important to fix, so should be given high priority
4 = Usability catastrophe: imperative to fix this before product can be released
Appendix F
Pilot Study Protocol

-Go to gooru.org

-Select “Explore as Guest”

-Select Student

-Select the course Skills for Employability.

Please explore the program. Pay close attention to the four different sections on the screen. As you explore each section, write down any UX issues that you note. Along with independently exploring the interface, please complete the following tasks and write down any UX issues that you note, rate their severity, and classify the problems in accordance with the provided constructs.

Task 1

- You want to begin a course on Skills for Employability where would you go to do that?
  What do you see as the learning objectives for that topic?

- Select the study once you have located it.

- Begin the lesson once you have located it.

- Note any usability issues that you believe are present, rate their severity, and classify the error.

Task 2
Task 2

- You wish to gain knowledge about the topic of ‘Professional Behavior’?
  - Where would you go to do that?
  - Please let me know when you are finished.
  - Note any usability issues that you believe are present and rate their severity. What change would you make and why?”

- Let’s say you want to document notes that you will revisit later.
  - Where would you go to do that?
  - Please let me know when you are finished.
  - Note any usability issues that you believe are present and rate their severity. What change would you make and why?”

**Return to Dashboard**

**Competency Gains**

Task 3

You wish to learn about which concepts you have mastered and identify what you still need to work on. Where would you go to do that? What is your next task in learning?

- Which concepts do you have the most masteries in?
- What concepts do you need to work on?
  - What information makes you think that?
  - How would you navigate to those concepts?
• Note any usability issues that you believe are present, rate their severity, and classify the error.

Task 4

You want to learn about the particular competencies for the Teamwork section, where would you go to do that?

• Which of those competencies have you mastered?
  
  o How do you know that?
  
  o What is your next task in learning?

• Note any usability issues that you believe are present, rate their severity, and classify the error.

Return to Dashboard

Review Progress and Performance

Task 5

You want to check your progress. Where would you go to do that? Based on the interface, what is your next learning task?

• What areas do you need to work on? Why do you believe that?

Follow Up

• Note any usability issues that you believe are present, rate their severity, and classify the error.

Next Task
**Pending Activities**

**Task 6**

You want to check if you have any assignments due. Where would you go to do that? What is your next learning task?

- Select one of the assignments
- You want to check how much time you have spent on the assignment. How would you do that?

**Return to interface**

**Go to the Learning Journey**

**Select “Professional Behavior”**

**Task 7**

Select the CFU – Professional Behavior

Answer and submit the questions.

- Where would you go to do that?
- Please let me know when you are finished.
- Note any usability issues that you believe are present and rate their severity. What change would you make and why?”
Appendix G
Full Study Protocol

-Go to goorulearning.com

-Select “Explore as Guest”

-Select Student

-Select the course Skills for Employability.

-Note: You are entering the interface as an “in progress” student.

Task 1

- You want to begin a course on Skills for Employability
  - Where would you go to do that?
  - Where would you select the lesson?
  - Please let me know when you are finished
  - Note any usability issues that you believe are present and rate their severity. What change would you make and why?”

Return to Dashboard

Select Learning Journey

Task 2

- You wish to gain knowledge about the topic of ‘Professional Behavior’?
Let’s say you want to document notes that you will revisit later.

○ Where would you go to do that?
○ Please let me know when you are finished.
○ Note any usability issues that you believe are present and rate their severity. What change would you make and why?”

Return to Dashboard

Competency Gains

Task 3

You wish to learn about which concepts you have mastered and identify what you still need to work on.

○ Where would you go to do that?
○ Please let me know when you are finished.
○ Note any usability issues that you believe are present and rate their severity. What change would you make and why?”

● Which concepts do you have the most masteries in?

○ Where would you go to do that?
○ Please let me know when you are finished.
○ Note any usability issues that you believe are present and rate their severity. What change would you make and why?”

● What concepts do you need to work on?
○ What information makes you think that?
○ Where would you go to do that?
○ Please let me know when you are finished.
○ Note any usability issues that you believe are present and rate their severity. What change would you make and why?”

Task 4

You want to learn about the particular competencies for the Teamwork section.

● Where would you go to do that?

● Please let me know when you are finished.

● Note any usability issues that you believe are present and rate their severity. What change would you make and why?”

● Which of those competencies have you mastered?
○ Where would you go to do that?
○ Please let me know when you are finished.
○ Note any usability issues that you believe are present and rate tNonheir severity. What change would you make and why?”
Return to Dashboard

Review Progress and Performance

Task 5

You want to check your progress.

○ Where would you go to do that?
○ Please let me know when you are finished.
○ Note any usability issues that you believe are present and rate their severity. What change would you make and why?"

Based on the interface, what is your next learning task?

○ Where would you go to do that?
○ Please let me know when you are finished.
○ Note any usability issues that you believe are present and rate their severity. What change would you make and why?"

Return to Dashboard

Task 6

You want to check if you have any assignments due.

○ Where would you go to do that?
○ Please let me know when you are finished.
○ Note any usability issues that you believe are present and rate their severity. What change would you make and why?”

Return to interface

Go to the Learning Journey

Select “Professional Behavior”

Task 7

Select the CFU – Professional Behavior

Answer and submit the questions.

○ Where would you go to do that?

○ Please let me know when you are finished.

○ Note any usability issues that you believe are present and rate their severity. What change would you make and why?”
### Interaction within the learning environment (UX)

<table>
<thead>
<tr>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Customization of the interface</td>
</tr>
<tr>
<td>2. Expectation of content placement</td>
</tr>
<tr>
<td>3. Functionality of component parts</td>
</tr>
<tr>
<td>4. Interface terms aligned with existing mental models</td>
</tr>
<tr>
<td>5. Navigation</td>
</tr>
</tbody>
</table>

### Interaction within the learning space (LX)
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Engagement with modality of content</td>
<td>Thoughts on learning design format, aesthetics, and users’ desire to engage with elements used on learning interface</td>
</tr>
<tr>
<td>2. Dynamic interaction</td>
<td>Interaction that engenders learners’ desire to continue or discontinue in their self-directed learning</td>
</tr>
<tr>
<td>3. Perceived value of technology feature to support learning</td>
<td>Perceived value of a specific technology feature impacted one’s learning</td>
</tr>
<tr>
<td>4. Scaffolding</td>
<td>Cues, hints, etc. that expanded the learners’ prior knowledge.</td>
</tr>
</tbody>
</table>
Appendix I

Experiment Checklist including Dummy Items
Note: Dummy items are bolded.

UX Study

For each task you complete for the experiment, you will review the checklist for each task. If you agree that the problem listed applies to the task you completed, then check the box for that problem. After you select the usability problems please rate the severity of the problem in the boxes provided.

1. Name (Please note. Individual responses will not be shared. Only the research team has access to names. We will then use the data to aggregate and present as averages with no identifiable information)

2. Task 1: Select all the usability problems you believe apply to this section. Multiple choice.

   Navigation around the interface is confusing
   The placement of the lessons should be more prominent
   Interface terms are not clearly defined
   How to use/navigate the Resume Study section is unclear
   The aesthetics of the interface are overwhelming
   Navigation from dashboard to lessons is difficult
   Icon representation is misleading
   **Links on the interface do not work**
   The font is too small
   Interface does not show your progress
   The interface needs to display more instructions
   The interface content is not engaging
   How to begin a lesson is not clear

3. For each box you selected for Task 1, please list your answers below and rate the severity of the problem. Please use the following rating system:

   1 = Cosmetic problem only: need not be fixed unless extra time is available on project
   2 = Minor usability problem: fixing this should be given low priority
   3 = Major usability problem: important to fix, so should be given high priority
   4 = Usability catastrophe: imperative to fix this before product can be released

4. Task 2: Select all the usability problems you believe apply to this section.
Video for How to develop positive work relationships does not work
The purpose of the time listed next to the lessons is not known
Meaning of interface terms are unclear
Spelling is inconsistent
Icon representing the notes is unclear
The placement of the notes when they appear on screen needs to be changed
Instructions are incorrectly labeled
Notes were not where you expected them to be
Navigation issues are present: clicking the progress bar takes you back to dashboard
Terminology is confusing because it is not defined
How to start the lesson is not clear
The content is poor quality and not engaging
Lesson interface uses too many colors
Study timer does not work
Content is not easy to understand
More scaffolding is needed for the learner
The note section does not add value to learning
The lesson uses too many videos
There is too much text
Learning content is not engaging

5. For each box you selected for Task 2, please list your answers below and rate the severity of the problem. Please use the following rating system:

1 = Cosmetic problem only: need not be fixed unless extra time is available on project
2 = Minor usability problem: fixing this should be given low priority
3 = Major usability problem: important to fix, so should be given high priority
4 = Usability catastrophe: imperative to fix this before product can be released

6. Task 3: Select all the usability problems you believe apply to this section.
The graph is confusing to the learner
The colors of the interface are inaccessible for colorblindness
The colors make understanding the interface difficult
The interface does not provide information about when you mastered a topic

7. For each box you selected for Task 3, please list your answers below and rate the severity of the problem. Please use the following rating system:

1 = Cosmetic problem only: need not be fixed unless extra time is available on project
2 = Minor usability problem: fixing this should be given low priority
3 = Major usability problem: important to fix, so should be given high priority
4 = Usability catastrophe: imperative to fix this before product can be released
8. Task 4: Select all the usability problems you believe apply to this section.

   Navigation for this section is difficult and problematic
   Where content is located is confusing
   **Competencies for each sections are not explained**
   **The topics for each lesson are not clear**

9. For each box you selected for Task 4, please list your answers below and rate the severity of the problem. Please use the following rating system:

   1 = Cosmetic problem only: need not be fixed unless extra time is available on project
   2 = Minor usability problem: fixing this should be given low priority
   3 = Major usability problem: important to fix, so should be given high priority
   4 = Usability catastrophe: imperative to fix this before product can be released

10. Task 5: Select all the usability problems you believe apply to this section.

    Text is small
    Interfaces aesthetics are unattractive
    Too much text
    **Says no assignments due but has assignments**
    **The labels for each section are confusing**
    **Interface does not show time spent on lessons**

11. For each box you selected for Task 5, please list your answers below and rate the severity of the problem. Please use the following rating system:

    1 = Cosmetic problem only: need not be fixed unless extra time is available on project
    2 = Minor usability problem: fixing this should be given low priority
    3 = Major usability problem: important to fix, so should be given high priority
    4 = Usability catastrophe: imperative to fix this before product can be released

12. Task 6: Select all the usability problems you believe apply to this section.

    Says no assignments due but shows assignments
    Items to Grade section does not make sense for a student
    **There is too much text on this page**
    **More instructions are needed**
    **Purpose of this section is unclear to the learner**
13. For each box you selected for Task 6, please list your answers below and rate the severity of the problem. Please use the following rating system:

1 = Cosmetic problem only: need not be fixed unless extra time is available on project
2 = Minor usability problem: fixing this should be given low priority
3 = Major usability problem: important to fix, so should be given high priority
4 = Usability catastrophe: imperative to fix this before product can be released

14. Task 7: Select all the usability problems you believe apply to this section.

No submit button until the final question
Instructions are needed for the quiz
No feedback on answers
Answer key is unclear as to what is right/wrong
Inaccessible for colorblindness
Placement of the quiz answers is confusing
**Interface does not show time spent on quiz**
**Background images are distracting to the learner**
**Quiz does not add value to learning**
**Quiz questions use to much text**

15. For each box you selected for Task 7, please list your answers below and rate the severity of the problem. Please use the following rating system:

1 = Cosmetic problem only: need not be fixed unless extra time is available on project
2 = Minor usability problem: fixing this should be given low priority
3 = Major usability problem: important to fix, so should be given high priority
4 = Usability catastrophe: imperative to fix this before product can be released
Appendix J

Novice v. Expert Severity Rating Averages Across Tasks

<table>
<thead>
<tr>
<th></th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>N4</th>
<th>N5</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>(N=7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(N=9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>(N=2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>(N=2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>(N=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>(N=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>(N=4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.14</td>
<td>0.14</td>
<td>0.86</td>
<td>0.86</td>
<td>0</td>
<td>0</td>
<td>0.86</td>
<td>0.86</td>
<td>2.29</td>
<td>2.29</td>
</tr>
</tbody>
</table>

Average

| SD | 0.38 | 0.38 | 0.90 | 0.90 | 0 | 0 | 0.89 | 0.89 | 1.11 | 1.11 |

123