Sensemaking With Learning Analytics Visualizations: Investigating Dashboard Comprehension And Effects On Learning Strategy

by

Halimat Alabi

M.Ed., University of Victoria, 2013

M.A., San Diego State University, 2007

B.S., Purdue University, 2001

Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in the

School of Interactive Arts & Technology

Faculty of Communication, Art, and Technology

SIMON FRASER UNIVERSITY

Spring 2021

Declaration of Committee

Name:	**Halimat Alabi**
Degree:	**Doctor of Philosophy**
Thesis title:	**Sensemaking With Learning Analytics Visualizations: Investigating Dashboard Comprehension And Effects On Learning Strategy**
Committee:	**Chair: Cheryl Geisler** Professor, Interactive Arts & Technology
	Marek Hatala Supervisor Professor, Interactive Arts & Technology
	Brian Fisher Committee Member Professor, Interactive Arts & Technology
	Steve DiPaola Examiner Professor, Interactive Arts & Technology
	Katrien Verbert External Examiner Associate Professor, Computer Science KU Leuven

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

In the provision of just-in-time feedback, student-facing learning analytics dashboards (LADs) are meant to aid decision-making during the process of learning. Unlike summative feedback received at its conclusion, this formative feedback may help learners pivot their learning strategies while still engaged in the learning activity. To turn this feedback into actionable insights however, learners must understand LADs well enough to make accurate judgements of learning with them. For these learners, LADs could become an integral part of their self-regulatory learning strategy.

This dissertation presents a multifaceted examination of learners' sensemaking processes with LADs designed to support self-regulatory learning. The in-situ studies detailed therein examine learners' understanding of the data visualized in LADs and the effects of this understanding on their performance-related mental models. Trace data, surveys, semi-structured in-depth qualitative interviews, and retrospective cued recall methods were used to identify why, when, and how learners used LADs to guide their learning. Learners' qualitative accounts of their experience explained and contextualized the quantitative data collected from the observed activities.

Learners preferred less complex LADs, finding them more useful and aesthetically appealing, despite lower gist recall with simpler visualizations. During an early investigation of how LADs were used to make learning judgments in situ, we observed learners' tendency to act upon brief LAD interactions. This inspired us to operationalize gist as a form of measurement, describing learners' ability to make sense of a LAD after a brief visual interrogation. Subsequent comparisons of the accuracy and descriptiveness of learners' gist estimates to those of laypeople repeatedly showed that laypeople were more apt than learners to produce accurate and complete gist descriptions. This dissertation culminates in a final study examining the evolution of learners' mental models of their performance due to repeated LAD interaction, followed by a discussion of the contextual factors that contributed to what was observed. Trends observed across this work suggest that learners were more apt to "get the gist" with LAD after repeated interaction. This dissertation contributes a novel method for evaluating learners' interpretation of LADs, while our findings offer insight into how LADs shape learners' sensemaking processes.

Keywords: Learning analytics; visualization; dashboard; gist; online learning; asynchronous discussion group

Dedication

This work is dedicated to three beloved individuals, Lorraine Alabi, Ifetayo Alabi, and Krusheska Quiros.

Mum, thank you for your unceasing belief in my ability to succeed, and for imbuing me with a deep respect for the many opportunities that education may provide. Your life – modeling a woman with equal parts ingenuity and kindness, grit and empathy, strength and intelligence – made everything I do and have done possible. You taught me how to work and how to hold hope, both incredibly important abilities when undertaking an endeavour of this magnitude. I am grateful for you and your endless demonstrations of love. I hope to continue your legacy for the rest of my days.

Ifetayo Zarine, my golden child, you are my inspiration and my light, and I love you with my whole heart. I believe that children choose their parents – you timed your entry into this world to be present for every one of my degrees. You contributed to this work; doing this with you by my side every step of the way was precious, as it reinforced who I am holding hope for. I want you to have the tools to be able to create opportunities for yourself, to never be afraid to take chances because you are solid in your belief in your many abilities. With this, my final degree, I pass the torch on to you. As you make your way in the world, I hope you never lose your curiosity. Your vision is uniquely yours; I hope you never stop questioning, dreaming, or making. You will always have my love and support; it is an honour to be your mother.

Krusheska, you joined this PhD journey at the most difficult point. Thank you for making me and this work a priority in your life from day one. You put in more hours with me at the library than many who worked there, and your intellectual contributions helped me translate my handwaving to text, with a very fancy pen to boot! You held it down at home while everything in our lives drastically changed, supporting me through one of the most challenging years of my life. Simply put, you are the ride or die of my dreams. I look forward to many more adventures with, and time to enjoy the fruit of our labour.

You are my favorite collaborators. Thanks to the three of you for your unconditional love, support, and encouragement.

Acknowledgements

I would like to express my deepest appreciation to Dr. Marek Hatala and Dr. Brian Fisher for your guidance over the years; I benefit from your wisdom, expertise, and the countless hours you have devoted to making me a better researcher. I would like to thank members of my committee, Dr. Steve DiPaola and Dr. Katrien Verbert, for your insightful feedback and the inspiration that your research has provided me.

Thanks to my lab for your feedback and camaraderie, with special thanks to Sanam Shirazi, Varshita Sher, and Sina Nazeri. Tiffany Taylor, thank you for your endless encouragement and enthusiastic support. I appreciate your patience with me, and with all of the other students, faculty, and staff in our department.

To the SFU library family – especially Leanna Jantzi, Soo Oh, Alisa Kuan, and Michael Chang – thank you for the sorely needed respite from my writing stress. Thank you for your commitment to supporting the SFU community – you are the unsung heroes of academia. I will return the 32 books I currently have on loan; I promise. I just need one more extension…

Mentorship is far too often a thankless task. I would like to extend my gratitude to two of my mentors whose early input in my career was particularly meaningful on this PhD journey, Jeff Vargas and Ken Allen. You are on a very short list of people I would follow into any venture, at any moment, anywhere in the world. You both created a space for me to experiment and grow, while modelling the utmost in professionalism. That legacy lives on in me, and the lives of the students I work with.

Ken, you saw something in me that I had yet to develop. All these years later, I still strive to lead with the grace and integrity that you exemplify. I did not get enough time under your tutelage but suffice to say, I think I did a good job with what I had.

Jeff, thank you for giving me the opportunity to innovate professionally for the first time in my life – you set these wheels in motion. Working with you was a master class in creativity, team building, and conflict management; I quote you to this day, my homie for life! Thank you for taking a chance on me, and for teaching me how important it is to be able to bring your whole self to work. I took that lesson forward with me to

every place I have worked since. I know that it has made a difference in many of my students' lives. I miss our Friday morning Boondocks routine like nothing else!

I would be remiss if I did not acknowledge my friends Cynthia Brooke and Desaraigh Byers. Cynthia you have held me up – both literally and figuratively – and made Vancouver feel like home. Desaraigh, your gifts of language and insight fed my soul, when you were not busy feeding me fried chicken. Thank you both for lending me perspective whenever mine was skewed and continuing to bless me with your friendship.

I am deeply grateful to SSHRC for the early recognition of my work and the funding to make it possible. Finally, I am indebted to the many students who participated in my research and whose learning paths have shaped me as an educator over the years.

Table of Contents

List of Tables

List of Figures

List of Acronyms

HIT	Human intelligent task
LA	Learning analytics
LAD	Learning analytics dashboard
LMS	Learning management system
MTurk	Amazon Mechanical Turk
MTurkers	Amazon Mechanical Turk workers
MOOC	Massively Open Online Course
SRL	Self-regulated learning

Chapter 1.

Introduction

This dissertation presents a multifaceted examination of learners' sensemaking processes with learning analytics dashboards (LADs), designed to support self-regulatory learning. As an online instructor I witnessed far too many intelligent, competent adults struggle in online learning environments – unable to accurately assess their progress, to find deficits in their academic strategies, or their place within their learning cohort. Of the many challenges inherent to learning online, the first and perhaps most significant is the inability to accurately ascertain assess one's own progress. If properly understood, learners would be able to use LADs to make timely learning judgments and meta-cognitively monitoring their progress. The fundamental question is then, ***how do learners interpret what they see visualized in LADs, and what do they do with this information***? To answer this question, this dissertation draws from theories of education, cognition, and information visualization. It contributes a novel method for evaluating learners' interpretation of LADs; our findings offer insight into how LADs shape learners' sensemaking processes.

1.1. History of learning analytics

Learning analytics (LA) is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs." (Siemens, 2011) LA has parlayed off the adoption of academic analytics in higher education, albeit with different goals. The term *academic analytics* came into use around 2005, it describes the use of student data by institutions for retention purposes (Baepler & Murdoch, 2010). Using predictive analyses and data from both high school and collegiate coursework, combined with recent learning behaviors and assessments, academic analytics have been implemented at a number of universities (Baepler & Murdoch, 2010) to support the operations, recruitment, student support and retention (van Barneveld et al., 2012). While

the primary goals of academic analytics stem from business needs such as resource administration, retention, and finance, Learning analytics target curriculum and instruction, addressing learner success and retention from *inside* the classroom during the process of learning (van Barneveld et al., 2012). Though the terms *academic analytics* and *learning analytics* are often used interchangeably, the primary difference between the two is that learning analytics are focused on informing and empowering educators and learners (Siemens & Baker, 2012). The term learning analytics dashboard (LAD) describes learning analytics designed for learners that "aggregates different factors about learners, learning processes, and learning context, into one or multiple visualizations." (Schwendimann et al., 2017) As one type of learning analytics, learner-facing LADs are specifically meant to inform learners' decisions during the process of learning.

One of the earliest implementations of LA was Course Signals. It is perhaps the most well-known and highly cited example of a LAD (Arnold & Pistilli, 2012; Devaney, 2010; Sclater et al., 2016). Course Signals has had a positive impact on thousands of learners since it was first deployed, helping them earn one letter grade higher than they would have without access to the LAD (Arnold & Pistilli, 2012; Devaney, 2010). Further, the successful adoption of this tool continues to exert great influence on the development of LA as a field, and the willingness of universities to implement learner-facing LADs.

Course Signals was originally designed for educators to alert learners that they may be at-risk of failing. It was later modified to visualize the status of learners' performance for them, though this required educators to manually trigger updates. The Course Signals LAD was based on a stoplight metaphor; performance to date was indicated by red, yellow, or green lights. It provided no personalized feedback; clicking on the signal yielded a standard list of learning resources. The overly simplistic visualization conveyed little information, making students reliant on educators to interpret the data for them before next steps could be determined. Its lack of scalability would make it difficult to utilize in fully online learning environments. While technically learner-facing, in its failure to provide timely, actionable insights directly to learners during the process of learning, Course Signals is an example of a LAD that was not designed for learners to control their own learning.

Visualizations designed for educators' use far outnumber those designed for learners (Verbert et al., 2020), perhaps for good reason. First, there are significant differences in the ways educators and learners utilize dashboards. Educators have historical knowledge of a course, its position within the curriculum landscape, and the skills that learners must develop to be successful in their program of study. They have a pedagogical grounding that influences how they value the individual learning activities as part of the overall learning experience. If the educator has previously taught the course, they are aware of performance patterns seen over its duration and are better able to situates an individual learners' progress in relationship to the class. For example, an educator often knows if a poor grade on a midterm is recoverable, or if the signifies a learner should repeat a course. This awareness makes educators better poised to gauge the impact of learners' study habits, interaction patterns, assessments, and overall performance. It influences how educators utilize dashboards, and it could be argued that educators are better able to leverage their logical expertise in the assessment of learning data. If it cannot be assumed that learners are able to perform visual analyses as adeptly as educators or properly determine what they should do next, then it is inappropriate to give learners the same dashboard visualizations without first ensuring that learners are able to reap similar benefits.

By the time they reach university, learners have had at least 10 years of experience reviewing their academic assessments and achievements. Learners' interpretation of LADs could be swayed by these longstanding impressions, or individual differences in goals or motivations (Beheshitha et al., 2015b; Dringus, 2012). It is important to know how learners understand LAD visualizations because this understanding effects the accuracy and quality of learners' mental models of their performance, which in turn influence learners' strategy enactment.

Jointly, these factors speak to the need to better understand how learners use LADs to think about their learning. Insight into learners' cognitions while using LADs will help to ensure that learners are able to use the formative feedback provided by LADs to support strategic self-regulatory learning. In the examination of learners' sense making processes with LADs, this research contributes to ongoing work in learning analytics that centers the goal of learner empowerment.

1.2. Potential impact of increased adoption & implementation

In an increasingly digital culture, the ability to reliably and accurately parse visual information is one of the most important aspects of digital literacy. Requiring both technological and cognitive skill, *digital literacy* is described as an ability to "use information and communication technologies to find, evaluate, create, and communicate information" (Digital literacy, n.d.). It requires the ability to source information and think critically about it; it also involves knowing how to use digital tools to engage, communicate, and collaborate (Digital literacy, n.d.). Recognizing the social and economic relevance of digital literacy, the policy makers of educational bodies around the world have been revising curriculum to incorporate the development of digital literacy, an important 21st Century skill (Emily R Lai, 2012; Forum, 2016; Foundation for Young Australians, 2017; Koenig, 2011; OECD, 2018).

MediaSmarts, Canada's Centre for Digital and Media Literacy, characterizes digital literacy three ways, as 1) skill and ability, 2) the capacity for critical understanding, and 3) the knowledge and expertise to create and communicate with digital technologies. One aspect of digital literacy is *visual literacy*, defined as an individual's ability to "use, interpret, analyze, and think critically about visual images and the significance of what they are seeing" (Bamford, 2003). Associated with traditional literacies such as language proficiency, visual literacy goes beyond comprehension, including the ability to critically analyze what is viewed in order to make judgements about its accuracy and validity. Children begin to accurately read visual images around one year of age; by three they are able to produce images to communicate graphically (Bamford, 2003). It is often assumed that learners will automatically understand graphical representations of data, even though graph comprehension is a learned skill (Glazer, 2011).

Workplace computational skill, including the capability to interpret data, is sought after by "every industry embarking on digital transformation." (Venkatraman et al., 2019) The need for a trained workforce who can derive value from data is unprecedented and exists across the range of industries for manufacturing to media, banking to entertainment. To address this need, institutions of higher education are revisiting their

curricula (Torii, 2018; STEM Partnerships Forum, 2018) to increase the employability of their student populations. In an effort to develop learners' visual and digital literacies, educational technologies – including LA – are being introduced as early as primary school (Jaakonmäki et al., 2020; Rodríguez-Triana et al., 2015; Molenaar, 2019). Even and perhaps especially in technology rich environments, learners cannot be assumed to possess digital fluency. Social discourse on so called *digital natives* forwards the idea that this generation is innately gifted with technological abilities (Selwyn, 2009). This attitude cannot be carried over into the design of applications for digital natives. Just as access to a basketball does not qualify someone for the NBA, the availability of ubiquitous technology does not imbue digital natives with the ability to perform analyses with these technologies. **It is important then, to better understand how learners' cognitions with LADs emerge, and how they evolve with repeated use.** Increasing learners' ability to make accurate judgements with visualizations builds their fluency with these kinds of information technologies, as they practice important workplace competencies. Addressing variations in the digital fluency of a populace has the potential to reduce the existing digital divide (Wei & Hindman, 2011).

1.3. Dissertation organization

The next chapter explores the pedagogies that LAD support in online learning environments, to better understand the theoretical, philosophical, and ideological underpinnings of LADs. Chapter 3 gives a contemporary perspective of student-facing LADs, touching upon recent trends, issues, design and evaluation methods. Chapter 4 provides a foundation in visual cognition and perception as related to the interpretation and evaluation of LADs. The concept of gist is introduced in chapter 4; it will be operationalized in the experiments described in chapters 5-11. Rather than a standalone methods section, the methods utilized in each quasi-experiment are detailed in the chapter about that experiment.

Chapter 6 describes the first experiment, a pilot study that asked learners to complete tasks and make gist assessments using three informationally equivalent LADs. Administered completely online, this study was undertaken to determine if learners' gist

recall or task accuracy were influenced by visualization type or factors of individual difference such as spatial acuity, cognitive reflexivity, subjective or objective numeracy.

Chapter 7 details experiment 2, a two-part exploratory study undertaken to determine how learners interacted with LADs during the learning process. In the first phase of this study, semi-structured qualitative interviews were conducted with learners after they used LAD during a 7-to-10-day small group discussion activity. Retrospective cued recall methods were used in the interviews to garner feedback on how and when LADs were used to make learning judgments during the discussion activity, using re-creations of the LADs learners saw during the discussion activity.

Learners who choose not to interact with LADs forfeit any of their potential benefits. From a design perspective, a better understanding of what forms learners' initial impressions of LADs may contribute to their adoption and ongoing use. In the second part of the exploratory study we compared learners' impressions of eight new LAD prototypes before and after interacting with them, to see if learners' initial impressions would be persistent. Learners performed a forced choice ranking before and after performing a cognitive walk-through with each visualization. Results were again controlled for factors of individual difference, to see if this influenced learners' preferences.

Feedback from the exploratory study interviews revealed that learners – even those who successfully utilized the LADs – only briefly attended to them. This prompted the subsequent study of gist in the next two experiments, to determine what learners understood from brief LAD interactions.

Chapter 8 describes Experiment 3, Conceptual Features of Abstract LADs, which compared the accuracy and descriptiveness of learners' gist assessments to those of Amazon Mechanical Turk workers (MTurkers), who represented laypeople in this study. The LAD visualizations were based on three types of natural scenes, representing 3 varying levels of complexity. Abstract visualization types were selected for this study to see if aspects of the human visual system prioritized one visualization type over others. Administered completely online, in this study participants were asked to describe gist after a brief 30 second exposure to each LAD. By interrogating gist, we contribute to the

design of future LADs by establishing an empirically validated baseline of what learners perceive in a glance.

Chapter 9 describes Experiment 4, Proportional Estimates of Gist, in which learners and MTurkers were again asked to describe the gist of three different types of LADs after a 30 second exposure. This experiment extended the work of the previous study by repeating the measurement of gist between learner and MTurker populations. Again, the 3 LADs represented three different levels of complexity; they were chosen based on their purported benefits in the facilitation of estimations of proportion. The study was undertaken to see if learners produced more accurate or complete gist descriptions than MTurkers, and if this varied due to numeracy or visualization type. Previously we evaluated gist assessments for their accuracy and/or descriptiveness. In this study we measured the completion of the gist assessment, which better described the phenomenon of interest.

Experiment 5, Stability of LAD-based Mental Models, described in chapter 11, was undertaken to better understand the role LADs played in shaping learners' mental models of their performance, and to see if the gist gleaned from the LADs persisted over the course of the learning activity. Learners used a LAD designed for this experiment during a 7-day learning activity; semi-structured interviews were conducted shortly after the activity to determine the stability of their mental models. Retrospective cued recall methods were used in the interviews to prompt rich descriptions of gist and the learning context. This experiment was conducted entirely online due to COVID restrictions.

Chapter 12 concludes the dissertation with a discussion of the research contributions, limitations, and recommendations for future research directions.

Chapter 2.

Student-facing LADs

2.1. Learning online

The term *online learning* is broadly descriptive and represents distance education delivered wholly or in part online. Most often, online learning is delivered through a learning management system (LMS) such as Canvas[1] or Moodle[2]. The LMS is where the majority of learning activities take place, as it houses the course content, resources, and means of communication between learners and educators. Instruction in online classes may be completely self-directed, educator directed, or some combination thereof. Online learners often have more demands on their time and attention and fewer ways to engage with their universities than face-to-face learners (Meyers, 2014), making engagement even more important in this modality. Given the high levels of autonomy required to successfully learn online, LA that support self-regulation may have a particularly large impact for online learners (Dabbagh & Kitsantas, 2004; Hartley, 2001; Schunk & Zimmerman, 1998).

Hirumi's (2002) interaction framework of online learning represents the interworking processes that together represent learning that happens in an online environment (see Figure 1). The framework consists of three levels, each connects to an individual learner. The first level describes interactions with the self. These learner-self interactions are the cognitions that happen within the learner, including metacognitive monitoring and self-regulation. The second level describes interactions with resources in the online environment, both learner-human and learner-nonhuman interactions. Human interactions include those with peers within the learning management system (LMS) and others outside the learning environment such as discipline specific groups and professional organizations. Non-human interactions include those supported by the LMS,

[1] https://canvas.instructure.com/

[2] https://moodle.org/

the LMS interface, the learning content, and the LAD. Learner-instruction interactions happening in the final level describe any action toward the achievement of learning outcomes. Though interactions with the LAD take place on the second level of interaction in this framework, they may induce the learner to interact on all the other levels of the framework. LAD interactions may prompt the learner to interact with the instruction, the environment, or their human counterparts in new and meaningful ways.

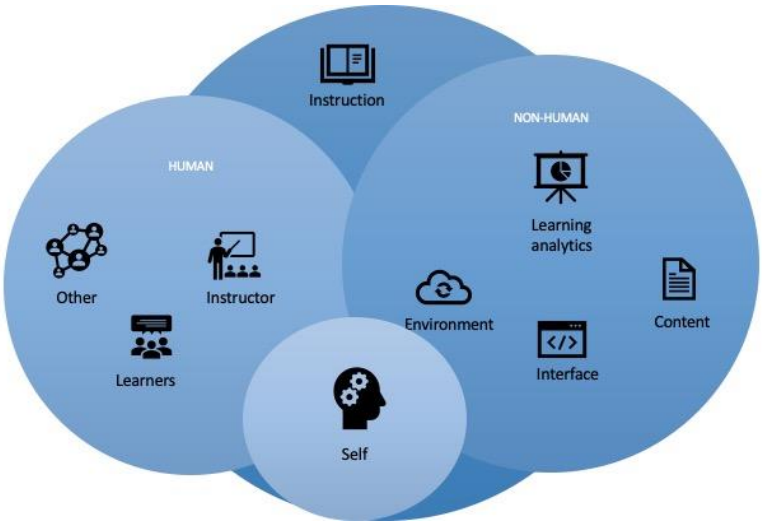

Figure 1. Hirumi's (2002) interaction framework of online learning

Student-facing LADs have the potential to help learners metacognitively monitor their progress throughout the learning process – from the development of a learning strategy, to the monitoring and regulation of its success. LADs are used by learners to make judgments that direct strategic learning. These judgments influence both planning and action. In the provision of feedback, LADs may guide learners in the identification and rectification of maladaptive behaviors as we saw with Course Signals, or used as educational interventions (Wise, 2014; Wise et al., 2014a; Wong & Li, 2019). Within the context of online discussions LA may be utilized to determine how learners interact with their peers, including who they interact with and when, when or how they engage with the learning materials or adopt domain language, and the planning of all such activities. This is of course, dependent upon multiple factors, such as if the learner chooses to use the LADs, if they understand the visualizations of their activity, and if they are aware of the next steps they should take to achieve their goals. LADs may affect the accuracy and quality of learners' mental models of their performance, both positively and negatively

(Dringus, 2012). Additionally, individual differences in goal orientation (Beheshitha et al., 2015b) have been shown to effect how successful learners are with LADs, so care must be taken to ensure that LADs do not negatively impact learning.

Learners do not have the same pedagogical grounding as educators to interpret visualizations of their learning efforts. Though learners may have access to peer results, overall, educators have a more global perspective. The design of student-facing LADs must take this into account, presenting the data in such a way that learners are able to assess their own performance, and situate it within the overall performance of their class. Further, learners must be able to appropriately infer the importance of the data visualized and the magnitude of their learning-related behaviors. This is challenging, since how learners draw conclusions from LADs is as yet poorly understood.

2.2. LAD trends and issues

Common criticisms of LADs include a lack of focus on the learners' perspective (Ferguson, 2012; Duval 2011), failure to incorporate pedagogical foundations in their design (Jivet et al., 2018), failure to measure the appropriateness of the incorporated visualizations for learners' visual literacy levels (Schwendimann et al., 2017), and a lack of evidence that LADs improve learning outcomes (R. Ferguson & Clow, 2017). Additionally, recent systematic reviews of student-facing LADs (Bodily & Verbert, 2017; Bodily et al., 2018; Schwendimann et al., 2017) have revealed that much of the research concerning LADs is exploratory, early stage or proof of concept work (Schwendimann et al., 2017). As more student-facing LADs are being developed, the number of LADs studied during deployment with learners is bound to increase.

There has also been a call for more diverse methods of study and evaluation of LADs to determine how they are being used during the process of learning to develop competency in one or more areas. In their review of LADs for learners, educators, and researchers, Viberg et al. (2020) highlighted the need for more mixed method and qualitative studies. In their review of 93 student-facing LADs, Bodily and Verbert (2017b) found only two papers that addressed how LADs impacted students' learning-related behaviors. In the work they reviewed, more commonly (34%) learners'

perceptions of the design were reported. The authors concluded that to advance the field, more tools need to be validated through use in situ.

Some reviewers have also offered advice to LAD designers and researchers on how to advance the field. For example, Schwendimann et al. concluded their literature review with a designer checklist. First, they directed researchers and designers to define how they define terms such as "learning dashboard." They suggested that designers provide details about the technologies used, the educational contexts of use, targeted learning constructs and or impacts including new practices that developed using the dashboard, and personal details about the users. Finally, they suggested that future research should include evaluation of the dashboard according to its impact on learning (Schwendimann et al., 2017).

Bennett and Foley's advice stemmed from their own research. They conducted interviews with 24 undergraduates in their final year of study, to see how these learners received LAD feedback after seeing their learning data visualized for the first time (Bennett & Folley, 2019). Bennett and Foley distilled the results of three questions – how learners felt reviewing their data, whose responsibility they felt it was to act on it, and if they would take any action as a result of what they saw visualized in the LADs into four LAD design principles supporting student agency and empowerment. First, learners should be able to customize LADs based on their goals. This advice was based on the interpretation of criterion-referenced data using a three-level rating, and how it was interpreted differently by learners. If the LADs may be customized, then a person whose goal is staying above average would not see the same visualization as someone striving for high marks. The second principle was to foreground student sensemaking, to aid in the interpretation of the LAD. The third principle was to aid in the identification of actionable insights, so they may understand exactly what they need to change to achieve their goals. The final principle was to embed dashboards in educational processes.

2.3. The user experience

An early criticism of LADs was that they did not consider the user experience, whether aesthetic or usability related (Verbert et al., 2011). Though much of the work

reported in the Schwendimann et al. review was exploratory, they did report that more dashboards were being evaluated in authentic settings (2017). While the majority of papers (58%) contained no evaluation, 15 papers (29%) mentioned evaluating the dashboards with stakeholders, and 19 papers mentioned evaluating the dashboards with learners. From the LADs evaluated with learners, these evaluations typically used a mixed-method approach (65%) and included 30 to 150 learners (Schwendimann et al., 2017). Many used trace data – including assessments and grades, date and time information from interactions with the content, peers, or educators –to capture differences in the ways individual learners organized their learning, from goalsetting and planning, to the choice, enactment, monitoring, and management of learning strategies (Selwyn & Gašević, 2020). Most of the papers (74%, 23 papers) addressed usability, often in the provision of feedback that would then be used to improve the LADs' design. Seven cited feedback on improved awareness, 5 evaluated changes in motivation or behavior resultant from dashboard use, and 4 evaluated its impact on learning. Schwendimann et al. (2017) found only 4 papers describing student-facing dashboard research that used multiple data sources, and that performed authentic evaluation with students.

The lack of trust that many students experience with LA is at this point, almost endemic of the field. Some learners, and even educators, fear that LA are an instance of "big brother," in the classroom, implemented as institutional surveillance, rather than as a pedagogically validated learning resource. It is a common misconception that the focus of LA is tracking (Selwyn & Gašević, 2020); the early focus of LA on prediction likely contributed to this. LA provides education – a discipline often reliant on self-report data that is vulnerable to bias and recall errors – a source of data that is able to quantitatively map change over time (Selwyn & Gašević, 2020). The data used in LA is a proxy for learning; in designing LA we must be specific about why the these indicators were chosen as representations of learning, and how they will be used (Selwyn & Gašević, 2020).

2.4. Sensemaking

Though sparse, there is ongoing research into how learners use LADs to think. Two recent works (Lim et al., 2019; Tervakari et al., 2014) specifically address learners' sensemaking with LAD. Both tested peer comparison, the most commonly used frame of reference, for student-facing LADs (Jivet al., 2017).

Lim et al. (2019) used qualitative and quantitative methods to compare learner differences in sensemaking with 4 LADs employing different reference frames. *Reference frames* describe the internal and external conditions that help learners interpret their visualized data (Wise et al., 2014). According to Jivet et al., the 3 reference frames used in LADs are achievement, progress, and social comparison with peers (2017). The 4 LADs used in Lim et al. (2019) were self, course, peer, and both course and peer referenced. The LADs were created with secondary data; learners were asked to respond as if the data were their own. Think aloud protocols were employed to probe learners' sensemaking with regard to their affect, planned learning actions, and motivation for learning based on the hypothetical data visualized in their LADs.

Of the four LADs, students focused the most on their own course activities using the peer referenced LAD. Perhaps more interestingly, researchers found that social anxiety was induced with both of the LADs containing peer referenced data. Even when course and peer referenced data were available, participants focused on the peer referenced information that caused their anxiety. The LADs were found to have a negative impact on affect overall, but this could be beneficial if it compelled learners to utilize their learning time more effectively. Based on their findings, Lim et al. suggested that LADs be personalized to the learning objectives with messages connecting course goals, and that learners be provided training and support to aid their sensemaking.

In their sensemaking study Tervakari et al. (2014) gathered the opinions of both learners and educators on 5 different dashboards in an online course using trace data, surveys, and focus groups. The results highlighted disconnects between what learners and educators considered motivating, and how learners actually utilized dashboards in situ. Learners preferred LADs depicting immediately actionable support, rather than those that helped them monitor and evaluate their own performance. Even if they found the

visualizations helpful, learners did not necessarily find them motivating. Of the 64 learners, 1 was very active, 4 were frequently active, and the rest of learners were largely inactive outside of the learning intervention, which required interaction with the LAD. These 5 students often filtered the data according to author, which essentially presented a peer comparison frame of reference. The behaviour of the most active student was markedly different from the others in that they remained active over the entirety of the course. The authors posit that perhaps this learner was motivated by competition. Though the educators involved in the study had no problem detailing multiple ways dashboards could benefit online learners, the learners themselves were less likely to find the dashboards useful or easy to interpret.

Lim et al. (2019) and Tervakari et al. (2014) have divergent opinions on peer-referenced LADs, likely stemming from the contexts of the LAD's use in each of their studies. In the Lim et al. laboratory study, the LADs visualized learners' hypothetical time expenditures spent on learning activities relative to their cohort (Lim et al., 2019). Lim et al. (2019) posited that peer-referenced LADs support a performance goal orientation (Elliot & McGregor, 2001), which leads to a surface-approach to learning. As such, they recommended the avoidance of peer-reference visualizations in the design of LADs. This is understandable, since the LAD did not visualize group learning activities. The Tervakari et al. LAD was part of a learning intervention that included group discussions; in it learners saw their own data visualized alongside that of their peers (Tervakari et al., 2014). The authors reported on learner's actual use of the LAD in a social learning environment, where peer-referencing is expected.

2.5. Pedagogy

LADs have many purposes, and as a result, many pedagogical foundations. Some LA have a stated purpose of simply increasing awareness (Scheffel et al., 2017a), while others prompt reflection (Arnold et al., 2017a; Fritz, 2011; Gibson et al., 2017a; 2017b), or suggest learning resources (Anaya et al., 2016). As we saw with Course Signals, some LADs are predictive, designed to identify and notify learners at-risk of academic failure (Baneres et al., 2019). Previously viewed as tools or techniques, educational technologies

such as LADS are increasingly being designed with pedagogical approaches in mind. Many learning theories exist; the pedagogies in this section are included because they are commonly used in online and blended learning contexts. Cognitivism, constructivism, transformational learning, self and socially regulated learning theories each conceptualize learning differently and often take a slightly different approach to knowledge creation and retention. We explore these pedagogies in this section to better understand the theoretical, philosophical, and ideological underpinnings of LADs, as well as the social and cultural contexts of their use.

2.5.1. Cognitivism, constructivism, transformational learning theory, self and socially regulated learning

Cognitivists describe learning as taking place within the mind; it is an internal process involving motivation and metacognition, abstraction, thought, memory, and reflection (Ally, 2008, p. 33). Applying information processing theory (Miller, 1995) in an online learning context, care must be taken to facilitate higher order information processing, helping learners transfer information from being seen, to working, and then long-term memory. This is a quick process – information that is seen is stored for less than one second and lost if not transferred to working memory in this time (Kalat, 2007). It takes approximately 20 seconds for information in working memory to be stored in long-term memory. Ally offers strategies for the presentation of online learning materials, such that they receive enough attention to be committed to long-term memory (2008). These strategies include chunking (Miller, 1995) to reduce the cognitive load required to process new information, using both intrinsic (learner driven) and extrinsic (performance driven) sources of motivation, the encouragement of metacognitive monitoring during the process of learning, and the provision of ample opportunities for reflection.

Constructivist theory is more externalized than the cognitivist perspective; a core belief of constructivism is that learners construct their own meaning through interaction with social and physical environments and then draw their own conclusions from these experiences (Hung et al., 2004). It prioritizes situated, contextual learning in which learners are active participants. Collaboration and cooperation are crucial (Hooper & Hannafin, 1991; Johnson & Johnson, 1996; Palloff & Pratt, 1999), as learners benefit

from their own experiences and the lived experience of others. New information is scaffolded upon previous learning. Social constructivism takes place in group learning scenarios; the newly constructed meaning is negotiated between participants, with each individual making the learning personally relevant and meaningful (Powell & Kalina, 2009).

Transformational learning theory incorporates aspects of both constructivism and cognitivism, and incorporates authentic real-world problems (Mezirow, 1981; Mezirow, 1991). Also interaction-based, this theory is unique in its multiple phases and types of reflection (Mezirow, 1991). This theory of adult learning describes a ten-step process of how perspectives are changed, through 5 contexts 1) the frame of reference for the learning, 2) the conditions of communication, 3) the learning process, 4) the learners' self-image, 5) the situation encountered during the process of learning (Mezirow, 2000). The multilayered reflection is particularly important to online learning because it helps learners to process information in relevant and meaningful ways (Ally, 2008).

In group-based learning activities knowledge is co-constructed through ongoing social interaction between learners. The learning is situated within not only the individual learner's mind, but also within the context of the group and the learning environment. Social models of learning – including cooperative, collaborative, co-regulated learning (CoRL), and socially shared regulation (SSRL) – emphasize goals, motivation, and social factors (Bandura, 1991; Pintrich, 2000; Schunk, 2001; Zimmerman, 1989, 2001) from both of these perspectives. For example, Bandura's social constructivist theory of learning recognizes the temporal, environmental, socio-economical, and cultural contexts of learning (Bandura, 1991). The model gives credence to measures of academic performance that include resources such as time management, information and help seeking behavior, goal setting, self-motivation, and emotional regulation.

Students who self-regulate their learning do so by monitoring, evaluating and adjusting their behaviors, cognitions, and motivations (Zimmerman, 2012). This self-regulation involves the coordination and control of cognitive and metacognitive thought; it also encompasses the selection and application of goal-directed learning strategy (Duncan & Mckeachie, 2005). It is one of the motivational constructs strongly related to

academic success (Boekaerts & Corno, 2005; Miltiadou & Savenye, 2003; Pintrich & de Groot, 1990; Zimmerman, 2000). A number of studies indicate that self-regulatory skill may be taught and improved with practice (Bembenutty, 2009; Cleary & Zimmerman, 2004; Greene & Azevedo, 2007; Paris & Winograd, 2001; Perels et al., 2009; Perels et al., 2005; Pintrich & de Groot, 1990; Stoeger & Ziegler, 2008).

Today, it is common for student-facing LADs to cite foundations in the self or social regulation of learning (Jivet et al., 2018; Matcha, 2019). Multiple theories of self-regulated learning (SRL) include phases of monitoring, reflection, and control. Bandura describes self-regulation as the way an individual influences their external environment through their self-observation, self-judgment, and self-reaction (Bandura, 1991). Zimmerman's theory of self-regulated learning describes it as a process that "occurs largely from the influence of students' self-generated thoughts, feelings, strategies, and behaviors, which are oriented toward the attainment of goals." (Zimmerman, 2012). This theory includes goal-setting, self-efficacy, and dispositional attributes such as affect and motivation. Pintrich's (2000, 2003) model of SRL stresses the importance of motivation in all SRL phases and similar to Zimmerman (1989, 2000), incorporates goal setting and self-efficacy. Pintrich (2000) states that SRL is "an active, constructive and goal directed process where learners monitor, regulate, and control their cognition, motivation, emotions, and behaviour, guided and constrained by their goals and the contextual features in the environment." Winne and Hadwin's definition of SRL emphasizes its transitory nature, offering a model of self-regulation that presents as a recursive pattern of metacognitive monitoring and feedback (Winne & Hadwin, 1998). Their model is based on events – occurrence, contingency, and patterned contingency (2000) – rather than mental states. The Canadian Consortium for Self-Regulated Learning defines it as a "complex metacognitive and social process that involves adapting thinking, motivation, emotion, and behavior… 21st century skills that extend well beyond academic work to support learning and success in such contexts as: work, social, sport, health, and recreation" (Canadian Consortium for Self-Regulated Learning, n.d.).

2.6. Formative feedback

Initially stated in 1985, Cohen's (1985) assertion that feedback "is one of the most instructionally powerful and least understood features of instructional design" still echoes true. Not to be confused with summative assessment, formative feedback is that which is received during the process of learning. Unlike summative assessments received at the end of an activity, formative feedback can still influence learners' engagement in an ongoing learning activity. Sadler's theory of formative feedback mandates that it must motivate learners to close gaps between their actual and desired performance; this theory is supported by their research studying the effects of formative assessment on the development of expertise (Sadler, 1989). According to Black and William, learning gains triggered by formative assessment are "amongst the largest ever reported for educational interventions" (Black & Wiliam, 1998; Black & Wiliam, 2009). Black & Wiliam's studies indicate that formative assessments must involve: 1) learners engagement with their self-assessment, 2) the provision of feedback by educators that tells learners how they may improve, and 3) is continuously adjusted by educators according to the aforementioned assessments (Black & Wiliam, 1998). The greatest difference between the theories of Sadler and Black and William and is the role of the educator. While valuing the educator's perspective in interactions with learners, Sadler (1989) regards the development of expertise as a self-directed – i.e. learner directed – endeavor that is independent of an educator's involvement.

By aggregating and visualizing learners' performance-related data, LADs may serve as the visual foundation for learners' judgments of learning. Judgments of learning (JOL) are the knowledge estimates that learners make to determine what to study, when the study, and how much to study (Metcalfe & Finn, 2008; B. L. Schwartz, 1994; Thiede et al., 2003). They are part of the metacognitive process that takes place during learning; as such they are instrumental in the monitoring and selection of appropriate learning strategies and resources throughout the learning process. As learners cycle between goal discovery and maintenance, their metacognitions about the status of their learning – especially when there is a discrepancy between desired and actual performance – are answered by these learning judgments. As part of learners' self-assessment, JOL are

instrumental to learners' self-regulation. When JOL are skewed, learners may take inappropriate action, or none at all. Overconfidence in one's abilities (Aghababyan et al., 2017; Alicke & Govorun, 2005; Dunning et al., 2003; Ehrlinger et al., 2008; Kruger & Dunning, 1999) could lead to inaccurate JOL, resulting in unrealistic metacognition about one's learning. Accurate JOL can result in positive self-regulatory learning strategies, like giving preference to difficult course content or that which learners feel they have not learned adequately when studying. Accurate JOL have been associated with strategic studying that led to improved academic performance (Thiede et al., 2003).

In Schwendimann et al.'s review of 55 LAD studies, 31 papers (56%) explicitly mentioned a pedagogical approach (2017). This review was followed shortly by that of Jivet et al., who in their recent review of the integration of learning science in learning dashboard research found self-regulated learning theory to be the most common design foundation, often used to either support awareness or to trigger reflection (Jivet et al., 2017; Jivet et al., 2018). In their review of LA-related SRL research, Viberg et al. found that LA were primarily used to measure SRL, rather than support it (2020). Self-reports are most often used to measure features of learners' self-regulated learning behaviors (ElSayed et al., 2019), using instruments such as questionnaires, interviews, think-aloud protocols and learning diaries. Still there are LADs that have been built to support specific learning theories such as Bloom's Taxonomy(Hu et al., 2017), Engeström's Activity Theory, Dillenbourg and Jermann's concept of social planes (Mejia et al., 2017), Vygotsky's Zone of Proximal Development (Anderson et al., 2001; Mendiburo et al., 2014), and self-regulation and reflective learning (Arnold et al., 2017; Lim et al., 2019). Ruiz et al.'s TEAMQuest attempted to move learners through the 4 phases of the learning analytics process model (Ruiz et al., 2016). Moving students through awareness, reflection, sense making, and impact, TEAMQuest supports students' tracking of their emotions in an attempt to identify emotional patterns that might indicate failed learning. Ruiz et al. posit that awareness of their emotions will positively impact learners' self-reflective processes and educators' teaching strategies (Ruiz et al., 2016). If they do not state a pedagogical foundation, LADs commonly cite their purpose as supplying formative feedback.

Some LADs – usually described as awareness tools – were designed to provide formative feedback. This is a common goal, particularly in literacy development. WiREAD helped learners monitor their reading relating progress with four different visualizations, including a social network (Tan et al., 2016). AcaWriter used formative feedback to help collegiate learners shift their focus from achieving a certain grade to improving their writing (Knight et al., 2020; Tan et al., 2016). Scholar Analytics supported didactic pedagogy in the provision of differentiated instruction and formative feedback with a novel visualization (Montebello et al., 2018). In their research, Gibson et al. (2017) designed a LAD grounded in theories of reflection and reflective writing for university learners, to guide students to reflect more deeply, and to utilize specific language in doing so. It was co-designed with the educators and experts in learning and academic language, before being deployed with learners for validation.

Scheffel et al. (2017) designed a collaborative learning activity widget designed to foster awareness and reflection for learners in a master's degree course. The widget explicitly informed learners about the activities of their group members in a project-based course using radar and bar charts. Learners were meant to reflect upon how their behavior influenced their position in the team, in addition to their course outcomes. The Evaluation Framework for Learning Analytics questionnaire was given to learners twice to gather feedback on the effectiveness of the tool, along with open-ended questionnaires. Students alerted the researchers to the reality of other students "gaming the system," to achieve a higher score, and in their comments, mentioned wanting to see ratings of the quality of their discussion posts in addition to their quantity. The GRAASP social media platform also supported collaborative, inquiry-based science lab activities in multiple ways (Vozniuk et al., 2014). In the evaluation of the GRAASP peer assessment component, the researchers used expert and student reviewers to mimic using "wisdom of the crowd" to achieve consensus on the reviews.

2.7. Design and evaluation

Student-facing LAD research for this dissertation was collected from the Web of Science and the top learning analytics journal and conference – the Journal of Learning

Analytics and the Learning Analytics and Knowledge conference. The Web of Science includes journals such as: Computers and Education, Internet and Higher Education, British Journal of Educational Technology, International Journal of Educational Technology in Higher Education, Educational Technology & and Society, Australasian Journal of Educational Technology, and the International Journal of Computer Supported Collaborative Learning. The initial keyword search combined terms such as education or learning, learners, visual or visualization, dashboard, and analytics. So, for example, a Web of Science search with the terms "learning analytic, education, visualization" yielded 160 papers. Duplicates were removed, then the abstracts of the papers were reviewed to see if they detailed learning analytics built for learners' use. Quite often the phrase "for teachers and learners" was used in the abstracts, but the papers did not detail any student involvement. Papers that did not evaluate the visualizations or dashboards with learners were omitted. This left a total of 29 papers that met the criterion, with another 14 papers that were notable in some regard, such as their design methods. In the inclusion of learners in the design and evaluation processes these studies represent the ideal, rather than the norm. In comparison, only 13% of 94 student-facing LADs included in Bodily and Verbert's 2017 literature review described the design process of the visualizations (2017a).

Many relied on questionnaires, using them to evaluate prototypes (Ahn, 2013) or fully designed and deployed systems (Arnold et al., 2017). Questionnaires and log data were also used to evaluate LADs after deployment (Mouri et al., 2017; Hu et al., 2017); some studies included additional methods (Broos et al., 2018) or planned to (Hu et al., 2017). Interviews were also a well-utilized method of information gathering, sometimes supported with questionnaires (Mendiburo et al., 2014), or questionnaires and trace data (Whitelock et al., 2015). Others involved instruments such as the User Experience Questionnaire (Laugwitz, 2008) or Evaluation Framework for Learning Analytics questionnaire (Scheffel et al., 2017b). The Evaluation Framework for Learning Analytics contained questions such as "This dashboard stimulates me to think about my past learning behavior" and "It is clear to me which data was collected to assemble this dashboard." Though validated instruments such as this one allow for comparison across studies, one of the drawbacks is that unless it relies on open-ended responses, it seeds

learners' minds with the desired responses. The studies primarily described exploratory work (Ho et al., 2016; Mouri et al., 2017) or exploratory analysis followed by an in-situ deployment (Ruiz et al., 2016b). The next largest category of studies described deployments during real learning processes visualizing learners' own data (Whitelock et al., 2015; Arnold et al., 2017; Broos et al., 2018; Hu et al., 2017; Khosravi et al., 2020; Mouri et al., 2017). Only one compared multiple visualization techniques (Ruiz et al., 2016). A few studies detailed iterative design processes (de Quincey et al., 2019; Ruiz et al., 2016) or evaluations involving hundreds (Broos et al., 2018), or even thousands (Arnold et al., 2017) of participants. These massive, often multi-university deployments required that the LADs be used by students, which does not give readers a sense of what learners would do of their own volition. Smaller studies often show wide patterns of interaction with LADs, such as that observed with OpenEssayist (Whitelock et al., 2015), detailed later in this section.

Design

Few of the reviewed studies referenced their design methods. Those that did tended to borrow methods from other disciplines such as User Centered Design (de Quincey et al., 2019), design-based research (Tan et al., 2016). Mendiburo et al.'s work is included here because of its instructional design methods, pedagogical foundation, integration into the learning activity, and extensive, multi-step testing with learners (Mendiburo et al., 2014).

Similar to Wise et al. (2014), Mendiburo et al.'s work had a clear pedagogical foundation and was integrated into the learning activity (2014). Mendiburo et al. wanted to better understand children's interactions with virtual manipulatives in the study of fractions (Mendiburo et al., 2014). The learning activity was built on the theory of the Zone of Proximal Development put forth by Vygotsky and Cole (1978). The long-term goals of the study were to create a system that would aggregate learner information, organizing it according to similar learning trajectories, and that the system would provide actionable instructional recommendations. The learner sample was derived from three sections of a math class. A three-day learning intervention was planned; the researchers led a discussion that introduced virtual manipulatives, demonstrated how to use them, and

then led multiple subsequent practice sessions. A multiple-choice pre- and post-test were administered to be able to compare the accuracy of their responses. From a sample of 41 learners who participated in the intervention, seven students who received low scores on the post-test were chosen to participate in in-depth interviews. The interviews began with general questions, followed by a talk aloud session solving problems from the post test. Using a variety of pedagogical strategies to identify misconceptions or incomplete understandings, the researcher asked opposing questions to determine contributing factors to learners' low performance scores. This information was then used to augment the design of future iterations of the LAD.

WiREAD use a design-based research approach, supplemented by a quasi-experimental design deployed with secondary school students (N=92) (Tan et al., 2016). Tan et al. (2016) evaluated the LAD by surveying learners' perceived ease-of-use, usefulness, and how helpful they felt the dashboard was to their learning or growth. This evaluation was supplemented with qualitative surveys (N=86) and focus groups conducted with of subset of learners (N=30). In the qualitative interviews feedback on the ability of the LAD to motivate the students was mixed; some students said that negative emotions felt as a result of viewing the lab would adversely affected their learning. Further, some students felt that the LAD would be more helpful if it was criterion-based and self-referenced, instead of being norm-referenced to their peers.

Exploratory work

Much of the reported exploratory work was performed *on* learners' data, not with learners themselves. Take for example SINQ, a social media application that supported scientific inquiry-based learning with LA. Ahn et al. (2013) performed a case study with six learners, using log data of their interactions with the application to generate visualizations used by the researchers to explore their learning trajectories. Some work, such as 3DLAV, used data generated from secondary sources to produce a LAD prototype, then had learners rate the visualization on features such as easiness, friendliness, motivation, encouragement, and collaboration (Ho et al., 2016). Evaluations such as these describe proposed, rather than actual use. In this kind of scenario feedback must be received critically, because it remains unknown what learners will do in an actual

learning scenario using their own data. Arguably, a needs assessment should drive or be incorporated with exploratory work, yet few studies included any kind of needs assessment. This is similar to Bodily and Verbert's (2017a) recent literature review, where only 6 of the 94 LADs included any form of user needs assessment. In that review 14 papers stated why they collected the data they did, and 10 performed a usability test of the LAD.

Multifaceted or novel evaluation methods

The following LADs utilized multifaceted, multiple instrument or setting evaluations (Knight et al., 2020; Whitelock et al., 2015), or novel evaluation methods such as paper prototyping with learners (Hillaire et al., 2016), laddering (Hinkle, 2010), and the expert-novice comparisons (Khosravi, 2020).

OpenEssayist is an example of a LAD that learners chose to use voluntarily; as a result, the observed interaction patterns varied significantly. Of the students who used the LAD, the majority accessed it two, three, or four times (11, 8, and 9 students respectively) for short amount of time (Khosravi et al., 2020). The mean session length varied from less than minutes to over two hours; almost half of the users (18 students) interacted with the LAD for less than 10 minutes near the assignment due dates. The interviewed student continued to use the LAD for additional writing assignments after the conclusion of the research study. The LAD did not initially meet his expectations for the type of feedback given; with the repeated use he recognized how the LAD could help him restructure his work. Though essay grades were reported for this research, it was difficult to determine what, if any, relationship existed between performance and the LAD interactions.

The researchers who created the LEA's Box open learning models (OLM) sought to investigate students' motivations for their initial interaction with the OLM, which featured 10 types of visualizations — skill meters, tables, stars, smiley faces, gauges, a histogram, network, radar plot, word cloud, and tree map (Bull et al., 2016). Similar to LAD, open learning models (OLM) visualize learning data directly to learners and have similar pedagogical foundations and goals, such as promoting self-monitoring, planning,

and reflection. The difference is that OLMs often have more data to rely on than LADs, and sometimes the model itself is negotiable.

LEA's Box OLM inputs data from a variety of activity and sources to reflect language learning accomplishment to students (Bull et al., 2016). Students were Italian language learners in a British university, who had a tendency to only utilize formative assessment quizzes close to exam periods, when they would be least beneficial for learning vocabulary. LEA's Box was designed to foster prolonged use, contextualizing learners' progress by visualizing their competency over time. This study was conducted to see how students might interact with the OLM when not required to do so. The researchers explained the purpose for LEA's Box before administering a questionnaire that interrogated learners' initial thoughts about the OLM without having used it. The questionnaire included questions such as which of the 10 visualizations did learners anticipate using, how they anticipated using the visualizations, the features they thought would be included, and their expectations about negotiating their personalized OLM. This particular study reported on student's intentions, to see if they would engage in self-directed formative activity if supported by the OLM. Participants' questionnaire responses were all graded on a five-point Likert like scale. Twenty-two of the participants indicated a tendency to use a mix of both structured and unstructured visualization types. Out of 25 participants, 23 said that they would use the OLM for all four of the stated purposes — comparing levels and topics, planning what to work on next, to think about their competency, and to note their strengths or difficulties (Bull et al., 2016). The same number of participants stated a desire to see the evidence for the OLM values when they disagreed with them, and 19 expected to be able to influence the OLM when they disagreed with it. Surprisingly, an additional 14 individuals wanted to discuss the learner model values even if they agreed with them. Of course, it remains to be seen if learners would indeed use the OLM as anticipated. This is however, information that Bull et al. desired, to know what visualizations students might find beneficial, and how they would anticipate using these kinds of visualizations before actually interacting with the OLM.

From their results, Bull et al. advised that though complex visualizations might be able to indicate multivariate relationships, the inclusion of simple visualizations may allow students to "identify a visualization they can envisage using" (Bull et al., 2016). All

of the participants wanted to use the OLM to plan what they would next work on, and nearly all (24) wanted to compare their levels across topics and to identify their strengths and weaknesses. The authors suggested that this could be seen as participants' recognition of the benefits of OLM, but there is another possibility. In asking this question directly, it could have also clued students into what the desired behaviors were, so students answered accordingly.

AcaWriter was evaluated in multiple settings (Knight et al., 2020). In one of these evaluations the LAD was used as part of the classroom activity that introduce the students to the concept of rhetorical moves in their writing. In this way, the instructor demonstrated its use, and the students were better able to understand the use and relevance of the LAD to their writing practice. To measure its impact on student writing, the researcher surveyed learners' perceived usefulness of the learning design on a five-point Likert scale, with and without AcaWriter feedback. Knight et al. also compared the scores of writing samples from students who did and did not receive this feedback, finding that students in the feedback group had a statistically higher number of rhetorical moves then did the control (2020).

Ruiz et al.'s TEAMQuest supported students' tracking of their emotions; they posited that awareness of their emotions will positively impact learners' self-reflective processes and educators' teaching strategies (Ruiz et al., 2016). Their paper detailed an exploratory analysis followed by an in-situ deployment over the course of two months. A questionnaire was created for TEAMQuest based on learners' responses to the extent that each emotion influenced their learning and their certainty in assessing their emotions, using a six-point Likert scale. Prototype visualizations were created; with traditional graphs such as bubbles, stacked bar charts, boxplots, and an innovative visualization based on small multiple squares. The bubble chart shows the individual students' emotions for each session; the other visualizations compared the emotions of the individual to those of the group. The LAD was evaluated with a three-step process, a satisfaction questionnaire, log data, and learner interviews. The satisfaction questionnaire employed a six-point Likert scale, however the researchers augmented the preference question slightly, asking learners to distribute 20 points among all of the proposed visualizations. The interviews consisted of eight confirmatory questions from the

satisfaction questionnaire, questions such as “did you have problems understanding the graphs?” Learners were asked to complete the emotion questionnaire twice each week, wants to reflect upon the previous week, and again to reflect upon the activities from present activities. The LAD was deployed with learners after their fifth week; at the end of the sixth week learners were asked to complete the satisfaction questionnaire. This kind of deployment — featuring tight coupling with the learning activity, repeated in-class exposure, and a graduated deployment such that learners knew how to use the LAD before being allowed to access it on their own — likely contributed to learners’ opinions about the visualizations. The common graphs and novel visualization all received positive ratings (four or above) on the six-point Likert scale for ease of understanding, emotional awareness, group emotional awareness, self-reflection due to own evolution, and self-reflection due to comparison to the group. Student agreed that tracking emotions could help their learning, but this value statement did not play out in actual use. While 10 out of 15 agreed that awareness of their emotions could influence their learning and that awareness of the group’s emotions could help them reflect on their own, only six out said that they would continue tracking their emotions after the conclusion of the activity. Further, the log data revealed that students only visualized their emotions when prompted to do so in class, even when the LAD was continuously available to them.

In the second phase of the TEAMQuest study the LAD was further integrated into the learning activities, by attaching it to university’s clicker system (Ruiz et al., 2016). This allowed educators to create emotional capture events in which teachers would elicit responses to the emotional questionnaire during times they deem significant to the learning process. The LAD was deployed into compulsory classes, with 97 and 81 enrolled students respectively. In this iteration the stacked bar charts were replaced by bar charts, and the novel visualization was omitted because non-expert students experienced difficulty interpreting it. Box plots and bar graphs were used for all events comparing individuals to the group. Again, log data was compared to learners’ responses to the satisfaction questionnaire. Participation was considerably lower, with 36% and 22% of learners participating in the subsequent study. Ruiz et al. (2016) stated that while the sample was too small to achieve significance, the answers from the sample provided insight into their experience of tracking their emotions. The interpretation of the results of

the study is circumspect however, because causal relationships between learners' emotions and what prompted them could not be identified. Though the researcher's interpretation of learners' individual emotional states were tied to the learning at hand, this interpretation does not take into account personal contexts such as the learners' baseline emotional state or their emotional states outside of the course. Looking at all experiments together, 55% believed it could positively impact behavior. The researchers concluded that the utility of the visualizations was good because the majority of learners agreed that the provided information was interesting, or that it prompted them to reflect on their emotions. However, what aspect of usability the term usability implied was freely chosen, and continued use was something few of the learners intended.

de Quincey et al.'s (2019)research is exemplary in how they included learners in every step of their design. In their research linking collegiate learners' engagement to their reasons motivating them to attend college, de Quincey et al. (2019) employed User Centered Design[3] standards in the co-design of student-facing LADs employing multiple novel visualizations. The initial visualization design was performed in focus groups, followed by deployment, and then contextual interviews with learners who use the LAD over the course of two semesters. The LAD mapped learners' course performance and engagement, predicting outcomes based on learners' self-selected motivations. The visualization was based on the results from a laddering technique (Hinkle, 2010) used during semi structured interviews to identify higher-level motivations such as job prospects, money, or social prestige. Multiple types of visualizations were tested in a focus group; a clear divide was seen between learners who preferred playful metaphors, such as a tree visualization, and those preferring more traditional visualization types. The resultant visualizations were initially individualistic; after focus group feedback a "comparison with peers" visualization was added, along with a visualization of class averages. The peer comparison visualization was simple, depicting data points that were connected by lines, with color differentiating individual from the class average.

[3] https://www.usability.gov/what-and-why/user-centered-design.html

In the contextual interviews, learners were asked to perform a think aloud while using an example of the dashboard to determine learners' understanding of the graphics and the data visualized. Additionally, they asked learners if they trusted the data or their scores, how they would feel if they received a negative report, and the criterion they would like to see used in a comparative visualization. Learners had difficulty understanding the graphics, and the relationship between the graphics and scores depicted. Learners wanted a clear explanation provided for how scores were calculated.

In a subsequent deployment of the LAD, learners were introduced to the dashboard in the first lecture of each module, with time given for questions. Even with this, the researchers found that learners needed more support, so they added interventions in the form of personalized text weekly emails that advised students on how to get more support. They also implemented a "lecturer-in-the-loop" process, such that slides were produced from the visualizations for the lecturer to use in class to prompt discussions about how the LA were being used.

In the next evaluation de Quincey et al. (2019) again used a mixed method approach. This time adding questions to determine the LAD's impact on learners' motivation in addition to usability, as measured with the User Experience Questionnaire (Laugwitz, 2008) and questions similar to the impact, awareness, and reflection portion of the Evaluation Framework for Learning Analytics (Scheffel, 2017). The questionnaire was completed by 35 learners. Only 49% of learners thought that the data were presented in the dashboard accurately reflected their engagement. Engagement with this version of the dashboard was high (between 87% and 89%), but this may be attributed at least in part to the participation of the lecturer and the embedded nature of the visualization in the class activities.

Seven learners completed contextual interviews to assess their metacognition using their own dashboards and emails to inform their responses. They responded to questions such as "do you feel that the learning analytics system had a direct impact on your performance" and "can you describe how the learning analytics system impacted your motivation." Learners said that the visualizations' effect on performance was slightly positive, that it would have more impact on more difficult modules. de Quincey

et al. (2019) noted that during the deployments the benefits of LA were explained to students, yet the overall sign-up rate to use the LADs was only 48%. They took this as an indication that perhaps not all students want to engage with LA. Perhaps not surprisingly, the module with the highest engagement levels was the one in which the in-class visualization was used the most, and the visualization was integrated with the course activities. As compared to engagement levels seen from a more mature LAD at a local university, they found their usage statistics to be promising. de Quincey et al. (2019) asked the undergraduate learners who used their dashboard (N=169) about their feelings of dependability and trust, interpreting the varied results as a need to provide learners with as much information as possible.

Chapter 3.

Visual cognition and perception

The tight coupling between visual perception and cognition is evident in the nearly synonymous meaning of the words understanding and seeing. To visualize is a mental process; a visualization is the tool for this processing, the knowledge discovery, sense making, and insight generation it entails. The design of a LAD is both an art and a science, informed in turns by graphic design, psychology, and human computer interaction.

Perception and cognition are the building blocks for visual cognition. While perception helps us physically process visual information, cognitive skill aids in its integration with prior knowledge, goals, emotion, and attention to make accurate inference with the visualized information before us. Advancements in neuroscience have given insight into the mental processes underlying vision, attention, emotion, and decision-making (Carrasco, 2011; Chatterjee & Vartanian, 2014; Kirk et al., 2009).

Visual processing is the result of visual routines. If one thinks of the brain as a computer, the processing of visual information takes an input, performs calculations, and return some output based on the series of routines enacted. Marr's (2010) prominent visual processing framework does just this, describing the routines as taking place on the computational, algorithmic, and hardware implementation levels. Rensink's (2000) framework separates the visual structure into two parts, a lower and higher-level system, with visual routines taking place in the each. The low-level system processes features from the higher level system; one of the higher level system has two parts, one that requires attention – focusing on objects of interest – and one that operates without attention, processing scene layout and gist.

Aspects of the human visual system have evolved to help us identify objects in our environments. Two unique aspects of the human perceptual system are statistical learning and the ability to make statistical summary representations. Statistical summary representations are the rapid averages of visual objects made when viewing items of

differing size or position; these averages may be improved with feedback (Fan et al., 2016). Examples of the features that may be extracted include mean size, weight, or position. These processes proceed in the absence of intention, awareness, or perceptual cues. Numerosity — at least in terms of averages and means — is a feature of the statistical summary representation. The perception of numerosity is often reinforced with additional marks, closure, or contours. For example, in a series of studies by Zhao and Yu, the perception of numerosity was influenced to a statistically significant degree by the regularity of perceptual features (2016). The numerosity of a group of identically coloured objects was consistently underestimated when the objects were placed in close proximity.

Statistical learning is the identification of visual regularity in an environment, such as the spatial configuration of a forest, a kitchen, or a street scene. It is a rapid process of discerning regular patterns in a space, such as objects that tend to be seen together when encountered in real life. Statistical learning differs significantly from learned semantic guides. It happens without attention, or even intent. This makes evolutionary sense. If a person can quickly detect what belongs in an environment, then the identification of danger irregularities is also quick. It contributes to why novelty draws our attention. In daily life we see similar objects occupying similar spaces, in often-similar configurations. This experience further hones this contextual aspect of perception. It is why we see a familiar scene in a sketch, though the image is essentially blocks of alternating tones. Statistical learning lends structure to a multifaceted visualization; the adoption of a familiar hierarchy lessens the overall complexity of the scene.

Fortunately, visual hierarchy is one of the most mature areas of visualization research. Unfortunately, results on regularity in visual search are mixed. Some studies claim that the eye is drawn to homogeneous rather than heterogeneous displays (Nowakowska et al., 2017), while others indicate that induced regularity produces no benefit and perhaps may even slow performance (Vaskevich and Luria, 2018). It is possible that the difference exhibited in search study results may be attributed to an unstated occurrence, such as the roles novelty or aesthetic attraction play, or if due to

textural differences produced by a stark contrast between background and foreground objects (De Vries et al., 2013).

3.1. Visual cognition

There is no agreement in the literature on how visual cognition transpires, though multiple theories exist to explain the phenomenon (Deller et al., 2007; Esterman, 2000; Healey & Enns, 2012; Healey et al., 1996; Kristjánsson & Egeth, 2020; Pomerantz & Portillo, 2018; Treisman & Souther, 1985; Wolfe, 2010). Treisman's feature integration theory (Treisman & Gelade, 1980; Treisman & Souther, 1985) is adopted for this work for multiple reasons. At the time it was proposed it combined contemporary research in cognitive psychology, visual psychophysics, and neuroscience to describe the role of attention in the identification of target objects from distractors. Though multiple alternatives and modifications have been suggested – including by Triesman herself (Treisman, 2006) – this theory has held up for 40 years, because and in spite of its ability to neatly combine aspects of interdisciplinary research. The initial theory (Treisman & Gelade, 1980) is built on the premises that 1) the perceptual process is hierarchical, that 2) humans are only able to visually encode a finite number of object features, and that 3) object features are detected automatically, and in parallel (Kristjánsson & Egeth, 2020). From there, individual object features are combined as they are processed – for example, combining a single object's spatial location, color, and shape in one's mind – and focal attention is what integrates these features and binds them together. This theory provided a framework for research in visual search and attention; the role of attention in visual perception research may be attributed to it (Kristjánsson, 2015; Kristjánsson & Egeth, 2020; Noudoost et al., 2010).

Though some subsequent empirical evidence conflicts with some of the originally proposed aspects of the feature integration theory, the most significant aspect of the feature integration theory to the present studies – that the ability of an object to capture attention is context dependent – remains true. Subsequent studies have supported the idea that preattentive features do influence one's ability to detect objects in a visual field, however attention is required to make sense of the objects. Further, to make sense of their

use in a visualization, these preattentive features must be integrated with additional stimuli in the visual field. The stability of the individual aspects of the model is more important for low-level psychophysics research – for example, a low-level task like the identification of a single object from a single type of distractor – than the series of studies in this dissertation because they are conducted with real, multifaceted tasks during the process of learning. The feature integration theory maintains that attention, an internal mechanism, can both filter and boost the selection of target objects from distractors in a visual field.

3.2. Visualization and graph comprehension

Reliant on both cognitive and visual perception, the strength of a LAD's visualization is in its ability to inform, and sometimes, persuade. Visualizations enhance important data, either reducing or omitting redundancies. It has been shown that compared to text, information visualized in graphical form may support different reasoning processes (Stenning & Oberlander, 1995) or make different aspects of the depicted data explicit (Larkin & Simon, 1987). While graphs support a static view designed for data consumption, interactive visualizations often provide a framework with which individuals seek the answers to self-generated questions. This interactivity, the main difference between graphs and visualizations, allow visualizations to support repeated search, hypothesis, and insight generation. Much of the research on graph comprehension informs visualization design, in part because users' expectations for graphs and visualizations are much the same. Further, visualization prototypes that are not interactive are identical to graphs.

Information visualization research has provided methods for organizing the display of quantitative information in graphs (Tufte, 2001; Ware, 2012), multivariate data (Hagh-Shenas et al., 2007), using glyphs (Demiralp et al., 2014), optimizing search and interaction with data (Shneiderman, 2002; Shneiderman & Plaisant, 2010), and assessment performed in lab and in situ (Isenberg et al., 2008; Liiv, 2010; Sedlmair et al., 2012; van Wijk, 2013; Zuk, Schlesier et al., 2006). Tufte (1990), who first championed the efficiency of ink usage by reducing data to ink ratio, began the culture of design

simplification that continues to prioritize the benefits of cognitive ease over aesthetic appeal, complexity, or novelty. This perspective is often at odds with proponents of aesthetics (Berlyne, 1970; Lim et al., 2007; Miniukovich & De Angeli, 2015) such as Bateman (2010), who extols the virtues of aesthetics to sensemaking with visualizations.

The affective response to aesthetic appeal is elevated arousal and pleasure, which is intimately tied to learning processes (Chatterjee & Vartanian, 2014; Ishizu & Zeki, 2013; Jacobsen et al., 2006; Tractinsky et al., 2000). Beauty fires attentional aspects of the brain that would otherwise not be engaged (Chatterjee & Vartanian, 2014), and people attend to beautiful things longer. Though influenced by culture and personal preference, beauty has an evolutionary basis that has resulted in commonalities in visual preference and response being seen across humankind (Falk & Balling, 2010; Hagerhall et al., 2004; Mealey & Theis, 1995). Though beauty may increase the complexity of a visualization, it may aid in the identification and selection of important information due to popout (Gillian & Sorensen, 2009) and framing effects (Sun et al., 2011). The use of visual metaphors may also aid sensemaking.

Gillian and Sorensen (2009) sought to directly compare the data-ink maximization maxim to Treisman and Gelade's (1980) findings indicating that visual search is improved by differentiation between a target and its background. In the study they compared participants' target feature selection accuracy, using bar and line graphs with and without background embellishment (Gillian & Sorensen, 2009). Accuracy was highest with the embellished graphs, prompting the researchers to reason that embellishment aided the popout effect, which made target features easier to locate (Treisman & Gelade, 1980).

With implications for visualizations meant to aid decision-making, Sun et al.'s research on graph framing effects suggests that manipulating graphical representations can have framing effects on the decisions made with those graphs. In their study, they showed that physical distance affected perceived numerical distance in both coordinate-based (line and bar graphs) and sector based (such as a pie chart) graphs; this has implications for visual depictions used to make preferential choices (Sun et al., 2011).

Though aesthetic appeal is commonly thought to be subjective, there are landscape properties with near universal appeal. Humans possess an innate preference for savanna-like settings, a preference modified through experience and enculturation with age (Falk & Balling, 2010). The "beauty" of this kind of landscape had an evolutionary purpose, because they were often indicative of resource rich, safe spaces for people to inhabit (Orians & Heerwagen, 1992). This preference for landscapes may have been an important factor in early humans' ability to perceive complexity (Gauvrit et al., 2014).

Graphical literacy, the ability to understand and make inferences based on graphically presented information (Shah et al., 2009), varies across a heterogeneous population. Take for example Shah and Freedman's study comparing the inferences generated with different graph types (Shah & Freedman, 2011). Using line graphs and bar charts, they found that users expected line graphs to depict interactions and bar graphs to display categorical differences. Shah and Freedman hypothesized that those with high graph literacy would be able to make inferences in all conditions based on their greater ability to mentally manipulate the graphs, however this hypothesis was unsupported. The researchers were surprised to find that participants with high graph literacy were only able to make correct inferences when the data was familiar and presented in a format supporting that type of inference. By using an open-ended question format, Shah and Freedman were able to identify differences in the responses generated by participants with low and high graph literacy, namely that those with low graph literacy tended to review them on a superficial level. Study results indicate that inference generation was not supported by graph familiarity, the participants' graphical literacy, or format alone.

The variability of graph literacy across the population and the varied effects their designs have on users means that data visualization methods should be selected based on the user, the task type, the users' expectations, and insights users are attempting to discover. Their effectiveness is evaluated in a number of ways, including error rates and time on task, eyetracking (Kurzhals et al., 2014), self-reports and user studies (Liu et al., 2014), and preference or satisfaction measures (Bangor et al., 2008). Error rates and time on task are traditional measurements used in information visualization, along with self-reports, user preference and satisfaction rates (Bangor et al., 2008). Verbal protocols may provide insight into individual differences in graph comprehension, while multiple choice

assessments are easily distributed to high numbers of participants. Subjective measures such as self-report or satisfaction rates can be subject to poor memory recall or self-deception biases (Yannakakis & Martínez, 2015) however, whereas physiological measures such as eye tracking require strict environmental controls. A final issue in the measurement of graph literacy is that they may be domain specific.

A recently created instrument, the Visualization Literacy Assessment Test (Lee et al., 2016), looks promising. It features 53 multiple-choice and true-false items and takes approximately 23 minutes to complete. The instrument was evaluated for content validity by domain experts, and iteratively tested with MTurkers with varying degrees of education In addition to the measurement of visual literacy, early research suggests that it may also be used a reliable predictive measure of an individual's ability to learn from an unfamiliar visualization (Lee et al., 2016).

3.3. Spatial memory

Visual spatial working memory is the ability to interpret and recall spatial information. This ability has been linked with academic and career success, especially for individuals in science, engineering, mathematics, technology, and design occupations. Intelligence tests such as those associated with Carroll's Human Cognitive Abilities (Carroll, 2009) include broad spectrum tasks that measure different aspects of spatial ability. Spatial ability is an amalgamation of three skills – low-spatial perception, visualization, and mental rotation.

The difficulty in measuring visual spatial memory most often lies in the lack of complexity of the tasks. Tasks often interrogate short, rather than long-term memory, and are limited by the number of factors that may be included in each task. Spatial scan tasks involve the recall of sequences of locations or objects within a space. Types of spatial memory tasks include reproducing a single location or configuration, a sequence of spatial locations or patterns (Claessen et al., 2014). The stimulus for these tasks is either auditory or visual and may possibly be made more difficult by task variables that are visual, spatial, or manual. Together these studies attempt to identify how spatial information is mentally encoded — whether it is influenced by the type of stimulus, the

task at hand, or a combination thereof. Eye-tracking studies support the idea that eye movements aid memory for sequences of spatial locations, though only in the case of focused rehearsal of the locations in the order presented (Gerard et al., 2009). This, in turn, supports studies in spatial attention — here meaning the act of attending to spatial information — that suggest that position is more memorable than serial order. Smyth and Scholey (1996) shows that in the absence of eye movement, shifting attention interrupts a spatial span task, without any eye movement. Shifts of locational spatial attention have been seen to interrupt spatial sequence, but not verbal sequential tasks (Gerard et al., 2009, Tremblay et al., 2006b).

Mental rotation tasks are also prominent (Hawes et al., 2015). Participants are usually asked to compare multiple stimuli to determine if they are the same after a number of rotations. A single match for multiple stimuli that have been rotated along any of their axes. In these mental rotation activities, response times have been seen to be an almost linear function of the angle of rotation (Sheppard & Podgorny, 1978). A number of studies indicate that the parietal cortex is activated and engaged longer as the amount of rotation increases, but it doesn't answer how the stimulus is processed in the brain (Heil et al., 1996; Rösler, 1995). The Revised Purdue Spatial Visualization Tests measures spatial visualization and rotation (Guay, 1976; Yoon, 2011). In the PSVT-R participants match symmetric and non-symmetric 3D objects with their rotations. As the test contains few words, it does not hinder individuals who may not speak the language of the test. The original version of the test included three subtests, with twelve items each. The revised version consists of 30 question items and must be completed within twenty minutes.

Kemps (2001) found that performance was influenced by the structure of the path, sequence of positions, repetition, and the absence of crossings. The positioning was also important, whether it be symmetrical, vertical, horizontal, or at 45° angles, which seems to indicate that the visual features of the path itself aid in memory performance.

In their summary of the effects of spatial information on visual working memory, Zimmer and Liesefeld (2011) cite special attention, eye movement trajectory, and the configurable and temporal aspects of spatial information as contributing factors to spatial

working memory. That said, they conclude that no single underlying mechanism exists to support or inhibit spatial working memory. They note that this does not account of individual differences, task demands, the time available to encode the information, or the number of items therein.

3.4. Gist

Visual understanding is achieved through two cognitive processes — overall understanding or gist, and deliberate analytical thought. Gist describes the global information one remembers about an image, including basic features, surfaces, objects, and the spatial relationships between them. Humans exhibit iconic memory for the gist of scenes, memory that is high capacity, short in duration, and precategorical (Dick, 1974; Sperling, 1960). The result of rapid visual processing (Simons & Levin, 1997), gist recognition is defined at different levels of detail across studies of attention, change blindness, object recognition, or long-term memory. In as little as 150ms one may surmise the gist, or semantic nature of an image; a behavioral response may be provoked in 250ms (Macé et al., 2005; Owsley, 2013; Rousselet et al., 2003; Thorpe et al., 1996; VanRullen & Thorpe, 2016). Ranging from the summary categorization to object-level detail, the definition of gist is related to the task being performed with it and is dependent on the relationship to the phenomena being studied.

To measure gist, comparisons are typically made by generating or finding sets of images of constant gist, then varying their graphical features in some way. To manipulate gist between images, the degree of change in the images' gist must first be identified. Expert and naïve raters have both been used to detect and quantify gist change between images, each with a unique set of benefits and drawbacks (Sampanes et al., 2008; van Montfort, 2007). A different group of individuals would then be used to verify the descriptions made by the previous group. It is good practice to analyze the fit of the gist descriptions again before testing them with the target population. Perceptual changes to a scene do not always change the viewers' understanding of the scene, so it is important to note the kinds of features or feature sets that would completely change an individuals' interpretation of gist (Sampanes et al., 2008).

In a series of 3 change blindness studies Sampanes et al. (2008) tested the speed at which participants recognized when a scene changed if the gist also changed. They compared the detection of changes in image pairs with differing gist with changes in image pairs with constant gist – as defined by participants in 4-5 words – reasoning that if the gist is automatically encoded when viewed, it should be consistent between 2 scenes of different images. This would make the detection of the change slower. Their findings indicated that changes in gist were detected faster than changes that did not modify the scene's gist, though change detection proved difficult if several features of a scene were changed. Sampanes et al. cited work by Ryan and Schwartz (1956) – comparing the perception of gist for photographs, shaded line drawings, line drawings, and cartoons from least to most detailed – as foundational to their own study.

In terms of gist, participants were most successful with the most abstract image, the cartoon, because it emphasized global properties and omitted irrelevant details. In another change blindness study, Tseng and Bridgeman tested the idea that gist is automatically encoded, by varying perceptual features with and without changing the gist of the image. They posited that perceptual changes not involving gist would be suppressed (Tseng & Bridgeman, 2010). Using natural scenes as the control, they compared two experimental versions of the same scenes. Changes that affected gist were more rapidly detected than perceptual changes that did not, leading the authors to surmise that gist is automatically encoded.

The colloquial understanding of gist — as the episodic interpretation of concepts' meanings, relationship, or inherent patterns — is also an informational construct with implications for decision-making (Brainerd & Reyna, 1990). Compared to verbatim information, gist is more memorable (Brainerd & Reyna, 1990; Reyna & Brainerd, 1995; Reyna & Kiernan, 1994). In situations of information overload, it has been suggested that decisions improve with psychological distance. Here distance refers to removing an object from the present time, space, or social distancing. In their comparisons of decisions made under information overload, Fukukura et al. (2013) found that the psychological distance induced with gist – it could be spatial, temporal, or an abstraction of the task at hand – tended to mitigate information overload and improve decision making. The authors offer psychological distance as an intervention for improving

decision-making, stating that the induced psychological distance mimicked the actual temporal distance created by the passage of time, a distance that seemingly increases the accessibility of gist memory. They reason that this distancing led participants to better organize the pertinent data features. The authors summarized that their research indicates that the exploitation of gist memory through psychological distancing could be used to improve decision-making.

In the following studies we define gist as a synthesis, the summative understanding of a visualization. It is similar to the term visual immediacy (Karabeg & Akkøk, 2004), used in HCI to describe the process of understanding of visualization "at a glance" (Culén, 2014). When shown to learners for a short amount of time, gist provides the first impression a learner has of their performance data. As the precursor to judgments of learning, gist—and by extension, the accuracy of these learning judgments—is likely subject to the individual differences that mediate the interpretation of visualizations.

Chapter 4.

Description of experiments

4.1. Learning activity, participants, and data

Learners, particularly those learning online, often have difficulty accurately assessing their learning. As an online educator, I have a unique understanding of how LADs could be used to support online learners' self-regulatory learning. This research was undertaken to determine how learners interpret LADs and what they do with this information, toward the ultimate goal of designing LADs that learners will be able to successfully utilize as part of their self-regulatory learning strategy. A mix of quantitative and qualitative methods were used to identify why, when, and how learners interacted with LAD to guide their learning. Following this section, an overview of each of the experiments detailed in this dissertation is provided. The methods utilized in each experiment are included within the associated chapter for that experiment.

Learning online is largely self-directed, and to be successful, a high degree of self-regulatory skill is often required (Dabbagh & Kitsantas, 2004; Hartley, 2001; Schunk & Zimmerman, 1998). As learners' interactions with student-facing LADs are unassisted, it follows that learners were the primary stakeholders for this inquiry. Each experiment touched upon at least one aspect of the learner experience with LADs. The learning context was selected to align with ongoing research in our lab, which meant focusing on blended, rather than fully online learners. Rather than visualizing cumulative learning progress the LADs were designed to support a single learning activity, the small group discussion, within courses offered from the same department. This homogeneity allowed the comparison of learners' LAD interactions jointly and individually.

The small group discussions used in these studies varied slightly in duration and size, lasting between 7 to 10 days, with 4-6 participants. The discussions were graded learning activities that required learners to demonstrate domain knowledge through a social learning activity, i.e. constructing knowledge as a group. In an ideal discussion, learners question their thinking, integrate new knowledge gleaned from the learning

resources and peers, and form new conclusions that they are able to articulate within the discussion thread. They are able to manage the timing, tone, and number of their group interactions, such that they are engaged and engaging in ways that create a transformative learning experience. Online discussions also follow unvoiced social norms, such as the four maxims of the co-operative principle of communication (Grice, 1975) that together, improve the efficiency of communication.

Of the four maxims, the maxim of manner is perhaps the most nuanced. It requires clear communication that is not ambiguous and is properly suited to the language level of the listeners. The maxim of quantity necessitates that no more is said than the conversation requires. The maxim of quality states that only factual information is shared, and the maximum of relation requires that contributions to the conversation are relevant.

The learning activity was supported by instructions, rubrics, and the LADs. LAD feedback may inform learners' decision-making during the process of learning, such as what learning resources to read or revisit, what domain-specific language to use, and what information should or should not be included in their future discussion posts. These judgments of learning may also determine how and with whom learners choose to interact.

Trace data from learning management systems (LMS) is commonly used to observe the behaviors of learners, including their interactions with LADs and in the learning activity. Message counts, timings, and quality were automatically captured from the LMS. This trace data was supported by survey and qualitative interviews. A mixed method approach gave me the ability to use learners' qualitative accounts of their experience to explain and contextualize the quantitative data obtained from trace data from the learning activity. This exemplified a top down and bottom-up approach – from the bottom we had the minutiae of clicks performed in the LMS during the learning activity and from the top, we had insight into learners metacognitions about their learning as they interacted with the LADs. Individual semi-structured interviews were conducted with learners to establish the context for their learning behaviors, to give insight into their mental models of their performance, and how these models in turn, influenced future performance. Crowdsourcing was added in experiments 3 and 4 to be able to compare the

gist assessments of learners to laypersons. Retrospective cued recall methods were used in the semi-structured interviews to garner richer feedback during the interviews (Pätsch et al., 2014; Eger et al., 2007; Frith & Harcourt, 2007; Harper, 2010).

4.2. Pilot study

A convenience sample of learners (N=22) answered very simple performance related, multiple-choice questions with 3 informationally equivalent visualization stimuli — a bar chart, a heat map, and a landscape-based LAD. The tasks were simple performance estimates, the kind of estimations that learners make to determine their performance at a glance, over time, or in relationship to their peers. After the passage of approximately 28 minutes, learners were asked to summarize the overall gist of the three LADs that they used at the outset of the study. Task accuracy and time on task were measured, along with aspects of individual difference that we thought would be associated with high task accuracy or visualization preference. These aspects of individual difference were represented by learners' spatial acuity, cognitive reflexivity, and numeracy, measured both objectively and subjectively. Participants also rated the visualizations on their aesthetic appeal and perceived usefulness. We hypothesized that quantifiable differences in task and gist accuracy would be seen between visualization types, and that learners would attend to the landscape visualization longer because of its novelty and aesthetic appeal. Further we hypothesized that factors of individual difference would mediate performance with the different types of LAD, and that high numeracy, spatial acuity, and cognitive reflexivity would all positively influence task accuracy.

4.3. Exploratory study

Conducted in two phases, the exploratory study employed a mixed-method approach. In the first phase of the study, semi-structured interviews were conducted with learners shortly after they used one of two LADs during a 7-10 day small group discussion activity. Trace data was captured and used to create snapshots of the LAD that participants saw during the discussion activity. Using retrospective cued recall methods

with these snapshots, we explored how learners used the LADs to make judgments of learning, and how they used the formative feedback from the LADs to metacognitively monitor and reflect upon their learning. After the interviews we showed participants the other types of LADs they could have used, to see if they preferred either LAD type. We also collected individual difference data on participants' cognitive reflexivity, spatial acuity, objective and subjective numeracy, to see if these factors were related to learners' performance with or preference to the LADs. In the second phase of the study, we explored learners' initial impressions of the LADs to better understand why learners choose to interact with LADs. Participants performed a forced choice ranking of eight LAD prototypes based on their perceived utility and aesthetic appeal, before and after performing cognitive walk-throughs with them. Learners were then asked to provide qualitative feedback on their reasoning.

4.4. Experiment 3

After a 30 second exposure to three LADs displaying abstract, natural scene-based visualizations, participants were asked to describe all that they understood from the perspectives of the fictitious students highlighted in the LAD. The accuracy and descriptiveness of these gist assessments were compared between learners and laypeople, here represented by MTurkers. We sought to determine which LAD prompted the most accurate or descriptive assessments of gist, the LAD that prompted the highest feature recall, and any mental models either participant group associated with any of the LADs. The LADs were created using secondary learning data; they depicted 7-10 day discussions, similar to the real discussion activity conducted in the previous study. This experiment was administered completely online; codes from this analysis were used in the subsequent study.

4.5. Experiment 4

Experiment 4 extended the work of the previous study by repeating the measurement of gist between learner and MTurker populations. Again, the three LADs represented three different levels of complexity; this time the LAD types were chosen

based on their usefulness in facilitating estimations of proportion. The study was undertaken to see if learners produced more accurate or complete gist descriptions than MTurkers, and if this varied due to numeracy or visualization type.

4.6. Experiment 5

Experiment 5 was conducted to evaluate how LADs shaped learners' mental models of their performance, and to see if these mental models were persistent throughout the learning activity. Learners participated in a graded, asynchronous small group discussion activity for 7 days using a LAD designed for this experiment. Soon after the conclusion of the activity, learners participated in in-depth semi-structured interviews utilizing retrospective cued recall methods and recreations of the LADs they saw during the learning activity. To better understand the factors that shape the conceptualization of gist, the interviews interrogated learners' goals, motivations, and self-concept of their performance. Trace data and interview data were analyzed to comprehensively address both the phenomena of interest and the contextual factors that shaped it. Each learners' LAD interactions served as an exemplar of how learners think with LADs, shedding light upon how learners create and sustain mental models about their performance using LADs.

Chapter 5.

Experiment 1 - Pilot study

5.1. Introduction

The first step in the development of an empirically validated model of learners' visual cognition with LADs was to understand how learners perceived and utilized different kinds of visualizations. This pilot study was undertaken to determine if, given informationally equivalent visualizations, learners exhibited differences in task accuracy or gist assessments made with them. Learners were asked to perform common visualization tasks, and then after the passage of some time, to make gist assessments of the visualizations that they use previously. Self-report information was collected on learners' goals and numeracy, to see if a relationship could be identified between these factors of individual difference and learners' performance or preference with a particular visualization type. Time on task measurements were collected to see if learners attended to any single visualization type longer than the others, and if this indicated a relationship with accuracy. At the end of the study participants were asked to rank the visualizations according to perceived usefulness and aesthetic appeal. The research questions addressed in this study were:

- RQ1: Given 3 informationally equivalent LAD visualizations, do learners perform differently on an immediate or delayed summarization task?
- RQ2: Given 3 informationally equivalent visualizations, do learners exhibit a preference for a certain type of visualization?
- RQ3: Do learners attend to one visualization type longer than the others, and is the amount of time on task correlated to the accuracy of their visual analyses?
- RQ4: Are learners able to remember the overall gist of the visualizations, and do they remember one type of visualization better than the others?

We hypothesized that quantifiable differences in task and gist accuracy would be seen between visualization types, and that learners would attend to the landscape visualization longer than the other two visualizations given its novelty and aesthetic appeal. We believed that relationships would be identified between factors of individual difference — namely spatial reasoning ability, numeracy, and cognitive impulsivity — that would mediate performance with the different visualization types. Specifically that 1) high numeracy would be associated with accurate task performance and gist assessments, 2) a similar relationship would be seen between poor numeracy and low accuracy, 3) participants with high spatial reasoning ability would perform well with all visualization types, 4) participants with low spatial reasoning ability would perform best with the least complex visualization, the bar chart, and that 5) that factors of individual difference would mediate performance such that high associations would be related to high accuracy, and low dispositional scores would be associated with low accuracy.

5.2. Methods

This pilot study was conducted with a convenience sample (Lewis-Beck et al., 2003) of collegiate students (N = 22) to determine if, according to visualization type, relationships could be identified between performance, preference, retrospective gist assessment, and factors of individual difference.

5.2.1. Participants

Participants were students currently enrolled in face to face or blended courses at a Canadian university, solicited through direct email and social media postings. As an incentive, the participants were given a $10 gift card for their participation.

5.2.2. LAV stimuli

Three different visualization stimuli were used to perform the learning tasks (Figure 2). Bar charts are perhaps the most familiar visualization type and are often used for comparison type judgments (Simkin & Hastie, 2012). Often used for making proportional judgments (Heer et al., 2010), heatmaps were used in this study to extend

the work being done concurrently in my lab (Beheshitha et al., 2015b). The landscape visualization was utilized due to the underlying metaphor of growth, the commonality of landscapes in everyday life, and their aesthetic appeal (Bateman et al., 2010; Howley, 2011). The embellishment of the landscape visualization would perhaps make it more memorable (Hullman et al., 2011).

All of the visualizations were mocked up using Tableau[4]. The landscape visualization was finished using Adobe Photoshop[5]. The same color scheme was used in each visualization to denote post quality. Color scheme was used to mitigate any individual differences in recall based on the influence of color. The data visualized was based on anonymized discussion data from an ongoing study, lending ecological validity to this experiment. In it, fictitious student messages were displayed according to the time the message was posted and quality of the message. Message quality was determined using latent semantic analysis in a previous study (Beheshithaet al., 2015b), and both positive and negative learning paths were visualized.

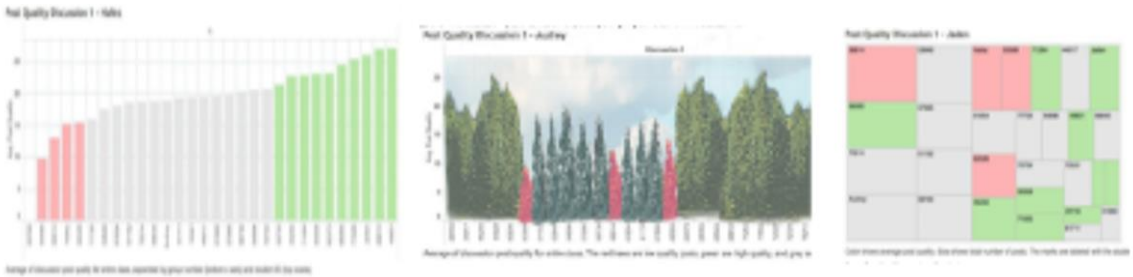

Figure 2. Pilot study visualizations

5.2.3. Additional study instruments

Study participants completed the following self-reports: Cognitive Reflection Test (CRT) (Toplak et al., 2011), the Purdue Spatial Visualization Test (PSVT) (Guay, 1976), the Berlin Numeracy Test (BNT) (Cokely et al., 2012), and the Subjective Numeracy Scale (SNS) (Zikmund-Fisher, 2008; Zikmund-Fisher et al., 2007). Together these tests

[4] Tableau is business intelligence software (tableau.com)

[5] Adobe Photoshop is image editing software (adobe.com/products/photoshop)

represented individual differences in the factors hypothesized to influence a person's performance with visualizations.

In the discipline of decision science, numeracy is often measured objectively and subjectively. The BNT objectively measures statistical numeracy, a subset of numeracy that describes decision-making when faced with risk (Cokely et al., 2014). The instrument is designed to perform best with college educated populations; 30 to 50% of the population used to validate the test held graduate degrees (Ghazal & Cokely, 2014). If a sample population is expected to be highly numerate, use of the BNT can prevent ceiling effects seen with other tests of numeracy (Ghazal & Cokely, 2014).

The Subjective Numeracy Scale SNS is a subjective measurement of numerical aptitude and preference (Zikmund-Fisher, 2008; Zikmund-Fisher et al., 2007). The SNS is strongly correlated with objective numeracy, and is often preferred by participants because it does not require calculation (Fagerlin et al., 2007; Peters, 2012; Zikmund-Fisher et al., 2007). Aside from being highly correlated with other tests of numeracy (Fagerlin et al., 2007), the primary advantage of the SNS is that it does not require an objective test of numeracy, which can create anxiety for some study populations. Both the BNT and SNS took approximately five minutes to complete. They were completed in this study to determine if one or the other had a higher relationship to task accuracy. The CRT measures cognitive impulsivity, the ability to resist reporting the first response that comes to mind when problem-solving (Frederick, 2005).The test is a well-known, and routinely used measure of an individuals' ability or disposition (Liberali et al., 2012; Pennycook et al., 2015; Sinayev & Peters, 2015; Toplak et al., 2011; Toplak et al., 2013; Welsh et al., 2013).

The PSVT-R objectively measures spatial processing ability, by asking participants to match objects with their rotated counterparts (Bodner & Guay, 1997; Guay, 1976). There is evidence to suggest that individuals with higher spatial reasoning ability perform well with spatial reasoning tasks, such as those seen demonstrated with graph literacy. The test is utilized in science, engineering, technology, and mathematics education research for this reason (Bodner & Guay, 1997; Branoff, 2009; Guay, 1976; Yoon, 2011). The revised 15 minute version of the test (Yoon, 2011) was included in this

study to see if high spatial reasoning ability correlated with learners' ability to correctly utilize visualizations.

5.2.4. Procedure

Participants reviewed three visualizations from the perspective of three fictitious students, answering three multiple-choice questions per visualization. The study instructions were as follows:

> "This is the first of three different visualizations you will use in this section to answer questions about a single student's performance in the discussion threads. This is the visualization that Student Name sees. You are answering questions as if you are Student Name reviewing your online discussion post results after participating in your classes' online discussions."

Participants were asked questions such as "for what discussion was Hafez' post quality above average," and "in discussion one Aubrey's post quality is better than what percentage of the class," and "over the course of the class discussions, Jaden's post quality is steadily doing what?" They were also asked true/false questions, such as "the majority of Jaden's posts were high quality." Answering these questions involved search, comparison, making inferences and gist assessments, covering the range of common tasks involved in visual analyses. When each question page was accessed, a timer began. It ended when the participant pushed the selection button for the question. This was measured in seconds and represented time on task.

The task related questions were administered first, followed by self-report questionnaires on numeracy, cognitive impulsivity, and spatial reasoning ability. Together the questionnaires took approximately 28 minutes to complete. Similar to Stone et al. (Stone et al., 2015), participants provided gist estimates, recalling aspects of the visualizations without being allowed to review them. At the end of the study participants ranked the visualizations on their perceived usefulness and aesthetic appeal. The entire experiment was administered online, using Fluid Surveys[6].

[6] Fluid Surveys is now Survey Monkey, a survey administration platform.

There were three tasks per visualization type and one gist estimate per visualization type, for a total of 12 tasks. Time on task was measured by the number of seconds spent on the single webpage containing all three tasks. The means of the correctly answered tasks according visualization type were compared using a one-way analysis of variance (ANOVA). Correlations were measured between the factors of individual difference, and between the factors of individual difference and the total number of accurate tasks completed. Linear regressions will be performed with three factors – using the CRT, PSVT-R, and either the SNS or BNT, to determine the best fit.

5.3. Results

Of the 47 participants who began the study, only 22 (13 female, 9 male) completed it. The majority of the dropout was seen approximately halfway through, during the BNT portion of the study. The majority of those who completed the study were master's degree students (N=11), followed by doctoral (5), and undergraduate (2) students. As part of the demographic questionnaire, participants were asked to rate their proficiency expressing themselves in English. This was included to see if language proficiency would add additional challenge to the experiment for participants. The majority (N=18) rated themselves as either excellent or good on this question. When asked how well they estimated the amount of work it takes to achieve their desired grade, again the majority of participants (N=17) rated themselves as either excellent or good.

There were not enough participants to perform a linear regression as planned. Using a medium effect size of 0.15 according to Cohen's (1988) criterion for a linear multiple regression, alpha error probability = .05 and error probability power = 0.80, the minimum projected sample size would need to be approximately N = 77.

Table 1. Pilot results

Correct tasks by visualization (out of 3 each)	Mean (SD)	Factors of individual difference	Mean (SD)
Bar chart	1.86 (1.13)	Average SNS	2.05 (0.64)
Heatmap	2.09 (1.02)	CRT (0-3)	2.27 (1.03)
Landscape	2.14 (1.04)	BNT (0-4)	1.5 (0.86)
Gist	1.27 (0.94)	PSVT-R (0-25, normally 30)	9.23 (3.46)

A statistical power analysis was performed with G*Power software (Faul et al., 2007) to estimate the required sample size for measuring differences in task accuracy with the 3 visualization types. Using a medium effect size of 0.67 according to Cohen's (1988) criteria for a repeated measures ANOVA, alpha = .05 and power = 0.90, the projected sample size would need to be approximately N = 33. For power = 0.80, the projected sample size would be N = 27. The analysis was performed since the number of participants (N=22) was close to the projected sample size.

The difference in task accuracy between visualization types (Table 1) as measured with a oneway ANOVA ($F(2,65) = 0.42$, $p = 0.66$) suggested that performance with the bar chart (M = 1.86, SD =1.12), heat map (M = 2.09, SD =1.02), and landscape visualizations (M = 2.14, SD 1.04) was functionally equivalent., Calculated using G*Power, with a medium effect size, the post-hoc power estimate of these results was power = 0.75. Similarly, the percentage of accurate gist responses across visualization types for the bar chart (45%), heat map (41%), and landscape visualizations (41%) indicates little difference between the gist estimates made with these visualizations. Concerning RQ1 and RQ4, neither learners' task accuracy nor gist response varied according to visualization type.

On average, participants spent much more time reviewing the bar chart (M = 237 sec., SD = 173 sec.) than the heatmap (M = 142 sec., SD = 123 sec.) or landscape visualizations (M = 100 sec., SD = 78 sec.). There was a significant difference between the amount of time spent on all three as determined by one-way ANOVA ($F(2,65) = 6.17$,

p = 0.04). This time difference was not related to task accuracy. With regard to RQ3, while learners did attend to the bar chart visualization longer than the others, this was not correlated to task accuracy.

Only two relationships were identified between the factors of individual difference measured. The total number of accurate visualization tasks performed (M = 6.09, SD = 2.51) and the Subjective Numeracy Scale score (M = 2.05, SD = 0.64) were correlated, r(22) = -0.45, p = 0.04). There was also a correlation between the BNT numeracy scores (M = 1.5, SD = 0.86) and the SNS scores (M = 2.05, SD = 0.64), r(22) = -0.44, p = 0.04).

After completing the tasks, questionnaires, and gist assessments, participants were asked to rank each visualization according to their perception of its usefulness and aesthetic appeal (Table 2). The bar chart was ranked first in both usefulness and appeal. Addressing RQ2, participants preferred the bar chart and heat map over the landscape visualization.

Table 2. Pilot study visualization rankings by count

	Bar chart		**Heat map**		**Landscape**	
Ranking	Usefulness	Appeal	Usefulness	Appeal	Usefulness	Appeal
1st	11	11	10	5	1	6
2nd	10	8	5	12	7	2
3rd	1	3	7	5	14	14

5.4. Discussion

The study didn't have enough participants to perform the linear regression as planned, and the comparison according to visualization type was not adequately powered. Still, several lessons were learned during this pilot study that impacted subsequent studies. In planning this study we assumed that a convenience sample would be appropriate, given that the population sampled consisted of collegiate students, which is

our target population (Check & Schutt, 2018). In retrospect however, this convenience sample may not be generalizable to our target population because the majority of the participants were pursuing master's degrees. The high English proficiency and ability to accurately determine the amount of effort required to achieve their academic goals is expected of graduate students, who have chosen to pursue advanced degrees. This may not necessarily be the case with undergraduates in their first or second year of study. It is also not clear if the visual analyses of graduate students are similar to those of undergraduates. It may be posited that as they have more experience using data, graduate students may have higher levels of proficiency performing visual analyses.

Static dashboards were used rather than interactive ones to mitigate the effects of interaction and navigation on the time on task measurements. The number of task trials was limited to three tasks per visualization, to be able to facilitate the later gist assessment, similar to Stone et al. (Stone et al., 2015). The tasks chosen for this study were common to the analysis of all kinds of graphs and visualizations. According to Lee et al., retrieving, filtering, and computing values are considered low-level task that can then be combined into higher level visual analyses (Lee et al., 2006). Though the type of tasks used in this study reflect the visual analyses learners do during the process of learning, they were greatly simplified due to the visualizations being static and the low number of trials.

Had there been more trials of task type, perhaps relationships between mean accuracy and time on task could have been identified. Though preliminary, it was notable that even though participants spent a significantly longer amount of time reviewing the bar chart their performance with that visualization type was no better than with the other visualization types. This seems counterintuitive; it seems as though the familiarity of the bar chart and novelty of the landscape visualization would both contribute to participants spending more time reviewing the landscape visualization and less time reviewing the bar chart. With more participants and task trials, we could determine if a relationship exists between time on task and accuracy according to visualization type.

Minute differences in accuracy were seen across the visualization types that made the accuracy of the tasks achieved with them functionally equivalent. The results across

visualization types for time on task and accuracy point to the importance of concisely defining the term performance. Indeed, participants did perform differently with the different types of visualizations. They spent more time with the bar chart even though the accuracy of their visual analyses was equivalent across visualization types. If performance is defined by the accuracy of the task estimates performed with the visualizations, then performance was equivalent, however the same tasks were likely achieved in different ways based on the visualization type. Our results could be interpreted to mean that it took more time to use the bar charts than the other visualizations to achieve the same result, i.e. equivalent accuracy in the visualization tasks. It would seem that given its familiarity, less time would be used or needed with the bar chart visualization. The results could also simply reflect a preference for bar charts (Edwards et al., 2006; Fortin et al., 2001; Marshall et al., 2004), and that participants chose to attend to their preferred visualization type longer. Without participant feedback and/or an alternate way to measure this, there is no way of knowing which interpretation was correct. Participants' ratings of usefulness and aesthetic value provided some context for these results. Participants clearly indicated a preference for the bar chart visualization, but the accuracy of their estimates was not markedly different when using them.

We anticipated display issues based on the type of browser used so to plan for future studies, information on the browser type was collected. Google Chrome was used most often with 17 choosing it, followed by 3 using Safari, and 2 using Firefox. An equal number of participants used Macintosh and Windows operating systems on desktops, and 2 participants used mobile devices. After the experiment concluded we learned that part of the landscape visualization was cut off, and many people did not see the scrollbar underneath the visualization. Some browsers automatically hide the scroll button unless one clicks on it directly or hovers the mouse over the area. Attempts will be made to lock screen widths in future studies so this does not happen again. There were also display issues with the PSVT; five responses were removed from the results. It is important to pilot studies to discover these very issues.

The inclusion of the Berlin Numeracy Test had the greatest impact on the study, as most of the participants who withdrew did so during this test. It is likely that the objective nature of the questionnaire discouraged some participants, since they were

required to make calculations to complete it. Without participant feedback we do not know if they were poorly motivated to continue, or the inclusion of this test signalled to them that the remainder of the experiment would require more effort than they were willing to expend. Though we are not sure what motivated the high dropout rate, it is possible that participants preferred the subjective test of numeracy over an objective one (Fagerlin et al., 2007). The construct is correlated with other tests of numeracy enough to fulfill our purposes (McNaughton et al., 2015; Zikmund-Fisher et al., 2007). What's more, the second dimension of it, the part that questions participants about their preference for the use of numbers as opposed to text in their everyday lives, may be extended to their preference for numerical visualizations of their performance data.

Chapter 6.

Experiment 2 - Exploratory study

6.1. Introduction

LADs provide formative feedback to learners by highlighting the effectiveness of their academic strategies. In this experiment we interviewed learners to better understand how and when they perform visual interrogations of LADs to monitor their progress, make learning judgments, and metacognitively reflect upon their self-regulatory strategies. Recognizing evidence suggesting that motivation varies according to individual differences (Beheshtiha, 2015, Beheshitha et al., 2015a; Beheshitha et al., 2015b), we collected auxiliary data on individual differences we hypothesized relevant, namely numeracy, cognitive reflexivity, and spatial acuity. In the first phase of this study we sought to learn:

- RQ1: Did the LAD influence learners' behaviors, and if so, how?
- RQ2: When controlled for numeracy, cognitive reflexivity, and spatial acuity, is there an effect of visualization type on posting behavior or learners' patterns of engagement?
- RQ3: Do learners have a demonstrated preference for visualization type, and is it based on their numeracy, cognitive reflexivity, or spatial acuity?

It was hypothesized that 1) learners would employ different learning strategies using the LADs based on factors of individual difference, 2) that numeracy, cognitive reflexivity, and message quality would be positively correlated, that 3) learners with low numeracy and cognitive reflexivity would exhibit a preference for the minimalist LAD prototypes, and that 4) given both LAD types, learners with high spatial acuity, high numeracy, and cognitive reflexivity would prefer the more abstract heatmap visualization.

The first phase of the study gathered feedback from real users about how and when the LADs were *actually* used to make learning judgments during the process of learning. The second phase of the study explored the persistence of learners' initial

impressions of the utility and aesthetics of LAD prototypes, to better understand why learners may choose to interact with LAD. The primary research questions for the second phase of the study were:

- **RQ1**: Do learners' initial impressions of a LAD's utility and aesthetics persist or change after interacting with them?
- **RQ2**: Are learners' prototype preferences correlated with any of the factors of individual difference?

From a design standpoint, it would be advantageous to know why a user chooses to engage with an interface, as this may contribute to its ongoing use.

6.2. Methods

The semi-structured interviews took place shortly after a discussion learning activity in which learners used one of two types of LADs. Using trace data from the activity, we used retrospective cued recall methods (Pätsch et al., 2014; Eger et al., 2007; (Frith & Harcourt, 2007; Harper, 2010) to discover how learners interpreted the LADs, how and when learners made learning judgments with them.

Learners performed a forced-choice ranking of 8 new LAD prototypes before and after exposure to these LADs to see if 1) their preferences were persistent, and 2) if these preferences were associated with any of the factors of individual difference. To replicate the exposure learners would normally get from performing tasks with a visualization interface, we performed animated cognitive walk-throughs (Mahatody et al., 2007) with the unfinished prototypes, using sequences of wireframes to show learners how they would perform common tasks.

6.2.1. Participants

Recruitment information was sent to instructors of first- or second-year undergraduate courses offered at a Canadian university. Two types of LAV were deployed in a single discussion in each of the four participating blended courses. The discussions took place in small groups of 4-6 students; learners were required to

contribute at least 4 cohesive messages to the discussion within its 7–10-day duration. Learners who participated in the discussion and who had accessed the LADs at least once were invited to participate in the semi-structured interviews. Interviewees were compensated with $15 or an equivalent gift.

6.2.2. LAD stimuli

The two LADs used in the first phase of the study were designed by Beheshtiha (2015) as part of her ongoing research; both compared learners' posts to those of their peers. They were accessed by clicking on a link in the discussion thread that then brought up the LADs on a separate page. Both updated every five minutes in real time. The top performers LAD visualized the number of the learner's posts compared to their top performing peers; the keyword heatmap visualized the quality of learners' posted messages against those of the entire class (Figure 3).

The keyword heat map visualized a grid of key concepts identified by the course instructor before the learning activity. The learner's keywords were presented on the left, and the class average was presented on the right side of the screen. The three levels of color utilized in each side of the keyword heat map indicate the three possible quality ratings. The quality ratings visualized in the keyword heatmap were evaluated with Latent Semantic Analysis (LSA), based on the coherence of messages employing key concepts. Coherence measures the relatedness of the sentences and paragraphs in a discussion post (Foltz et al., 1998). The post quality thresholds were based on previous cognitive presence research (Garrison et al., 2001) and learning analytics studies undertaken in our lab (Beheshitha et al., 2015a; Beheshitha et al., 2015b; Beheshtiha, 2015).

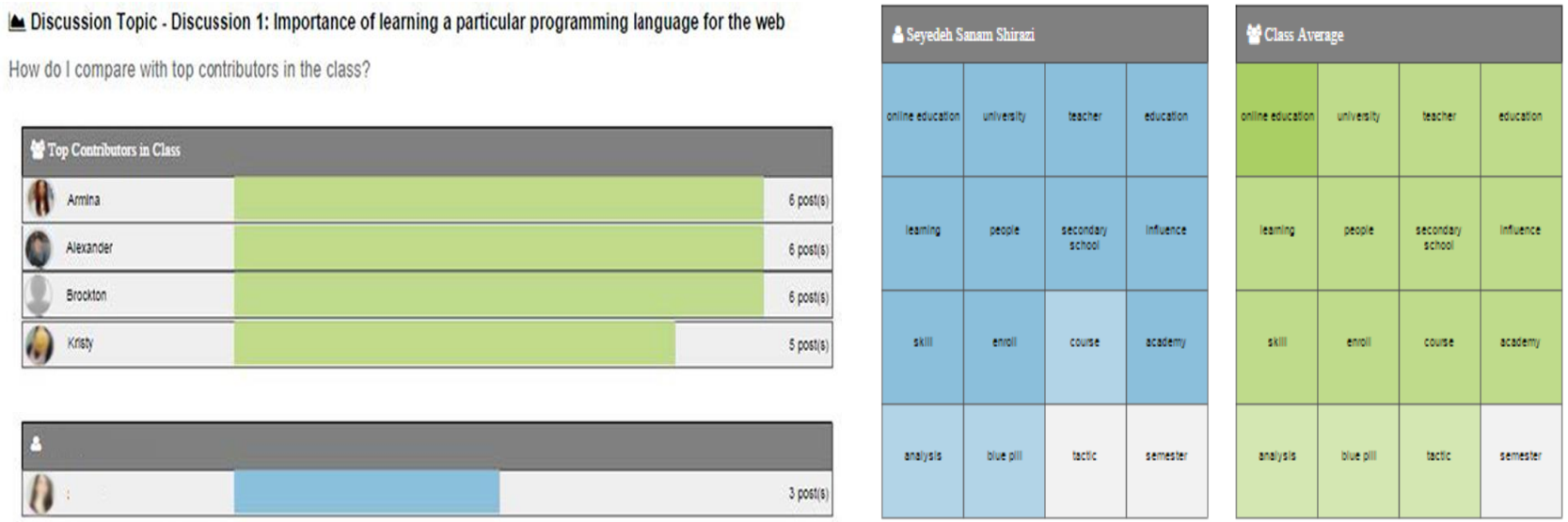

Figure 3. **Exp. 2 top contributors LAD (left) showed the number of posts from the top 5 contributors to the discussion. Keyword heatmap LAD (right) compared learners' average message coherence to class average.**

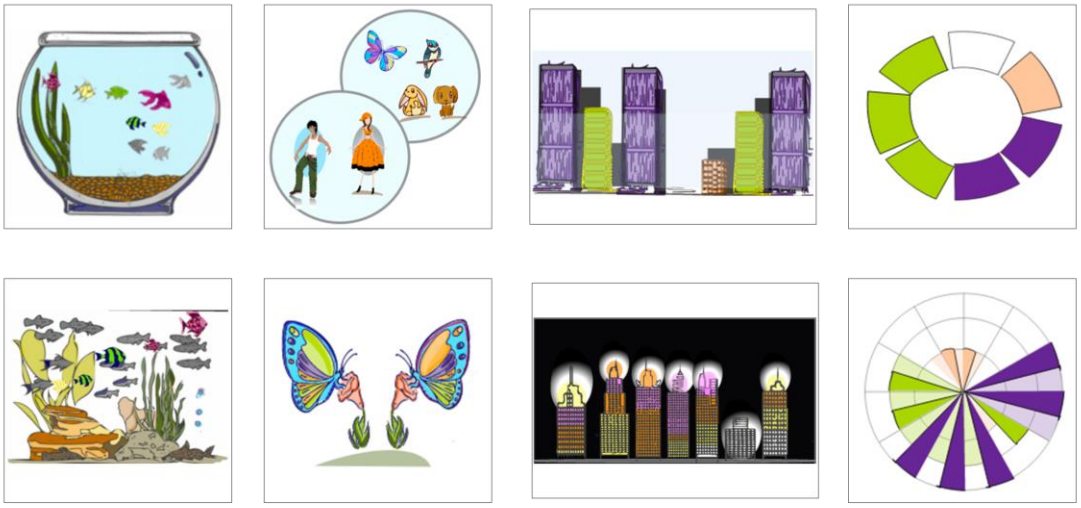

Figure 4. **Exp. 2 proposed individual LAD prototypes (top, right to left) were a polar graph or "flower", buildings, avatars, fish bowl. Proposed comparison visualizations (bottom, right to left) were a bouquet, cityscape, butterflies, fish tank.**

Eight LAD prototypes were introduced in the second phase of the study (Figure 4). All of the prototypes employed gamification as a means of depicting post coherence. The term gamification is used herein to describe the "the use of game design elements in non-game contexts" (Deterding, 2011). The rewards and incentives presented by game elements positively impact intrinsic and extrinsic motivation (Richter et al., 2015), which is why they have been employed in educational environments to motivate students to engage with learning content (Xiao et al., 2018; Krause et al., 2015). The leaderboard is a common application of gamification; leaderboards visually rank participant performance, allowing for direct comparison between individual members of the group. This kind of visualization has been found to foster competition, which in educational contexts, has not been universally well-received by learners (Domínguez et al., 2013). In this study the game-like elements visualized peer contributions in aggregate and did not name individuals, to avoid the induction of feelings of direct competition. Avatars are game design element that offer a way to visualize a personalized representation of students within the learning environment (Krause et al., 2015; de Quincy et al., 2019). In this study the human and animal avatars' state reflected changes in the coherence level of the learners' messages.

Produced in Photoshop and animated using Adobe XD[7], static screenshots of the LADs were initially presented to learners without explanation. Cognitive walkthroughs were performed with the animated version of the LADs that would change when elements within the LADs were clicked on. The LADs shown in Figure 4 are most to least abstract from left to right; the top row represents the single person's visualized results and the bottom row measures the individual against others. Though the interaction patterns differed between the LAD prototypes, they were informationally equivalent.

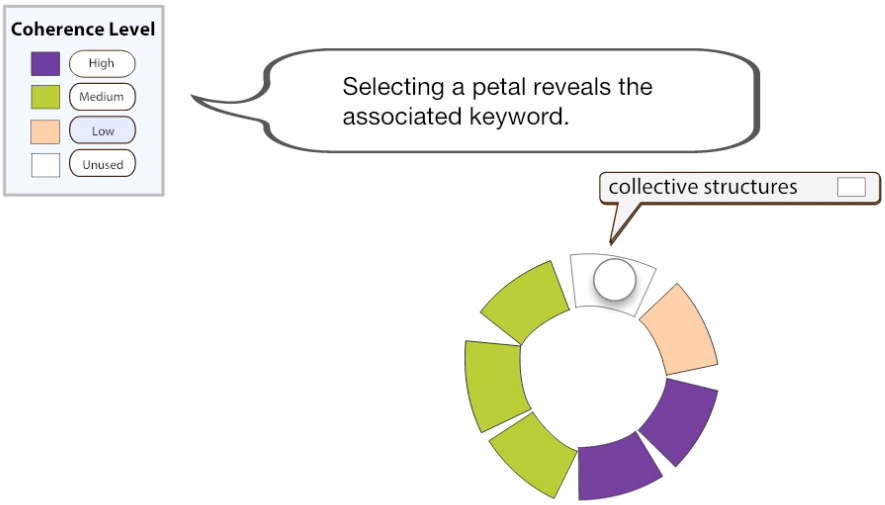

Figure 5. Exp. 2 single flower LAD prototype from cognitive walkthrough indicating that collective structures was an unused key phrase.

The single flower (see Figure 5), city (see Figure 7) and fishbowl LADs (see Figure 8) worked in much the same way. Clicking on petals turned the keywords on and off. Clicking on the coherence level buttons in the legend highlighted all of the petals with the corresponding coherence level. In the city LAD (see Figure 7), the high coherence button was selected, so the three keywords used by the participant with a high coherence level were *abstraction*, *technology*, and *performance*. The fishbowl LAD showed keywords associated with medium and high coherence levels. The flower used only color to indicate coherence. The city and fishbowl used color and height. Line fill and color represented coherence in the avatar LAD (see Figure 9).

[7] Adobe XD Prototyping software https://www.adobe.com/ca/products/xd/details.html

Figure 6. Example walkthrough screenshots with single flower LAD.

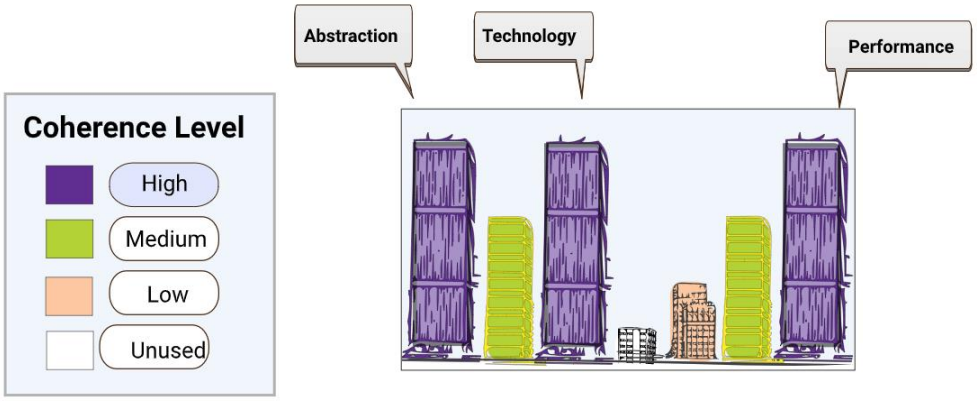

Figure 7. Exp. 2 single city LAD prototype from cognitive walkthrough indicating that message posts with high coherence used 3 keywords.

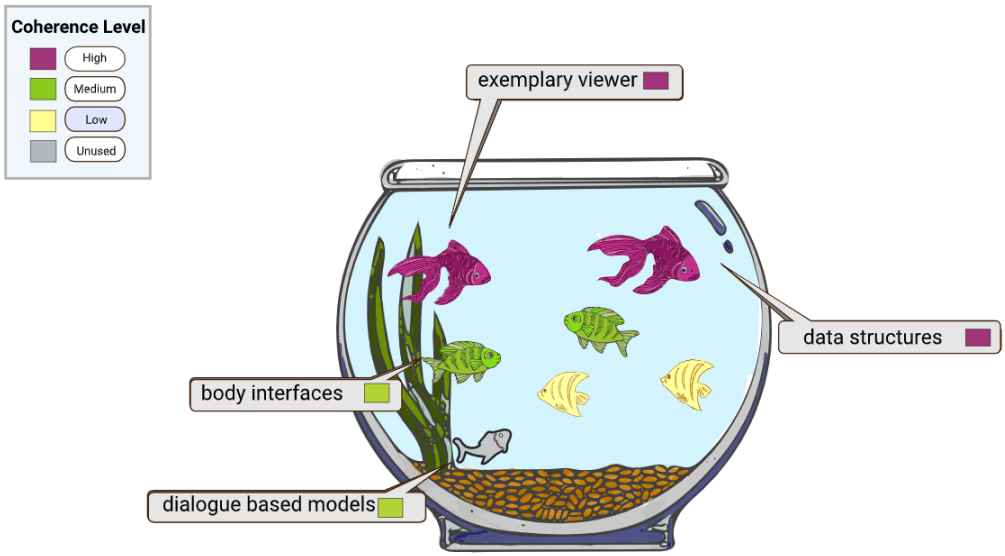

Figure 8. Exp. 2 fishbowl LAD prototype from cognitive walkthrough indicating message posts with medium and high coherence.

The avatar-based LAD (see Figure 9) represented the average coherence level of all of the learners' discussion posts. The participant could choose from three animals or three human avatars. At the outset, when no messages had been posted, the avatar would be represented by a partially complete, colorless line drawing. A low coherence rating would show a partially developed avatar with small elements of color. A medium coherence rating would display a fully complete outline and partially colored avatar. A high overall coherence level would be represented by a full-color, fully outlined avatar.

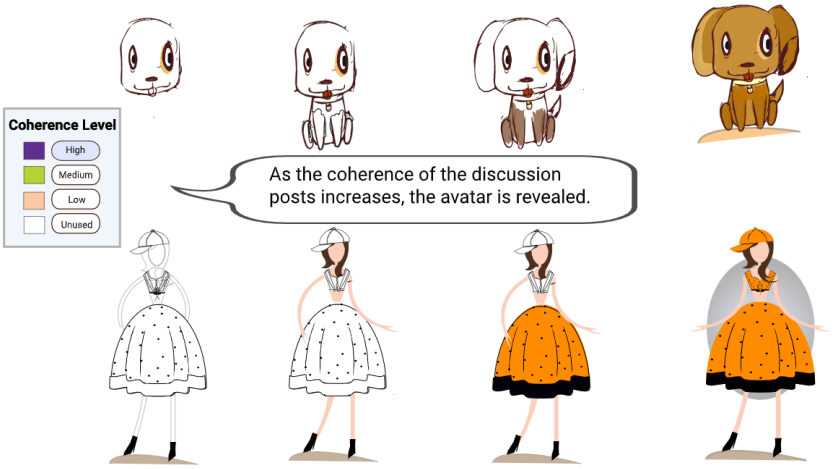

Figure 9. **Exp. 2 four coherence levels of single avatar LAD prototype. Not pictured – additional decorative items continued high coherence "earned" by the avatar.**

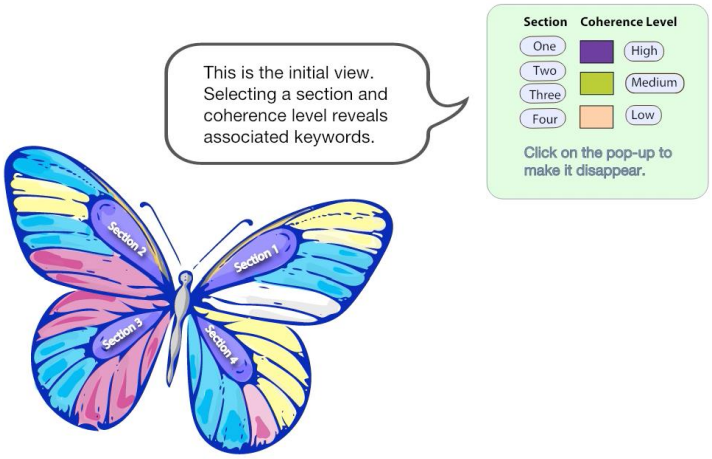

Figure 10. **Exp. 2 single butterfly LAD prototype**

The single butterfly LAD (Figure 10) was arranged differently from the other avatars in that it was arranged to show keywords that were associated with four different sections. This would be used in the learning activity that, for example, might have multiple sections or assigned readings. The visualization showed all of the sections in aggregate, offering the user the opportunity to turn the sections on and off individually or as a group, to reveal associated keywords.

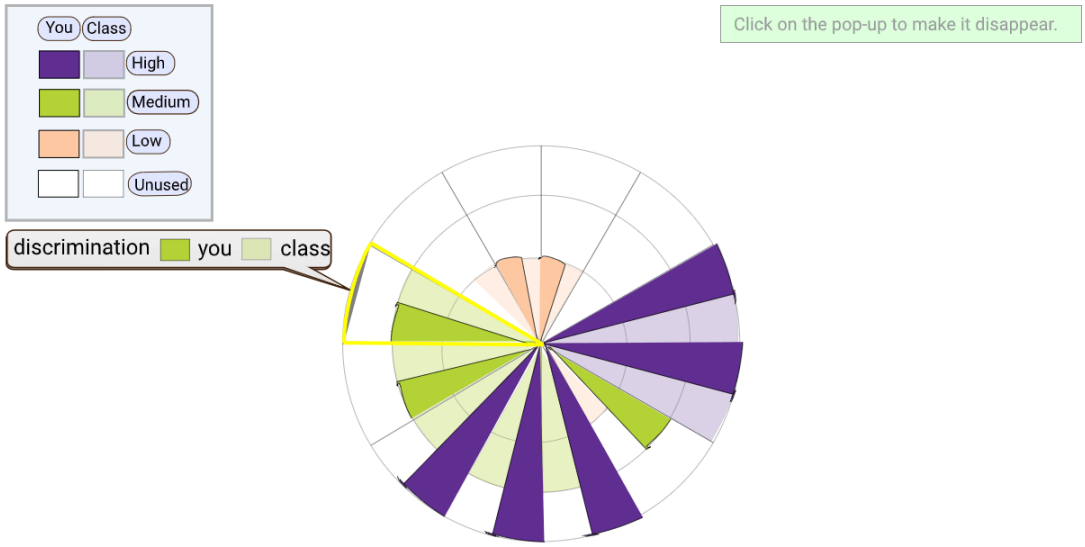

Figure 11. Exp. 2 big flower LAD prototype comparing the coherence of the keywords used by the learner to the class.

Similar to the single flower, each petal of the big flower (Figure 11) represented a keyword. The learner's coherence was on the left side of the petal, and the class was on the right. Clicking on each petal highlighted the section and displayed the keyword.

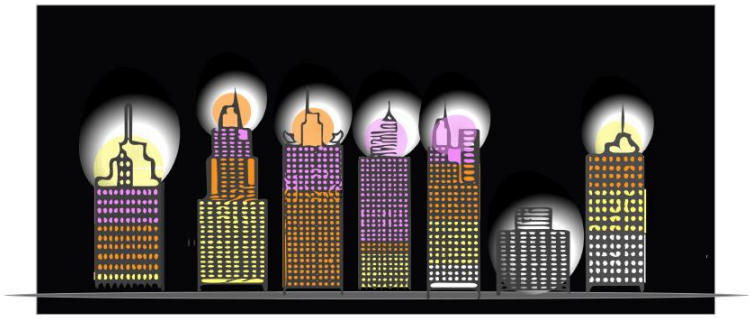

Figure 12. Exp. 2 cityscape LAD prototype.

The cityscape visualization (Figure 12) displayed the class average against that of the individual learner. Each building represented an individual keyword. The window represented each student in the class using that keyword; the color of the window represented the coherence level of the messages utilizing that keyword. The lights at the

top of the building represented the individual learners' use of each keyword. In this way, individuals could compare their coherence to that of the class.

6.2.3. Additional study instruments

Study participants were surveyed about their academic achievement goals, numeracy, cognitive reflexivity, and spatial acuity. Study participants completed the following self-reports: the Achievement Goal Orientation (Elliot, 1999; Elliot & McGregor, 2001; Elliot et al., 2005; Elliot et al., 2011), the Cognitive Reflection Test (CRT) (Frederick, 2005), the Berlin Numeracy Test (BNT) (Cokely et al., 2012), the Subjective Numeracy Scale (SNS) (Fagerlin et al., 2007), and the Purdue Spatial Visualization Test (PSVT-R) (Bodner & Guay, 1997; Guay, 1976). Together, these tests were meant to represent individual differences in the factors hypothesized to influence a person's performance with LAD visualizations. The demographic portion of the survey also included questions regarding familiarity with learning management systems (LMS) and technology, their approach to studying, and their previous experience with visualizations.

The Achievement Goal Orientation (AGO) instrument describes learning orientations, the underlying motivation for learners' performance of academic, achievement-based tasks (Elliot, 1999; Elliot & McGregor, 2001; Elliot et al., 2005; Elliot et al., 2011). The 2x2 achievement goal framework presents the goals as dichotomous, with either mastery or performance-based orientations. A mastery orientation indicates a desire for task mastery or the development of competence, while a performance-based orientation indicates goals based on the demonstration of normative competence. The goals are further described by their valence, with approach having a positive valence and avoidance having a negative valence. Thus, learners are either approaching success or avoiding failure.

For example, learners who are motivated to perform better than their peers are described by the performance approach orientation, while those who are motivated to avoid poor performance in comparison to their peers are described by the performance avoidance orientation (Elliot& Church, 1997). Similarly, learners with our performance

approach orientation want to have the highest mark in the class while those with performance avoidance orientations tried to avoid getting the lowest marks. The four possible categories resulting from the AGO are mastery approach, mastery avoidance, performance approach, or performance avoidance.

Two tests of numeracy were used, one objective and one subjective, to see if either was correlated to the other proposed factors of individual difference. The eight item Subjective Numeracy Scale (SNS) is a self- assessment with questions split into two groups, numerical aptitude and the preference for the numerical presentation of information (Fagerlin et al., 2007; McNaughton et al., 2011). It is correlated with objective tests of numeracy (Fagerlin et al., 2007; Peters, 2012), and takes 5 minutes to complete. The four item Berlin Numeracy Test (BNT) objectively measures statistical numeracy (Cokely et al., 2012), a subset of numeracy that specifically describes the accuracy of decision-making in the face of risk. The BNT also takes 5 minutes to complete.

The CRT measures cognitive impulsivity and could reflect one's tendency toward impulsivity when making decisions (Frederick, 2005). The 3 short self-report questions measure an individual's ability or disposition to resist using the response that first comes to mind when solving a problem (Frederick, 2005).

The PSVT-R is a subtest of the PSVT that objectively measures spatial processing ability, by asking participants to match objects with their rotated counterparts (Bodner & Guay, 1997; Guay, 1976). The test is routinely utilized in science, engineering, technology, and mathematics education research due to the strong predictive validity of the test in these disciplines (Bodner & Guay, 1997; Branoff, 2009; Guay, 1976; Yoon, 2011).

6.2.4. Procedure

In the first phase of this study, trace data was collected from online small group learning discussions that took place during four different blended university courses. All the courses were geared towards first- or second-year undergraduates at the same university, in the same department. Across the discussion activities of the participating

classes, learners (N=178) were assigned to one of two LADs, with remaining learners in each course serving as the control. In courses 1 and 2, one third of participants saw the visualizations. A crossover design was used in courses 3 and 4; one quarter of participating learners in each of two discussions was assigned either the top contributor or quality visualizations. LAD assignments were random; the course instructors assigned students to their discussion groups. This aspect of the design was determined by a separate ongoing study in our lab (Beheshitha, Hatala, Gašević, & Joksimović, 2015b). The interviewees (N=32) for this study were purposefully sampled from the learners who used the LAD at least once, with at least one participant from each discussion.

This study borrows from retrospective think aloud (Pätsch et al., 2014), retrospective cued recall (Eger et al., 2007), and photo elicitation (Frith & Harcourt, 2007; Harper, 2010), retrospective research methods that are each used to improve the accuracy of participants' memories and to help avoid the production of false memories. In retrospective think aloud, participants perform a series of tasks while being recorded. Often no more than 2 hours later, participants perform a think-aloud verbalizing their thoughts while watching the recording of their previous performance (Van den Haak et al., 2007). The recording acts as a visual cue to help participants recall the steps taken to complete their tasks. The major benefit of this HCI method is that the actual task execution takes place in a natural manner. Similarly, retrospective cued recall uses some form of queuing to aid participant memory. In the case of Eger et al's usability study (2007), eye tracking results were used to cue the think aloud. This study also borrows from photo elicitation, an established research method in the social sciences, that uses found or created images from the researcher or the participant (Carter & Mankoff, 2005; Frith & Harcourt, 2007; Harper, 2010) to stimulate memories, guiding interviews toward the production of richer descriptions. The combination or hybridization of these retrospective research methods is not uncommon. For example, in lieu of a diary study, House et al. combined digital visualizations and photo elicitation to study the social uses of camera phone images (Van House, 2006; Van House et al., 2005; Van House et al., 2004). Retrospective studies fall victim to bias, post hoc fabrication and rationalization, so they are best used in conjunction with other methods, as seen in this study.

Interview questions focused on the experience of learning with LADs – from the initial motivation that prompted participants to access the LADs, to the subsequent actions taken as the result of making learning judgements with them. In the interviews we sought to better understand how participants performed visual interrogations with the LADs, and how the LADs influenced participants' decision-making processes. Questions spanned initial perceptions and motivations for use, learning judgments made, and perceptions of learning at key points during the discussion. Discussion trace data – including message posts, replies, LAD and discussion views – were collected directly from Canvas. This trace data allowed us to re-create decision-making instances to review with participants in the interview. Participants used the recreations of the LADs they saw during the learning activity to discuss their experience with LADs. In conjunction with the trace data collected, these interviews provided contextual information from participants about their feelings, motivations, and cognitive processes – details of their lived experience impossible to glean from trace data alone.

A grounded coding approach was taken in the analysis of the interview data (Corbin & Strauss, 2020). The interview coding took place in stages, beginning with open coding, followed by axial and finally, selective coding (Corbin & Strauss, 1990). Preliminary informational concepts came naturally from the interview questions, which encompassed motivations for initial and continued use, the experience of use, and any difficulties experienced with LADs. In particular, several questions were asked to address judgments of learning made with the LADs. Followed by a round of axial coding, the initial concepts were reduced to overarching thematic codes that reflected the relationships between the themes identified.

The factors of individual different were collected to be able to categorize students according to any of these factors. The results from this learner subset were compared to results from the pilot study to begin building a profile of the characteristics of our student population. The factors were compared with Pearson product-moment correlations to identify relationships between any of them, particularly the BNT and SNS.

6.2.5. Interview script

The first part of the interview script included specific questions about participants' use of the first two LADs during the learning activity. Participants were asked general questions about their experience of the learning activity before being questioned about each time that they accessed the LADs. To facilitate their recall, the LADs were recreated for learners to view as they answered the questions. Participants were asked if they viewed the LADs before they began writing their response to the discussion, if they used them to make judgements of learning (JOL), how the LADs influenced their JOL, and how interactions with the LADs influenced their learning, if at all. They were prompted to recall what they were thinking at the time, and if this perspective changed for subsequent views. In the second phase of the interview learners were asked to explain the rankings given to each of the eight prototypes before and after the cognitive walk-through, and if they would use each of the prototypes in a similar kind of learning activity.

6.3. Results - Phase one

Participants (N=32) were undergraduate students of an interdisciplinary art and technology program at a Canadian university, with most (69%) in their second year. "Moderate to somewhat familiar" was the most often cited response for familiarity with the online discussions (56%), learning environments (62%), and Canvas in particular (66%). They described their general technology skills as moderate (42%) or somewhat familiar (29%), which was lower than anticipated. Participants were asked about their familiarity with both online learning environments and Canvas to see if the participants' experience with LADs could be related to their experience with technology, online learning, or the learning management system. This relationship could be particularly important for learners experiencing difficulty with the LAD, which could ostensibly be related to a lack of familiarity with technology or learning online. This relationship was not evident, as experienced or not, our participants had difficulty using the LADs for a range of reasons that will be discussed further in subsequent sections.

Table 3 lists results for the SNS, BNT, CRT, and PSVT-R. These results represent the individual differences of this student sample. This data was collected to be able to categorize and compare learners according to their abilities. The BNT results (M = 1.35, SD = 1.11) were lower than the general population used to validate the test (Cokely et al., 2012), but similar to results seen in the pilot study (M = 1.5, SD = 0.86).

Participants' SNS (M = 4.16, SD = 0.84) scores reflected a slight preference for numbers over words; these results reflected higher numeracy than those from the pilot study (M = 205. 0.64). The average results for the CRT (M = 1.06, SD = 0.73) were low in comparison to results from the pilot study (M = 2.27, SD = 1.03) and in comparison to the average of the respondents used to verify the test (N = 3,428, M = 1.24) (Frederick, 2005).

The results for the PSVT-R (M = 19.81, SD = 5.58) were higher than the sophomore biology and pre-med majors used to validate the test (M = 14.16, SD = 3.8). They were also higher than the results from the pilot study (M = 9.23, SD = 3.46) though 5 items were removed in the pilot study because they displayed incorrectly.

A Pearson product-moment correlation coefficient was computed to identify any existing relationships between subjective and objective numeracy, cognitive reflexivity, spatial acuity. A relationship was seen between subjective numeracy and spatial acuity ($r = 0.496$, $p < 0.005$), but not between objective and subjective numeracy as expected. Objective numeracy was correlated to cognitive reflexivity ($r = 0.465$, $p < 0.008$), and spatial acuity ($r = 0.425$, $p < 0.017$). The correlations observed between objective numeracy with cognitive reflexivity and spatial acuity were as expected, as it aligned with multiple studies that relate these factors. (M = 1.35, SD = 1.11)

As a highly educated sample, we expected participants to be highly numerate, more than the general population. The average SNS (M = 4.16, SD = 0.84) was as expected, but only because the SNS preference subscale (M = 4.42, SD = 0.92) skewed toward higher-than-average numeracy. The CRT (M = 1.06, SD = 0.73) was correlated with only the SNS ability subscale, $r(89) = 0.21$, $p < 0.04$. Breaking down the results of the CRT further, only 3 participants exhibited the highest cognitive reflexivity; conversely, 16 participants' scores fell at the lowest end of the scale.

Table 3. Exp. 2 SNS, BNT, CRT, PSVT-R Results (N = 32)

	Mean	Standard Deviation
SNS ability subscale (1-6)	3.89	1.1
SNS preference subscale (1-6)	4.67	0.71
Average SNS (1-6)	4.28	0.70
CRT (0 - 3)	1.06	0.73
BNT (0 - 4)	1.35	1.11
PSVT (0 - 30)	19.81	5.58

6.3.1. Causal conditions - Why they looked

Learners were asked to describe their general approach to online learning discussions to provide a baseline experience to compare to the experience of learning with LADs. The majority of those interviewed (N = 17) cited waiting for others to post, before sharing their own work. One student said,

> "I looked at the discussion roughly 4 times. The first time I looked at it to see if people had posted and I didn't [post then]. Then I waited a couple of days and looked at it again."

This was a significant period of inactivity during a 7-10 day discussion activity. In a small group this impacts not only the individual, but also the interaction patterns of the small group.

Over half of the participants (N =17) mentioned looking to their peers for guidance during the discussion activity. The reasons given were encapsulated by the comments of Participant 25 (P25) who said,

> "I don't want to be the first just in case I was wrong… and then maybe I completely misread the question and I was answering it wrong. I might also give the impression to other people that that was how you were supposed to take the question and maybe it might set our group on the wrong path."

It follows then that the LADs could provide some sort of social learning support, similar to that learners sought before the provision of LADs. Social influence was an ongoing

theme throughout the interviews. Four participants accessed the LADs because their friends told them about it; six participants accessed the LADs to be able to compare their performance to their peers. Still, the majority of participants said they had no reason for accessing the LADs. They commented that they were just curious, clicking on the icon representing the LADs "because it is colourful and big," or "to see what would happen." Only two participants stated that they initially accessed the LADs to assess their own performance. One of them mentioned curiosity about the differences in the keywords depicted in the keyword heatmap LAD stating,

> "[I]t looked pretty interesting. First of all I wanted to understand the main concept of this. Then I will think about 'why did he think this way,' with the colours, and why did they put those words but not the others."

The other participant who accessed the LADs to assess his performance was motivated by his misgivings, "…[W]as I not knowledgeable enough, like other people, or is my answer out of the range completely?" Though he said he initially accessed the LAD to judge his *own* performance, this response indicated that he would likely judge his performance in comparison to that of his peers.

6.3.2. Central phenomenon: How LADs were used

The keyword heatmap LAD was used to reflect upon the number of topics discussed, to find new unaddressed topics, and to improve the overall quality of participants' posts. As one participant said, "the keywords provided can help me find a closer answer to the discussion." In this case the LAD provided an immediate remedy to a judgment of learning (JOL) found lacking, because the LAD displayed the sought-after keywords. The keyword heatmap helped one participant find new conversational directions. She had this to say:

> "[I]t felt like at certain points we, the conversation got very stale… I was like oh wow, we have all these things that we haven't even talked about [after talking about the same keyword several days], there's more to go on this discussion. It was a good way to see, 'okay we've talked about this, what else is there to speak about?'"

Another participant used the keyword heatmap as a checklist, saying, "if there is a key word that is not coloured, I will try to go back and try to mention it in the discussion." One participant used the LAD's keywords in aggregate, to compare the

average number of topics her posts contained to the class. Several participants use the keyword heatmap LAD to identify the keywords that they should have used in their posts, or to find new things to say that their peers had not addressed. In this way, they used the LAD to find the language necessary to complete their assignment, or to differentiate their posts from their peers. For one participant the keyword heatmap prompted reflection on the subject matter and in turn, the cultural experience of her peers. She had this to say,

> "It is sort of interesting to see what areas people focus on more. I wanted to know what the percentages of the cultures were that brought up certain words. So like looking at this visualization, it shows that I touch on immigration more than the class average does and I think that maybe says quite a bit about myself and how their cultural standpoint is. Like maybe they are immigrants, so its not as big of a deal for them."

As a tool for reflection, the LAD helped her place herself within her immediate academic community. This student added more perspective with the following statement:

> "I'm not really good at reflecting on myself. It's hard for me to see all the good stuff. It's easy for me to see the bad stuff, so in this way I might be able to judge myself in that way. And the other side of is that if I see other people I might be able to see where I'm missing out as well compared to the collective of the class. Gave me a direction for my research."

This, being able to find and place oneself within a class of strangers, is an important aspect of LAD not often discussed.

Use of the top contributors LAD was more straightforward than the keyword heatmap – participants only cited using the visualization to view their work in comparison to the class. In the comments on this visualization type, the underlying metaphor of the leaderboard was mentioned several times. As expected, the top contributors LAD fostered feelings of competition. P30 said,

> "The board helped me to, understand see, which people might have some better opinions. Not really like the judgment "he's the best, he's better or something," but it could be a reference for me. It made me change my opinion slightly."

Many participants (N=12) mentioned their group's participation as having influenced their own behaviours. For the majority feelings of competition motivated them to post more. Even the person who aptly noted that knowledge of the class average had no influence on grade and thus did not make him feel competitive, stated that he did feel competitive when he recognized the name of one of his friends as a top contributor. This

motivated him to review the assignment instructions again, since he knew that this friend often submitted high quality work. Again this points to social influence – personally knowing the contributors motivated him to put more effort toward his posts.

6.3.3. Intervening conditions: Problems with use

The recognition of how to utilize the LADs was by no means immediate for all participants. Few learners understood the LADs or the information depicted within without difficulty. Some people were unclear about how it was updated, if this was automatically computer-generated or done manually by the teaching staff. One participant mentioned wondering if the LAD's usage was being monitored and would somehow be used in the grading of the discussion. Some misconceptions were considerable – one person thought the LAD was a calendar, another thought it was a recommender that would tell them what to do. It was surprising that someone who made the effort to sign up for the study – including the part that said that they used LADs – say, "I thought this was a schedule or something. I didn't know that the table provided the keywords to help us with the discussion."

Smaller misconceptions, such as thinking that the keywords provided in the keyword heatmap were topics that were brought up by the students, did not impede use. For many of these people, it took several times accessing the LAD before they understood how to use it. P2, who initially accessed the LAD because of curiosity, exemplified this experience. The second time she accessed it she had this to say "I think it did shape my response because I was like "oh, okay this is the context we are supposed to be looking at it from."

Numerous comments indicated that learners did not read the title of the LADs or the instructional tooltips provided underneath. These participants noted aspects of the LADs that they didn't understand, but they didn't make any effort to better understand the LADs or the information presented. Those who did make an effort tended to be more inclined to ask their friends for explanation, rather than reading the text provided with the visualizations. Said one participant, "I had to ask someone, 'what is this for, how are we supposed to use it?' I think there were instructions, I just skipped them." Another

participant gave a longer explanation, explaining the many steps they went through rather than read the LAD instructions.

> "I actually did not analyze the visualization. I looked at it and then I quickly went to the discussion question to see if I needed to include a bar chart. I was actually confused for a moment; I even texted someone to see if we needed to put in a graph somewhere. I don't read stuff. I just look and I skim. I see what I have to do and then I do what I have to do."

This quote amplified a major disconnect. While some students didn't put in any effort to understand the visualizations, some did – that effort just did not involve reading. One of the challenges encountered in these interviews was a significant language barrier. Even with prompting, the text of the discussion threads, and recreations of the LADs to queue recall, many of the interviewees struggled to understand what was being asked of them. For example, when one participant was asked if they used the visualization to judge their work against their classmates, this was the response received.

> "Somebody use the subtitles on it. They posted the subtitles on here. Of all this personalization kind of thing. I don't know, the linear model I could not catch them back. They organized really well. For me I can understand them. There's nothing I can remember most."

In this case, attempts by the researcher to inductively understand what the participant wished to communicate failed. Rich information was lost because the participants could not adequately describe their experience in English.

Feelings of affect and mistrust also influenced how the LADs were used. One participant whose visualization showed less than she had hoped for had this to say,

> "I shouldn't have been surprised but I was. I don't know why I was surprised. I knew I had not contributed anything, well one thing, and the other person contributed two things and they seemed to be far ahead."

Though the visualization did not match the evaluation of her work that she personally held, it nonetheless prompted positive regulatory behaviors. She continued,

> "[T]hat got me to read her comments three, four, five times. I'm like 'what did she say?' Okay, maybe I should do that. I went back and I looked at it. I'm like 'what exactly did she write?'"

The participant then revised her own post, based on what she saw in the visualization. Trust is a requisite for this kind of behavior however. While a few participants were able to directly articulate their feelings of mistrust, saying things like

"*numbers lie*," the majority of participants couched their mistrust in uncertainty. Speaking about the keyword heatmap visualization, P22 said,

> "I would not judge other people based on the chart, or the information there. I would only use them to see how things are laid out, it wouldn't necessarily affect my work as well. It depends."

When encouraged to elaborate he couldn't describe why the visualization would not affect his work, only that he was adamant that he would not use it to judge anyone, his peers or himself. If the participants saw visualizations that did not match their internally held evaluations of their posts, they tended to persist with their personal valuation of their work. Further, their feelings about their work – not just their internally held valuation of it – also swayed their opinions. P3 said that viewing the visualization of her group's high activity motivated her to post more, but if the visualization had shown that the rest of the class was more active than her group, her attitude, and likely her posting behavior, would have been negatively impacted. She was not the only participant whose affect seemed to mitigate what otherwise might have been a positive experience. Said P24,

> "If I was first [in the ranking of the top contributors visualization] I would've checked it more frequently, but since I was at the bottom…"

His statement describes a clear interest in the LADs, but only if he was doing well. In other words, he only wanted to view the visualizations if they were giving him positive feedback. In this sense his motivation to utilize the LADs was rooted in his achievement in the learning activity.

6.4. Methods - Phase two

In this phase participants individually ranked the LAD prototypes before and after performing a cognitive walk-through with them, to see if their impressions changed after this interaction. All eight visualizations were ranked individually. The first rankings were of perceived usefulness and aesthetic appeal; the later rankings represented the apparent appeal and usability of the LADs (Kurosu & Kashimura, 1995). After viewing their aesthetics, the cognitive walk-throughs allowed participants to test a simplified version of the interaction methods for each LAD prototype (see Figure 13). For each prototype the participant answered task-based questions, similar to the estimations of their performance

that they made with the previous LADs during the online discussion activity. This ensured that they understood how to navigate each visualization; participants clicked through animated frames of the prototypes to ask questions as needed. Once this was complete participants were asked to again rank the visualizations, this time using knowledge gleaned from the cognitive walk-throughs to inform their choices. A Pearson's chi-squared test was performed on the results.

6.5. Results - Phase two

The results of the initial forced-choice ranking are displayed in Table 4. One participant abstained from doing any rankings, saying that they did not care for any of the visualizations. For the individual-based visualizations, the fishbowl had the highest number of first-place rankings for both aesthetic appeal and usefulness, while the flower had the highest number of last-place rankings for both. For the visualizations that compared the individual to their peers, the cityscape had the highest number of first-place rankings for aesthetic appeal, while the bouquet of flowers had the highest number of first-place rankings for usefulness. Of the comparative visualizations the bouquet of flowers had the highest number of last-place rankings for both aesthetic appeal and usefulness.

Viewing all of the first-place rankings together in Table 5 shows a great deal of disagreement between the participants. The top-ranking votes were scattered amongst all of the visualizations. The fishbowl received the most first-place rankings for both aesthetics and usefulness, but it received an almost equivalent number of last place readings for usefulness. The single flower received the highest number of last-place ratings for both aesthetics and usefulness. When comparing the group-oriented visualizations, the cityscape got the most first-place rankings for aesthetics, and the bouquet got the most first-place ratings for usefulness. The bouquet also got the highest number of last-place ratings for both aesthetics and usefulness. The highest levels of agreement – for the people who least appreciated the cityscape for its aesthetics or the bouquet for its usefulness – represented only eight of the 21 voting participants, or 38%.

Additionally, only the fishbowl received identical ratings for aesthetics and usefulness, with 24% of the votes.

Table 4. Exp. 2 initial LAD prototype forced-choice rankings

	Fishbowl		Avatars		City		Flower	
Rank	Aesthetic	Usefulness	Aesthetic	Usefulness	Aesthetic	Usefulness	Aesthetic	Usefulness
1st	5	5	3	2	1	3	1	4
2nd	5	2	6	8	3	2	2	3
3rd	8	1	4	3	6	8	1	3
4th	1	3	7	3	4	5	4	4
5th	5	6	3	2	6	5	1	0
6th	5	3	5	6	8	4	1	5
7th	1	5	1	4	1	3	11	5
8th	1	6	2	3	2	1	8	7

	Fish tank		Butterflies		Cityscape		Bouquet	
Rank	Aesthetic	Usefulness	Aesthetic	Usefulness	Aesthetic	Usefulness	Aesthetic	Usefulness
1st	5	3	4	1	8	5	2	8
2nd	6	0	5	8	3	4	1	4
3rd	5	4	2	6	2	2	3	4
4th	4	6	8	3	2	5	1	2
5th	3	6	2	2	5	7	6	3
6th	1	2	5	6	4	2	2	3
7th	4	6	2	2	5	4	6	2
8th	3	4	3	3	2	2	10	5

Table 5. Exp. 2 LAD prototypes ranked 1st

	Fish bowl		Avatars	
	Aesthetic	Usefulness	Aesthetic	Usefulness
Rank 1	5	5	3	2
	Fish tank		**Butterflies**	
	Aesthetic	Usefulness	Aesthetic	Usefulness
Rank 1	5	3	4	1
	City		**Flower**	
	Aesthetic	Usefulness	Aesthetic	Usefulness
Rank 1	1	3	1	4
	Cityscape		**Bouquet**	
	Aesthetic	Usefulness	Aesthetic	Usefulness
Rank 1	8	5	2	8

All of the participants' first choices changed after the cognitive walkthroughs, but what is most important is *why* they changed. The majority of participants selected the "flower" prototype (i.e. polar coordinates) first as the one that they would want to use. Said one participant, "I chose the polar coordinates because they were similar to a pie chart and they are really easy to see to compare to the class." After exposure, this person chose the avatar-based visualization, citing motivation as the primary factor for his choice.

> "Of course you want to perform better to be able to see the avatar, to give it more stuff. Because obviously if you are performing better then you have more ways to get stuff for your avatar and that's pretty cool."

Without prompting one participant had this to say about why their initial choice changed,

> "I know I picked that one first just because I thought their first purpose was just to show some information right up front, but obviously there is more to it. It is not just like a newspaper article where you are just scanning through. It is for learning purposes, so I would choose that last or second to last."

Several participants who chose the simpler visualizations did so because they found the other visualizations distracting or confusing. Overall, the participants liked the

visualizations that fit their mental models, i.e. if they understood them. Several individuals based their choices on aesthetics or affect, the way the visualizations made them feel. Fun was a common theme, as was novelty. Novelty and fun attracted P2 who said,

> "[I want to] see something different because sometimes looking of the representation of the bar chart, it is common. You open the newspaper it's there. 'Oh, okay oil has gone down, using the line going down we know that.' This [the fish tank] is more fun to look at."

The avatar visualization was often perceived as fun and rewarding, even if it was not the participants' first choice. Said P13,

> "I like that [the avatar of a dog or human] it gives a sense of completion to the work as a whole. It gives you something to work towards. Maybe even initially you don't know what its going to look like... Its also sort of like, obviously this is gamifying it."

P29 said,

> "And also it feels like a game; if I do my work I can get a nice avatar out of it. It is a bit more rewarding. Of all the visualizations I'd go with the avatar, because it seems more rewarding, more appealing and attractive, making the task something that I'd want to do."

In this way, the visual appeal was a reward. This is a common mechanic in game design; graphics and audio are inexpensive ways to reward players for their persistence. Rewards in turn, perpetuate a positive feedback loop that keeps players engaged. Of all the participants, a single person cited their feelings as a reason for not selecting a certain LAD. They had this to say:

> "The fish are kinda cool but I don't really like anything that has an emotional connection, because I don't want to be judging how I feel about myself based on what an algorithm is saying my posts are. I just feel like I would prefer more separation. Look at something like this one [indicating avatars] where you get to have a cute puppy. With the example of the girl or the dog I just feel like there's more of an emotional connection to that. Whereas if you're not performing well on this task then you have a shell of a woman that is representing your status in this context. Something like a building if you're not doing well there's less of an emotional connection to a visualization like that."

Assuming that the participants were equally likely to rank the visualizations at any level, a Pearson's chi-squared test was performed on the rankings after the cognitive walkthrough. The avatars were a clear favorite for the single view in terms of aesthetics

χ^2 (3, N=31) = 3.710, p = 0.295. The flower was deemed the least aesthetically pleasing χ^2 (3, N=31) = 26.161, p = 0, and the least useful of the single views, χ^2 (3, N=31) = 7.065, p = 0.07. In the comparison view there were similar results. The fish tank was deemed most aesthetically pleasing, χ^2 (3, N=31) = 2.419, p = 0.49, but the least useful χ^2 (3, N=31) = 2.935, p = 0.402. Conversely the bouquet was deemed the least aesthetically pleasing, χ^2 (3, N=31) = 13.516, p = 0.004, but the most useful χ^2 (3, N=31) = 7.839, p = 0.049.

6.6. Discussion

Many results from this study were unexpected. The measurement of the individual difference constructs lent insight into participants' use of the LAD, just not as hypothesized. Thinking numeracy an influential construct for learners' successful use of LADs, we included both the Berlin Numeracy Test (BNT) and the Subjective Numeracy Scale (SNS). Each measurement had its own benefits and drawbacks. The BNT was likely the cause of the high dropout rate in the previous study, but that study population was largely composed of master's degree students, not our target population. The BNT and SNS were again used in this study to see which measure would be most appropriate for our learner population. Our participant's BNT results were low but correlated with the CRT results, which were also low. Conversely, their results on the SNS skewed toward higher numeracy, due largely to learners' stated preference for numbers over words. This raised a question of validity, which test was accurate?

Though conflicting, we believe both the BNT and SNS results were valid. As evidence we offer the lower than normed results from the CRT, verified by the behaviors described in the interviews. The higher SNS results could represent a genuine preference for the display of numerical information, one that does not need to coincide with numeric aptitude. It is possible that this difference is the result of the unskilled and unaware effect (Kruger & Dunning, 1999). The unskilled and unaware effect describes what happens when an individual's self-assessment is inflated, due to a lack of awareness of one's true ability. It is plausible that learners in the first or second year of college could struggle with their ability to discern their skill level. It is also reasonable that early in their studies,

undergraduate's numeracy may more closely resemble the general population over that of individuals who have completed advanced degrees. Another plausible explanation exists. Many participants for whom English was an additional language had difficulty expressing themselves in English during the interviews. This preference for the display of numerical information could be motivated by the challenge of translation, rather than a genuine preference for numerical information. Though the BNT and SNS results conflict at face value, these suppositions all give a reasonable explanation for these results and are supported by indicated by observations from the interviews and trace data.

The re-created LADs were instrumental in helping learners recall what transpired during the learning activity. While preparing the LAD cues for the interviews, we reviewed the discussion activity experience of each participant. This allowed us to get a sense of each participants' experience and to personalize the questions accordingly – for example knowing when they participated, what posts they read before they contributed to the discussion, etc. This procedure helped us to identify experiences with the LADs that we wanted to investigate more deeply. Having the LADs, these snapshots in time, we were able to probe learners' responses more in-depth, especially when participants' inferences included inaccuracies. The LADs helped participants to identify what transpired during the learning activity, especially when they had trouble remembering the exact sequence of events. Further, they recalled not only their motivations, but also their feelings during the activity. The levels of descriptiveness witnessed in the interviews varied greatly, however these difficulties were primarily due to English comprehension issues. For future interviews with this population, it would be advantageous to include a variety of response elicitation techniques, and to include probes of varying levels of English proficiency.

The primary contribution of this study results from the feedback on how and when learners used the LADs to regulate their learning. The perspective taken in the interview preparation assumed that if exposed, learners would use the LADs to change their learning strategies. There were instances when the learners used the LADs to revisit their learning strategies as expected. In these instances, the most common use of the heatmap LAD was as a list of keywords to include in the discussion posts. For those

whose behavior changed due to the top contributors LAD, the competitiveness invoked by the leaderboard-like iconography motivated the learners to post more.

The learners' difficulty interpreting the LADs and/or the information depicted therein was perhaps the most surprising aspect of this study, followed by their attending to the LADs so briefly. By and large the participants did not attend to the LADs for a length of time – even if there was a disconnect between their perceived performance level and what was depicted. Rather than reflecting upon the discrepancy or reviewing the learning activity instructions, their posts or the posts of their group members, the participants exhibited a greater tendency to proceed with their own, often erroneous, perspectives. In not attending to the LADs or actively thinking through difficulties experienced with them, the participants demonstrated low control of their cognitive impulsivity, similar to their results on the CRT.

Few participants used the LADs without experiencing some sort of difficulty. Many misunderstood the heatmap LAD, which featured uncolored keyword squares until half the class contributed to the discussion. This confusion would be easily remedied by the provision of a tooltip explaining how the LADs populate. This LAD type won't be used in future studies, so the takeaway here was to provide instructions based on learner's expectations. Their expectation of the LADs to "show something," even the learning activity had not yet begun, is an important thing to note when designing future LADs. Since this experience dissuaded some participants from returning to view the LADs a second time, it is important to meet this expectation from the start. Learners' perception was possibly associated with ideals of perceived value; learners expect the LADs to immediately offer usable information because that is what they have experienced with other types of visualizations in their daily lives.

There are several reasons why the participants had difficulty using the LADs, with trust being a major factor. If the LADs did not match the participants' perception of their performance, they tended to either proceed with their internalized beliefs of their performance or to ask another person for guidance. There are a few ways to address trust in the design of LADs, aside from messages and tooltips that help set and maintain learner expectations. Another reason learners had difficulty with the LADs was the lack

of attention paid during visual search. In this instance, the designer's challenge is to work with the brief amount of attention that a user allocates to a visualization. Here the amount of time learners devoted to using the LADs was also related to the individual's valuation of the class, the learning activity, their engagement with the subject matter, and their small group interactions. All of these factors influence how much effort learners devoted to their learning strategies, and by extension, their LADs. Now aware of the brief amount of time learners devote to using LADs, the challenge is to bolster the amount of information learners can glean from LADs within a scant amount of time, and to extend the amount of time that learners are motivated to attend to LADs.

For learners to successfully utilize LADs, they must first choose to engage with them. The second phase of this experiment explored LAD features that might influence learners' decision to initially engage with LADs, on the premise that perceived usefulness and aesthetic appeal influenced this choice. Forced-choice rankings were performed before and after exposure to the LADs, with learners explaining their choices in subsequent interviews. Our results indicated that learners' perceptions of usefulness and aesthetic appeal changed with exposure. To replicate the exposure learners would have using LADs in situ, we performed cognitive walk-throughs with wireframes of LAD prototypes before they were fully developed. Though it is always best to give participants access to a visualization through direct manipulation if possible, it is not uncommon to do heuristic walk-throughs with unfinished prototypes. The practice of assessing users' perceptions of an interface before and after use was utilized in Tractinsky et al.'s seminar work (2000), and in subsequent experiments exploring the relationship between aesthetics and usability (Hamborget al., 2014). The results of such work has been mixed however; the relationship between perceived aesthetics and perceived usability could be influenced by additional determinants, such as the interaction implemented by the interface (De Angeli et al., 2006). As was seen in the first phase of our study, intentionality also matters. While only two participants professed to initially accessing the LADs to assess their own performance, many of the learners' intentions changed with use. Similarly, learners' rankings of the prototypes on the basis of perceived utility and aesthetic appeal also changed with use.

The less abstract visualizations were initially selected for their perceived utility; we believe this was related to the cognitive load required to parse an unfamiliar visualization type. After exposure, participants were more willing to rate an abstract LAD type highly for both utility and aesthetics. This change in rating was, as evidenced in the interviews, because participants felt that they understood how to use the LADs after briefly being exposed to them. Kurosu and Kashimura (1995) described this as the difference between inherent and apparent usability, with apparent usability – i.e. how easy to use an interface appears to be – being more affected by aesthetics than inherent, or functional, usability. Though it remains to be seen if participants' rankings would change again after an extended period of use, this study provides evidence for the use of embellished visualizations to entice learners to initially interact with LADs, and to attend to them longer during the process of learning.

Chapter 7.

Experiment 3 - Conceptual features of abstract LADs

7.1. Introduction

This between- and within-subjects experiment was conducted to compare the accuracy and descriptiveness of gist assessments made after a brief exposure to LADs between two populations, undergraduate learners and Amazon Mechanical Turk workers (MTurkers). In this setting the MTurkers represented laypeople, depicting a wider swath of the general population than university students represent. We also sought to understand if their gist assessments differed according to LAD type when presented with LADs displaying visualizations based on three types of natural scenes. Exploiting the familiarity of regularly occurring scenes and the statistical learning aspect of human vision, we believed that participants' statistical learning systems might prioritize one type of visualization over others, resulting in faster or more accurate gist assessments with one of the three visualization types. Further, the visualizations' novelty or familiarity may aid their memorability. The research questions addressed in this study were:

- **RQ1:** In a comparison of abstract visualizations based on 3 natural scenes, which one prompts the most accurate gist assessments?
- **RQ2:** Which visualization type prompts the highest number of recalled features?
- **RQ3:** Which type of abstract visualization prompts the most descriptive gist assessments?
- **RQ4:** If stated, what are the conceptual features, i.e. mental models, associated with each type of visualization?

This experiment carried three hypotheses. First, it was hypothesized that the accuracy of the gist assessments would be the same between populations. Secondly, all participants would better attend to the nature-based mountain and tree visualizations more than the abstract city visualizations, as evidenced by the accuracy or descriptiveness

of their assessments. Finally, it is hypothesized that the learners' descriptions of gist would be more detailed, since they are currently engrossed in learning activities and have recently seen their own data presented in this manner.

7.2. Methods

We compared the accuracy and descriptiveness of gist assessments made by learners and MTurkers with three types of abstract, natural scene-based visualizations, to see what could be understood from them. As in the previous study, participants completed tasks from the perspective of a fictitious student. Unlike the previous study, every trial involved the same task – describing the gist of each visualization after a 30 second exposure. Participants were then asked to describe all that they understood from the fictitious student's perspective. We sought to determine which LAD prompted the most accurate or descriptive assessments of gist, the LAD that prompted the highest feature recall, and any mental models associated with each LAD type. The LADs were created using secondary learning data; they depicted 7-10 day discussions, similar to the real discussion activity conducted in the previous study. Results were compared between learners and MTurkers, and between the three types of visualizations. This experiment was administered completely online; codes from this analysis were used in the subsequent study.

7.2.1. Participants

Learners were solicited from first- or second-year undergraduate courses. Once permission was obtained from the course instructor, an email was sent out asking learners to participate. As part of an in-class solicitation, a recruitment presentation was shared that gave information about the study and instructions about how to login to the department's research study platform. To closely match the experiences of learners, MTurkers sought for this study were North American residents who had not achieved a bachelor's degree, and who had participated in at least one online learning course that utilized LADs.

7.2.2. Amazon Mechanical Turk

Launched in 2005 Amazon Mechanical Turk[8], or MTurk as it is often called, has increasingly been used for social science research (Paolacci, 2014), information graphics and visualization research (Skau et al., 2015), behavioral research (Mason & Suri, 2011), and education research (Follmer et al., 2017). In this online marketplace anyone with an Amazon account can post a task and set a wage for its completion. Assigned tasks are called HITs, which stands for human intelligence tasks. Any person over the age of 18 with an Internet connection and enabled device may work as an MTurker to complete HITs online.

Techniques to ensure quality data include hiring workers with high reputations and adding qualification tasks or attention checks to surveys. Each MTurker has a reputation based on the number of HITs that they have accepted or rejected; this reputation may be used when soliciting workers. Since MTurk is largely unregulated and MTurkers are working as independent contractors, researchers must ensure that they are paid a fair wage. HITs can be automatically or manually approved by the requester. The benefit of manually approving HITs is the ability to quality check each before paying for the work.

Study results on the attentiveness of MTurkers has been mixed. In a comparative study across four North American colleges and MTurk, Klein et al. (2014) found that MTurkers had a higher rate of completion than any of the college students, even when compared to students who were physically supervised while completing the survey. In other studies MTurkers have been criticized for behaviours such as inattentiveness or failing to read instructions (Crump, 2013), multitasking (Chandler et al., 2014), and working while distracted (Clifford & Jerit, 2014) – behaviors similar to those demonstrated by college students.

Hauser and Schwarz (2015) attribute this to the nature of the MTurker participant pool in the marketplace itself. Specifically, they posited that the MTurk sample

[8] https://www.mturk.com/

population was a non-replenishing subject pool that learned over time, based on the incentivized nature of their work. As the attention checks in online research studies became more common, MTurkers learned that they had to pay attention to get paid.

7.2.3. LAD stimuli

The LAD stimuli were created using secondary data from previous LA studies. The visualized data was selected from 8 different discussion activities carried out over 7-10 days. The selected data reflected different patterns and levels of participation over the course of a discussion activity. Only the timestamp and quality ratings of discussion posts were used in the LADs; fictitious names were generated for each student to reflect a wide diversity of ethnicities. Thirty-two visualizations were created, with two used as examples, followed by 10 trials of each visualization type. Based on pilot feedback, the two examplevisualizations were later incorporated into the main survey.

Scatterplots of all the data were created using Plotly Chart Studio[9] and augmented in Adobe Photoshop or Illustrator, depending on the graph type. Color and grayscale versions of the city buildings, mountains, and trees were drawn in Illustrator. Their shapes were informed by actual landscapes from an existing computer vision dataset[10], then modified for this study. For example, the trees were selected for their differing shapes — one each that was columnar, pyramidal, oval, conical, and an irregular open shape – rather than trees that would be found grouped together in nature. Similarly, the colors used in the LADs were also semi-realistic. For example, the trees were green, but not necessarily the colors one would expect of an oak, pine, maple, or chestnut. The color versions of the graph objects were used to denote the student performance being described; the grayscale versions of the objects represented the other members of the small group. The graph object's sizes were normalized to match the heights of the three different coherence levels depicted in the visualizations; this set of icons were then drawn over each of the data points in the 32 visualizations. A reduced opacity blue sky with

[9] Online software to create visualizations and charts https://chart-studio.plotly.com/

[10] Datasets for Computer Vision Research, http://www-cvr.ai.uiuc.edu/ponce_grp/data/

clouds was added to the background of each visualization, on top of reduced opacity gridlines at each date and coherence level. Figures 13, 14, and 15 are examples of the visualizations.

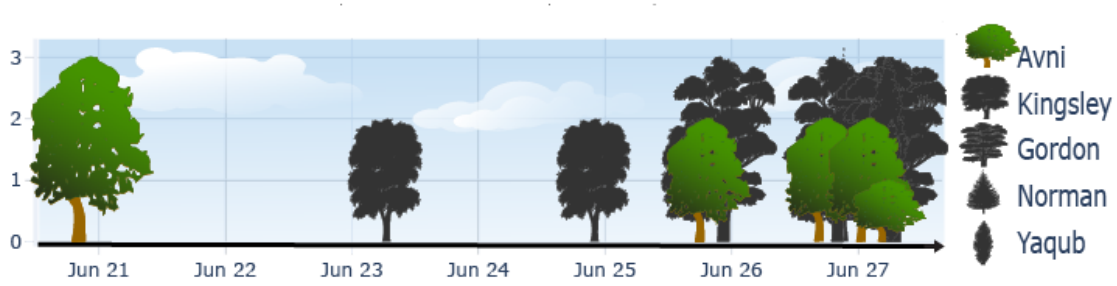

Figure 13. Exp. 3 Avni's tree visualization

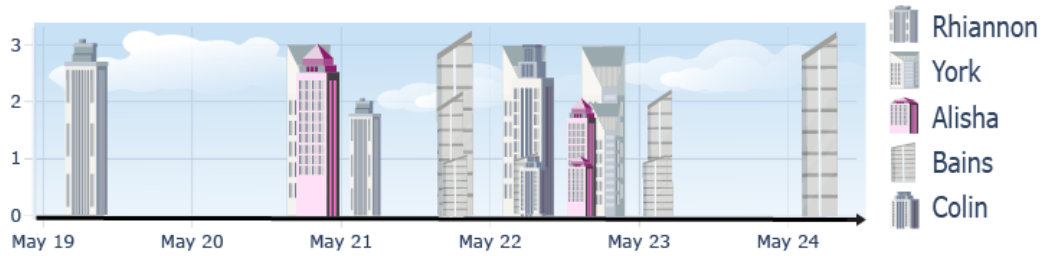

Figure 14. Exp. 3 Alisha's cityscape visualization

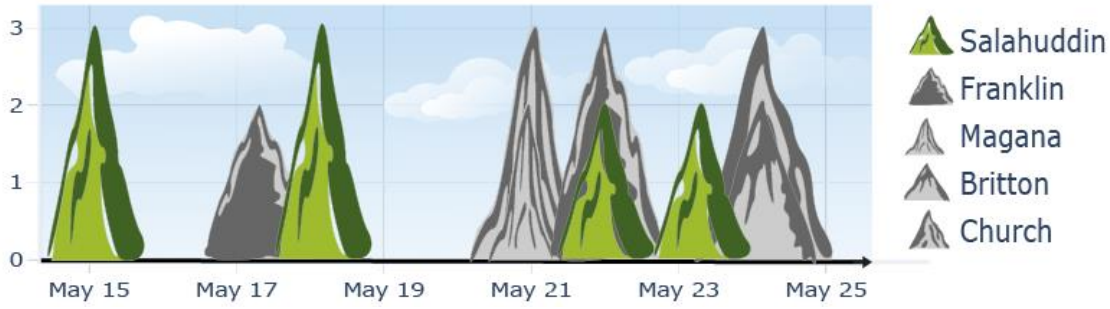

Figure 15. Exp. 3 Salahuddin's mountain visualization

A workaround had to be created to be able to display the LADs for 30 seconds. Earlier studies were hosted on FluidSurveys[11], which allowed videos, animation, and the inclusion of one's own code. With the university's transition to Survey Monkey[12] we lost

[11] Software for hosting online surveys now owned by Survey Monkey.

[12] Online survey hosting software www.surveymonkey.com

the ability to include videos or animated file types in surveys. Thirty second animated GIFs were created with the PNG versions of the visualizations. They were not animated in the sense that the data moved, only that another frame was added such that the LADs were no longer available to view after 30 seconds. Rather than ending this period with a white or black screen, which might create confusion, an additional frame of instructions was added at the end. The frame reiterated the instructions, telling participants to "summarize the image in a 4-6 sentence paragraph." Above the LADs the instructions stated,

> "After you review the graph, write a paragraph that describes everything you see and understand from the graph from the perspective of the highlighted student."

A large open textbox appeared below each visualization for their description, with a button underneath to advance to the next LAD when ready. The gist description was left open ended to avoid biasing responses, and to gather as much qualitative information as possible about what the participant understood from the scene.

7.2.4. Additional study instruments

The Subjective Numeracy Scale (Zikmund-Fisher et al., 2007) was used to describe participants' numeracy. This factor of individual difference was also used to categorize gist responses. Questions about the types of visualizations participants experienced in everyday life were used in the MTurker prescreen to mask the desired worker qualifications. For consistency these questions were also included in the learner version of the questionnaire.

7.2.5. Procedure

This was the first crowdsourced study undertaken in our lab. To ensure its usability it was piloted with both study populations. The learner pilot resulted in changes to the LAD instructions. The MTurk pilot led to the creation of a qualification survey, minute changes to the questionnaire to allow tracking across both surveys, and to a

switch in hosting services, from Amazon Mechanical Turk to TurkPrime[13]. The study was identical for both study populations, but the functional needs presented by crowdsourcing resulted in the different data collection procedures.

LAD instructions

The pilot study identified a need to make the main study instructions more user friendly. The instructions were lengthened to improve clarity, and the word visualization was replaced with the word graph. As one participant mentioned, "*I know how to read graphs; I'm not sure about visualizations."* Another said*, "not sure I know the difference between a graph and a visualization – and I have a master's degree and English is my only language.*" When asked which word they preferred, all of the learners who pilot tested the survey preferred the use of the word graph over the word visualization. An informal survey using 2 versions of the instructions was conducted with MTurkers, to determine which set of instructions were easiest to understand. Both sets of instructions included the types of shapes that could be encountered in the visualizations, but the longer instructions also noted that the position of the shapes could indicate the relationships between them. Of the 21 responses, this longer set of instructions received the most votes (see Figure 16), so this was the set of instructions used in the survey, with the line "*from the perspective of the highlighted student*" added at the end.

[13] TurkPrime is now CloudResearch https://www.cloudresearch.com/

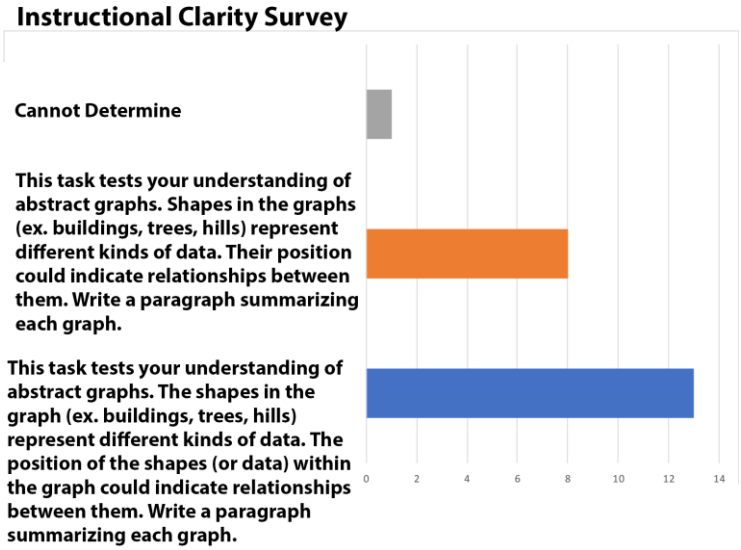

Figure 16. Exp. 3 instructional clarity survey results

After the animated GIFs displayed each visualization for 30 seconds, participants could take as much time as they needed to write their gist responses. Once submitted, participants were not allowed to go backwards in the survey. The initial instructions reminded participants that although they had to complete the survey in one sitting, if they had to leave for any reason they should do so before submitting the current gist response. This allowed participants to attend to personal needs without adversely impacting their responses.

Setting up the MTurk study

HITs are prepaid; MTurk holds workers' payments until the researcher approves their HITs. Any amount can be paid for a HIT, however MTurk charges a 20% fee on the amount paid to MTurkers. If more than 10 MTurkers are assigned to a task, MTurk charges an additional 20% fee. MTurk allows requesters to automatically include or exclude workers based on certain qualifications, for an additional fee. Premium qualification fees start at $0.05 per person. For example, the fee for an MTurker having or not having a bachelor's degree would be $.50 per person. MTurkers may be accepted or limited by almost any type of qualification; MTurk provides a list of the most commonly used qualifications on their website.

MTurkers find the studies they want to complete based on the HIT's description, keywords, worker requirements, the number of respondents required, the maximum time to complete the study, the maximum time the survey will be available on MTurk, and the amount of time the requester has to pay for the HIT. Qualification HITs are often used to screen MTurkers, but these HITs must be carefully worded to avoid priming MTurkers for the skills being sought. All MTurkers can see and preview the tasks for public HITs. All MTurkers can see private HITs too, but only qualified individuals can preview them. Giving HITs short easy to understand descriptions and posting them often helps to ensure that they are seen. New HITs are posted at the top the list, pushing down all previous HITs. Frequent reposting tends to result in faster data collection, but it is an involved, manual process. We learned from piloting this study that it is a good practice to ask MTurkers to provide their worker ID early in a survey. Asking for the ID later allows MTurkers who dropped out to restart it later. It may also be used to remove duplicate responses, to block or track workers across studies.

The maximum time a worker has to complete a survey, the HIT allotment time, is an important choice. MTurk advises requesters to be generous so MTurkers are not rushed. Too much time allows multitasking, which can adversely influence study results, and too little time may keep MTurkers from successfully completing the study. These MTurkers would not get paid, and both the MTurker and requester's reputations may be tarnished. As with other gig economy jobs, reputation has a direct impact on the type and number of jobs available to a worker. The chosen survey expiration time limits workers who batch surveys, or who may only periodically login to MTurk. Finally, it is good practice to not force workers to wait a long amount of time to be paid.

Soliciting MTurkers

The identification of participants in crowdsourced studies is uniquely challenging. While MTurkers may want to contribute to research, their primary motivation is financial renumeration. To maximize the return on the investment of their effort and time, MTurkers often look for the highest paying jobs that require the lowest amount of effort. Workers could also lie to make themselves eligible for a study. To participate in this study, MTurkers had to have participated in at least one online learning course that

utilized LADs and be North American residents who had not yet completed a bachelor's degree. The bachelor's degree qualification was available from MTurk for a fee, but the online course requirement was not. It could be set up as a custom qualification, but the qualification would rely on MTurkers' self-report. When presented as a custom qualification, it would be obvious to MTurkers that this was the sought-after qualification. Rather than do this we offered a separate qualification survey. MTurkers who met the qualification were then emailed an invitation to the main experiment. While this cost a bit more money, it allowed more control in participant screening.

Qualification Survey

In addition to standard demographic questions, the qualification HIT included questions about MTurkers' online activities and the types of graphs or visualizations that they had used online, so the qualifying qualifications were not immediately apparent. There were nine options for online activities, making it more difficult to guess that the desired experience was online learning. Though residency and education were included in the qualifications provided by MTurk, questions about the state or province of residency and the highest level of education completed were included in our qualification HIT. There were 8 possible responses available under the current enrolment question, making it difficult to guess at the intended enrolment option.

The qualification survey took under 5 minutes to complete and paid twenty cents U.S. for completion. It was made public to attract a diverse participant pool. In the HIT description MTurkers were informed that this survey could qualify them for a larger HIT paying $8. To mitigate quality issues that may arise from not limiting the participant pool, the HIT approval rate qualification — an MTurker's successful HIT completion rate — was set to 70%. It cost $5.60 U.S. for 20 people to do the qualification survey; this cost included $1.60 in fees charged by the MTurk platform.

Part way through the study we switched from hosting on MTurk to TurkPrime. MTurk charged extra fees for HITs with more than 10 assignments[14], making it prudent to frequently deploy multiple small HITs. This took a good deal of time and made

[14] Amazon Mechanical Turk Pricing https://www.mturk/com/pricing

tracking MTurkers more difficult. MTurk's standard usage policy is for each person to do a single HIT, rather than a series of HITs, as required for longitudinal or multi-part studies such as ours. To present the studies as seamlessly as possible meant constantly monitoring the qualification surveys to invite qualifying MTurkers to the main survey as quickly as possible. MTurkers then had to be contacted individually, for an additional fee. To mitigate these issues we found a secondary MTurk hosting service, TurkPrime, to administer the surveys. TurkPrime was created to support social and behavioral science MTurk studies (Litman et al., 2017). Its benefits included timed HIT release and micro-batching, allowing requesters to break up HITS and launch them throughout the day rather than all at once. TurkPrime tracks dropout and engagement rates, offers enhanced sampling options, and the exclusion of MTurkers based on previous study participation. Most importantly, TurkPrime made it possible to automate invitations to the main study from the qualification survey. The micro-batching and automated email features made using TurkPrime less expensive than MTurk, even after paying TurkPrime's fees.

7.2.6. Data collection

Participant recruitment for this study happened simultaneously with recruitment for the following study. We anticipated approximately 20 learners and 20 MTurker participants for each of the two studies. Learners were assigned to this study or the subsequent one, with two-thirds of the first study filled before learners were assigned to experiment 4. The MTurker solicitation was more involved. The first two weeks of the study many MTurkers qualified, but so few moved on to complete the main study that we were concerned about a high dropout rate. Anticipating a high dropout rate, all of the MTurkers who qualified in the first three weeks of recruitment were invited to the two studies running, with the goal of having at least 20 MTurker participants in each. A large number of MTurkers (N = 599) were prescreened for this study. In the end, 32 MTurkers and 20 learners participated.

7.2.7. Data coding

Student and MTurker data were collected from SurveyMonkey, cleaned and uploaded to NVivo for hand-coding and analysis. All responses for each visualization were coded at the same time. The visualizations were numbered and named using an alphabetical naming system, according to the fictitious student of interest. For example, all of the responses for 02_AvniT were reviewed at the same time. The naming convention signified that this was the second visualization presented in the survey, it was a tree visualization, and the focus was on the performance of the fictitious student Avni. Responses were coded in the order the visualizations were presented in the survey, then reviewed by participant, to see if any responses were duplicated in an attempt to game the system. Reviewing all of the participants' responses at the same time before moving on to the next visualization made it convenient to compare all of the responses at once. Each visualization was automatically coded according to the fictitious student name, the visualization type, and a participant number. All of the other codes used were manually added; each code then represented its own node.

The gist responses were coded with an open, emergent coding scheme. Each response could potentially be coded at an unlimited number of nodes. The first round of manual coding noted the descriptive aspects of the gist responses – the axes, features, trends identified, etc. (see Table 6). Non-gist related codes were used here too. The *possible omit* code was used to identify responses that sounded as if they were written to game the system, such as,

> "It shows 5 different things. it shows the time period on the bottom. the height on the left side. it shows the growth of those different things."

This response was so vague it could have been used to describe all of the visualizations; responses like this were omitted. The instruction code was used to identify any part of a response that mentioned misunderstanding instructions. An uncertainty code was added to any responses that cited uncertainty, either in the veracity of their gist description or some other aspect of the survey. Statements in this node ranged from participants saying, "I don't know," to expressing difficulty distinguishing the objects within the visualization or the survey itself. Though the last survey question was reserved for participants to be

able to give feedback, some described their experience of the survey while taking the survey, making this experience description part of the visualization response. Creating the experience code allowed data with this code to be extracted from the gist descriptions and analyzed separately. Table 6 contains examples of the initial coding scheme and the responses coded at these nodes.

Table 6. Exp. 3 gist response initial coding

Code Node	Initially coding scheme examples
Description-Axis	These responses describe one or more aspects of the X or Y axes.
"The trees were evenly distributed among the x axis and mostly reached up to 3 on the y axis. The x axis had dates from May 15 to May 25."	
Description-Color	These responses commented on the colors of the objects or the background of the visualization.
"One person Brenda had a different colored building than the others. Brenda's was blue while the rest were grey."	
Description-Group description	These responses focused on the group as a whole, tending to provide a summary of the group rather than the performance of the highlighted individual.
"Buildings represented people. One person had a green building as opposed to the others who had grey. The buildings were all different in structure. Most buildings were placed in the middle of the x axis which represented dates in late may (16-27). The y axis was 1-3 again."	
Description-Individual description	These responses describe the objects of the highlighted person, and sometimes – directly or indirectly – provided details of performance of that person.
"Franklin has the green tree appearing twice with one at 2 on the y axis on may 17th and one at 3 on a later date. He has the only green tree with the others being black."	
Self	Descriptions with this code focused primarily on the student of interest, often to the exclusion of other class members, or the group as a whole.
"Brenda is the highest building. It is the most highest. It is easy to understand." "I am discussing on may 21st, may 23rd and may 24th. The level I am discussing at is at 3 on the 21st and 23rd while at 2 on the 24th. I am overlapping someones discussion on the 21st."	
Self-comparison	Self-comparisons noted differences or similarities between the values of students' contributions, or the timing of these contributions.
"I am doing not as well as the other participants. I only show up twice and one of my trees is considerably lower than the others around me. I fall in the less than average category when comparing my tree to others."	

The second coding pass began with the review of all the nodes identified in the first pass. At this point in the coding, it was clear that learners provided more description than MTurkers, in terms of the number of features mentioned, overall word count, and the number of sentences provided. The majority of the responses from both participant populations were coded as incomplete, and as descriptions. These descriptions included comments on the colors used within the visualizations, the axes, and the visual appearance of individual or group features. All of the description codes were condensed and combined under a single *description* code; going forward this code signified responses that did not interpret what was visualized. The *complete* code was omitted; it was redundant since it was presumed that the provided gist responses were as complete as the participant could produce at the time.

Table 7. Exp. 3 second pass gist response coding descriptions with examples

Code Node	Description of coding scheme with examples
Gist-Accurate (yes/no)	Completely accurate description of gist.
Gist-Class description complete (yes/no)	Complete description of gist using description of the small group, either in whole or part.
Gist-Self description complete (yes/no)	Complete description of gist using only descriptions of the performance of the highlighted individual.
"In this image, I am doing worse than an average performer. I start out strong even though another group member is in front of me. In the second half of the image, I am behind the others with a decreasing contribution to the group. I am not doing well, and I would place myself in the bottom half of the group."	

All of the initial gist related codes were revisited and the nodes were recoded with the codes in Table 7. In the provided example, this response would be coded at all three of the gist related codes. The entire response would be coded at gist-accurate. The gist-self description complete code would be used for the parts of the response saying, “[i]n this image I’m doing worse than an average performer,” and “I am not doing well.” The rest of the response would be coded gist-class description complete. Though these codes work for this example they would not work for responses that were only partially

accurate, such as in a response where the given gist-class description was accurate, but not the summarization of the fictitious students' performance. Gist-accurate was originally used as a code because it was assumed that the provided gist descriptions would be accurate, but this was not the case.

In the final coding scheme gist could be described as one of the following: *accurate, inaccurate, complete, incomplete, details (of self),* or *overview*. Accuracy and completion were separated from other aspects of gist to give clarity to the codes. Their opposites – inaccurate and incomplete – were also coded. These four codes were unique – for example, a response could be either accurate or inaccurate, but not both. To be considered accurate, the gist response must accurately describe performance from the perspective of the fictitious student. The entire description of gist must be accurate to be coded as such. The *details (of self)* code described as a response in which the participant summarizes the performance of the fictitious student (or themselves), without mentioning the performance of other members of the group. The *overview* code described a response in which the participant described the performance of the fictitious student (or themselves) in relationship to one or more members of their small group. The final coding scheme is represented in Table 8.

Table 8. Exp. 3 final gist coding scheme with examples

Code	Coding scheme examples
Gist – Accurate Gist-Inaccurate	Gist responses were either completely accurate, or inaccurate. Below is an example of an accurate response.
"Here, we can see that Brenda is a top and consistent performer. Though she is only represented in 3 days of the data, she is in the green in each day of representation. She represents 1/4 of all the green blocks on the graph."	
Gist-Complete Gist-Incomplete	These responses were either complete or incomplete. Below is an example of an incomplete response.
" Bains scored mildly in the middle of the week."	
Gist-Details of self	These responses described gist in terms of the highlighted person, described the objects of the highlighted person, or remarked upon details of the performance of that person.

Code	Coding scheme examples
"Pacheo's results are highlighted in color in the bar graph above for the week of May 19th. On May 22nd, he is the only one who reported, with a value of 3 (largest green bar). On May 24th, he reported a value of 1 (smallest orange bar). May 24th had the most complex graph, with multiple bars for multiple reports from other people listed."	
Gist-Overview	These responses focused primarily on the group as a whole. They tended to summarize performance in regard to the group, or to describe the performance of the highlighted individual in reference to the group.
" Brenda looks to be present from May 22nd to 24th. When compared to her peers her numbers never fall beneath the maximum amount. She's always at 3 while her peers' numbers change constantly. She takes up a major part of the days she's present."	
Uncertainty	Responses with this code tended to express uncertainty in the instructions, or in the understanding or response for the given visualization.
" I hope I contributed enough information, but it was quite difficult to be thorough when I had less than a minute to view the graph."	
Feedback	This code was used to separate portions of the response referring to the experience of the survey that did not pertain to the gist response.
“This was harder than I expected going in.” “This graph is very esoteric to me.”	

7.2.8. Descriptions of gist

If using the colloquial meaning of the term gist – i.e. all that was understood from the visualization – any response could be understood as gist. For the purpose of this study, gist was defined as the description of performance from the perspective of the fictitious student. Even with this definition gist could be described in a number of ways, making the determination of its accuracy challenging. Take the LAD in Figure 17 as an example. In the visualization there are 8 buildings – few enough to be counted and tallied within 30 seconds. Harper, the person of interest, has 2 entries totaling 5 points. The participants could provide a response saying that Harper made two entries, one each at a medium and a high rating. Participants could count the number of points Harper achieved, or Harper’s points as compared to their group members. Their response could

also state how Harper is doing compared to any one of her group members, or the group in aggregate. The simplest response could state that Harper is doing better than most of her peers. Harper's posts could also be described temporally, according to the time of the week that they were posted. Here she started off posting the first day with the highest-level post, followed by a medium level post, with no posts the following two days.

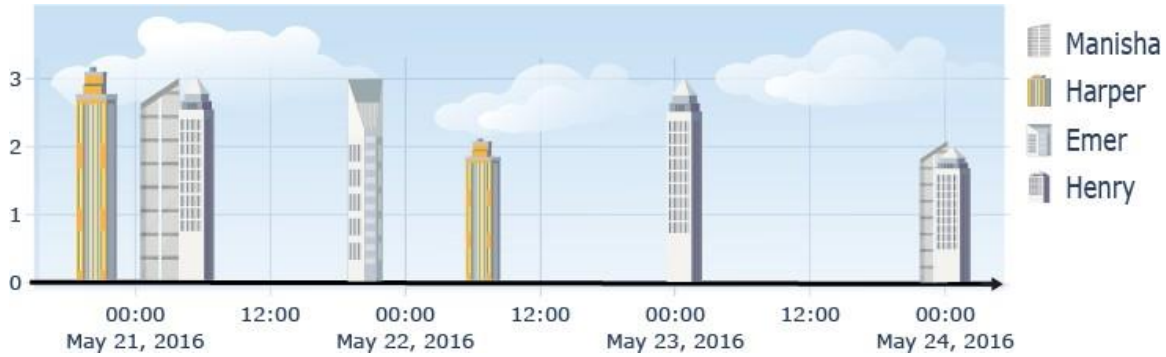

Figure 17. Exp. 3 Harper cityscape visualization

Each of these gist responses approaches Harper's performance in a different way. They illustrate the fact that accurate gist responses could be qualitative, numeric, comparative, temporal, or trend based. All are accurate but use different aspects of the visualization to assess gist. Conversely, a response such as the following is descriptive, but does not give enough information to be able to understand how the fictitious student is performing in the activity.

> "Harper has 2 yellow buildings. The first is on May 21st at 3 on the y axis. The second is on May 22nd at 2 on the y axis. There are 4 names total. The dates on the x axis are May 21 - 24. Henry has two buildings showing. One is on an early date and at 3 on the y axis and one at the end of the x axis the is below 1 one the y axis."

This response represents the majority of those generated by participants – responses that provide detailed visual descriptions of the visualization, without attempting to analyze or summarize performance from the perspective of the highlighted student. Another frequently observed response pattern was the provision of gist responses that were accurate but incomplete, like the following statement.

> "Harper did much better on May 21, 2016, but by the time May 22, 2016 rolled around he was doing worse. At the point of May 24, Harper is no longer even in the running. This chart shows how the students are doing in their discussions depending on the time that the sample was taken."

Harper did better on May 21, but better than whom? Is the perception that Harper was doing better than the group because they posted first, or because the quality of their post was a 3? Is the participant referring to Harper doing better on the 21st than they did on the 24th, i.e. that "better" refers to a comparison of Harper's posts? It is not clear if the participant is comparing Harper's performance to own earlier post, or to their group. While accurate, this gist assessment was incomplete, as it was unclear how Harper was doing. It was an accurate visual description, but an incomplete assessment of performance from the perspective of the fictitious student.

7.2.9. Data analysis

The coded data was visually reviewed for patterns and trends, such as which participant group provided the most accurate and complete responses, or which visualization type received the highest number of accurate responses. The responses of each individual over time were visually reviewed to see if trends could be identified across the duration of the survey – not just across visualization types or participant groups – to see if performance improved over time. A t-test was used to compare the total number of accurate and complete responses produced by either learners or MTurkers. For the within-subjects portion of the analysis, a oneway analysis of variance (ANOVA) was performed to compare the means of the number of accurate and complete responses within each participant group. This was followed by a post-hoc chi-squared analysis since the results were not normally distributed.

7.3. Results

Table 9 contains demographic information for learners, the prescreened MTurker participant pool, and the MTurkers selected for this and the subsequent study. On average the MTurkers were a decade older than learners, and the majority were not currently enrolled in college. The results are reported in this way – with the screened and selected MTurkers reported in large groups – due to difficulty tracking individual MTurkers through the prescreen and main surveys. MTurker IDs are a mix of alphanumeric characters; if they made a typo or input the wrong ID in either of the surveys, it was

impossible to track them. Had we administered the survey entirely on TurkPrime, these IDs would have been automatically collected, allowing for easier tracking.

Table 9. Exp. 3 participant demographics

	Surveyed MTurkers (N = 599)	Selected MTurkers (N = 63)	Learners (N = 16)
	Demographic Information		
Female	309	39	6
Male	291	22	10
Transgender, two-spirit, agender	7	2	0
Age range in years	18-72	18-63	19-25
Mean Age (SD)	35 (10.7)	33 (9.8)	21(1.6)
	Highest level of education		
High school degree or equivalent (e.g., GED)	76	4	10
Some college but no degree	141	42	5
	Current enrollment		
Full time at a 4- year undergraduate college/university	54	5	14
Full time at a 2-year undergraduate college/university	12	5	1
Part time at a 4- year undergraduate college/university	24	8	1
Part time at a 2- year undergraduate college/university	14	4	0
Not currently enrolled	442	41	0

Results of the SNS for the MTurkers are below, in Table 10. Unfortunately, a mistake in the survey set up resulted in the SNS not being administered to learners. The average SNS scores of the learners from the previous study (N=32) at 4.3 (SD 0.7), falls

within the two MTurker averages, being just slightly higher than the average SNS of the selected MTurkers.

Table 10. Exp. 3 MTurker SNS results

	Screened MTurkers (N = 599)	Selected MTurkers (N = 63)
Avg SNS- Performance	4.4 (SD = 1.1)	4.2 (SD = 1.1)
Average - Preference	3.9 (SD = 0.5)	4 (SD = 0.5)
Avg SNS	4.1 (SD = 0.6)	4.1 (SD = 0.7)

Participants were asked if they used common visualizations found in everyday life such as banking, utility and bill payment, health, time management, and educational applications. As may be seen in Table 11, these populations have had extensive exposure to visualizations. Zero learners and 3 of the selected MTurkers – representing 0% and 5% of their groups respectively – had never used any of these types of visualizations.

Table 11. Exp. 3 prior visualization experience

Visualization type	Mturkers (N = 599)		Screened Mturkers (N = 63)		Learners (N = 16)	
Banking graphs (ex. a checking account balance, bill payments)	371	62%	45	71%	9	56%
Educational graphs	281	47%	38	60%	15	94%
Utility graphs (ex. electricity or gas usage)	312	52%	35	56%	5	31%
Telephone or internet usage graphs	340	57%	38	60%	9	56%
Loan payment graphs (ex. mortgage, student loans)	264	44%	35	56%	1	6%
Time planning or tracking software graphs	217	36%	26	41%	8	50%
Laboratory result graphs	195	33%	21	33%	5	31%
Health or exercise tracking graphs	377	63%	50	79%	6	38%

Visualization type	Mturkers (N = 599)		Screened Mturkers (N = 63)		Learners (N = 16)	
None of the above	53	9%	3	5%	0	0%
Other (please specify)	18	3%	0	0%	1	6%

7.3.1. Study completion rates

Twenty learners began the study; 16 completed it. The completion rate for learners, was 80%; for MTurkers it was 59%. The four learners who didn't complete the study stopped just after providing their demographic information. Of the 32 MTurkers qualified, the responses of 8 MTurkers were omitted during data analysis. One MTurker stopped before providing their first response; 5 stopped after the first visualization, presumably using the first question to preview the survey before deciding not to participate. Two MTurkers attempted to game the system, using one response for all 32 visualizations. The responses of these 8 participants were completely omitted. Five MTurkers partially completed the survey. Their responses were included in the analysis, but not the completion rate. Two MTurkers stopped at LAD 11, one at LAD 12, one at LAD 22, and the last at LAD 25.

7.3.2. Accurate gist responses

To see if the provision of accurate responses was related to visualization type, we compared their distribution responses across visualization types. Since the number of usable responses differed for each LAD the accurate gist responses are reported as percentages (i.e. the number of accurate responses out of all responses provided). As seen in Table 12 below, MTurkers provided more accurate responses than learners, across all visualization types.

Sorted by type and percentage of accurate responses garnered, the table also suggests that participant performance differences according to graph type may exist. To illustrate the differences in accuracy between learners and MTurkers, responses with accuracy over 50% are highlighted. Overall learners provided fewer accurate responses

than MTurkers. Further, learners tended to provide accurate gist responses for many of the same visualizations that MTurkers performed well with.

Table 12. Exp. 3 accurate gist responses for learners and MTurkers

City Visualizations			Mountain Visualizations			Tree Visualizations		
	Learner	Mturker		Learner	Mturker		Learner	Mturker
03 BainsC	14%	16%	01 AlishaM	0%	6%	02 AvniT	9%	13%
04 BrendaC	5%	12%	11 EmerM	13%	13%	05 BrittonT	5%	17%
06 ChurchC	42%	55%	14 GordonM	31%	39%	07 ColinT	0%	61%
08 DaniaC	6%	16%	17 HenryM	25%	55%	09 DerekT	0%	24%
10 EamonC	0%	17%	20 LevisonM	25%	59%	12 FranklinT	6%	9%
13 GiuliaC	19%	25%	23 MaganaM	33%	67%	15 GuyT	31%	36%
16 HarperC	19%	55%	24 ManishaM	27%	48%	18 KayleyT	19%	64%
19 KingsleyC	38%	59%	26 PachecoM	40%	62%	21 LidiaT	40%	64%
22 LilithC	7%	45%	28 RhiannonM	36%	79%	27 PattersonT	13%	73%
25 NormanC	53%	67%	31 YaqubM	54%	55%	29 RoyT	54%	70%
30 SalaC	62%	75%				32 YorkT	69%	50%

7.3.3. Accurate and complete gist responses

Of all of the gist responses produced by both participant groups, the fewest were accurate and complete (Table 13).

Table 13. Exp. 3 accurate and complete gist responses for learners and MTurkers

City visualizations			Mountain visualizations			Tree visualizations		
	Learner	Mturker		Learner	Mturker		Learner	Mturker
03 BainsC	0%	0%	01 AlishaM	0%	6%	02 AvniT	9%	6%
04 BrendaC	0%	3%	11 EmerM	0%	4%	05 BrittonT	0%	8%
06 ChurchC	25%	45%	14 GordonM	25%	22%	07 ColinT	0%	43%

City visualizations			Mountain visualizations			Tree visualizations		
08 DaniaC	0%	8%	17 HenryM	19%	32%	09 DerekT	0%	4%
10 EamonC	0%	4%	20 LevisonM	13%	32%	12 FranklinT	0%	4%
13 GiuliaC	0%	8%	23 MaganaM	7%	33%	15 GuyT	19%	23%
16 HarperC	19%	23%	24 ManishaM	13%	38%	18 KayleyT	13%	23%
19 KingsleyC	19%	27%	26 PachecoM	13%	48%	21 LidiaT	7%	23%
22 LilithC	0%	27%	28 RhiannonM	21%	58%	27 PattersonT	13%	55%
25 NormanC	20%	43%	31 YaqubM	15%	25%	29 RoyT	23%	60%
30 SalaC	15%	55%				32 YorkT	15%	23%

An a priori statistical power analysis was performed with G*Power software (Faul et al., 2007) to estimate the required sample size using a medium effect size of 0.5 according to Cohen's (1988) criteria for a t-test with alpha = .05 and power = 0.90, the projected sample size would need to be approximately N = 34. For power = 0.80, the projected sample size would be N = 26. The participant sample size met this criterion.

A two-tailed t-test[15] was conducted to compare the total responses that were both accurate and complete between the two participant groups, learners and MTurkers. There was a significant difference between the responses of MTurkers (M = 7.33, SD = 7.99) and learners (M= 3, SD= 5.38); t(38) = 2.05, p = 0.005. This result suggests that a difference exists between these two populations, but not in the expected direction, since MTurkers produced more accurate and complete responses than learners.

A oneway analysis of variance (ANOVA) was performed to compare the means by visualization type within each participant group (Table 14), to determine if any of the means of the number of accurate and complete responses differed from the others for each of the participant groups. There were no statistically significant differences between

[15] The test conducted was the Aspin-Welch-Satterthwaite-Student's t-test using JMP 15.

group means for MTurkers as determined by one-way ANOVA ($F(2,69) = 0.15$, $p = 0.86$), or for learners ($F(2,45) = 0.13$, $p = 0.88$).

Table 14. Exp. 3 accurate and complete gist means by visualization type

	Learner	***Mturker***
City Visualization	M = 0.88 (SD = 1.75)	M = 2.21 (SD = 2.48)
Mountain Visualization	M = 1.19 (SD = 1.97)	M = 2.63 (SD = 2.87)
Tree Visualization	M = 0.94 (SD = 1.73)	M = 2.5 (SD = 2.83)

The distribution of these results was not normally distributed. The distribution of accurate and complete responses for learners had a skewness of 0.04 and kurtosis of -1.49. For MTurkers the skewness was 0.26 and kurtosis was -1.13. The skewness for both participant groups was acceptable, but the kurtosis values for both were less than -1. This meant both distributions were too flat, making the distributions non-normal (Hair et al., 2017, p. 61). Viewing the graph of the results confirmed that they were not normally distributed, making the assumption of normality not viable, so a follow up nonparametric test was conducted.

To determine if the provision of accurate and complete responses was related to visualization type, we performed a chi-square analysis for two or more independent samples. Assuming independence between visualization type and response type, we put forth the following hypotheses for both groups of participants, learners and MTurkers:

H_0: Visualization type has no relationship to the provision of accurate, complete responses

H_1: Visualization type is related to the provision of accurate, complete responses

The two categories of responses used for the analyses are 1) accurate and complete, and 2) inaccurate, incomplete, or inaccurate and incomplete. Using the chi-squared values from the contingency tables below (Tables 15 and 16), 2 degrees of freedom, and the value from 0.05 probability from the chi-squared critical value table:

For learners, $\chi^2(2, n = 512) = 1.9, p < 0.05$

For MTurkers, $\chi^2(2, n = 798) = 4.5, p < 0.05$

Since we cannot reject the null hypothesis, that the visualization type would have no relationship to the provision of accurate complete responses by either learners or MTurkers, these results support our earlier hypothesis that participants would perform better with some types of visualizations.

Table 15. Exp. 3 contingency table for MTurker's accurate and complete responses

Mturker contingency table			
	Accurate and Complete responses	Inaccurate, Incomplete, or Inaccurate and Incomplete	Total responses
City	53	214	267
expected	63	204	
	1.51197582	0.46441142	
Mountain	63	162	225
expected	53	172	
	1.94072126	0.59610287	
Tree	60	197	257
expected	60	197	
	0.01634597	0.00077303	
Total	176	573	749
so χ^2 = 4.530330367			

Table 16. Exp. 3 contingency table for learners' accurate and complete responses

Learner contingency table			
	Accurate and Complete responses	Inaccurate, Incomplete, or Inaccurate and Incomplete	Total responses
City	14	164	178
expected	17	161	
	0.43281835	0.04477431	
Mountain	19	139	158
expected	15	143	
	1.18380802	0.1224629	
Tree	15	161	176
expected	17	160	
	0.13636364	0.01410658	512
Totals	48	464	512
so χ^2 = 1.9343338			

7.3.4. Visual analysis of learning progression

It was observed that as both populations proceeded in the survey, their gist assessments became more accurate as time went on. Thinking that this was evidence of a learning effect that would be seen with regularity, the results were rearranged and re-visualized in cell plots (see Figures 18 and 19). Each cell has a square for accuracy, completion, details of self, and overview. Red cells indicate a negative value, for example an inaccurate response, and green cells represent positive responses. Responses for each individual were plotted as a horizontal line, then reorganized and grouped according to when the first accurate response was seen.

In Figures 18 and 19, participants in Group 1 produced no accurate responses for the duration of the survey. Four MTurkers (17%) and 3 learners (19%) were in this group. Of the MTurkers in this group, 2 of 4 quit the survey approximately halfway through.

Participants in Group 2 produced at least one accurate response within the first 5 responses. Four learners (25%) and 3 MTurkers (13%) were in this group. Group 3 is comprised of individuals who provided accurate responses within their first 10 responses. This group is the largest and most successful for both groups of participants, 13 MTurkers (54%) and 8 learners (50%).

Group 4, the final group, represented those who provided no accurate responses within the first 10 provided, but who did provide an accurate response at some point. Four MTurkers (17%) and 5 learners (31%) were in this group. Looking at the pattern of when accurate responses were provided, it was interesting to note that accurate responses seemed to be "activated." If an accurate response wasn't provided within the first 10 responses, it was likely that one would not be given. Further, there was a noticeable group of participants in both groups who provided accurate responses within the first ten, followed by a brief period of inaccurate responses, and then by responses that were largely accurate.

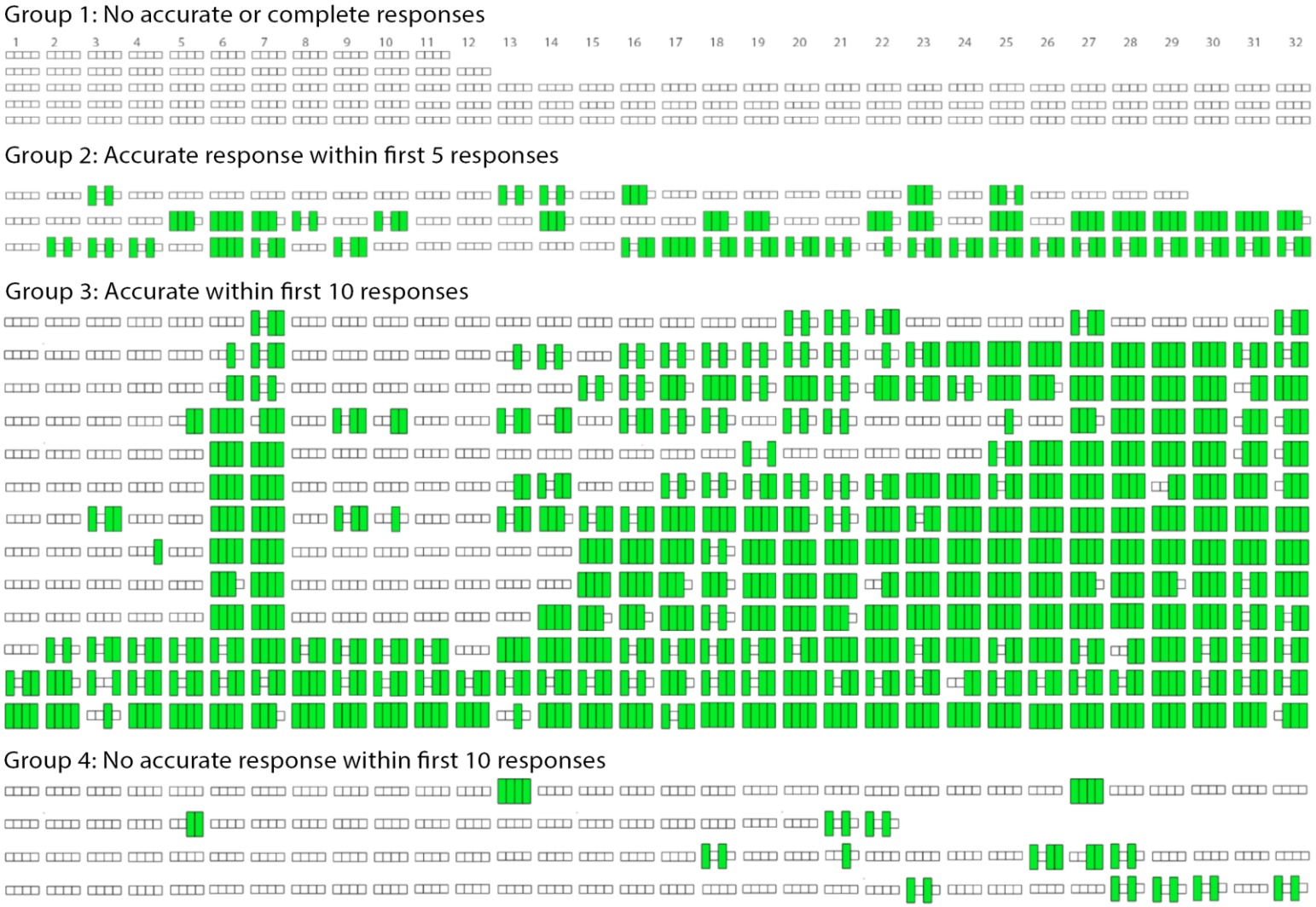

Figure 18. Exp. 3 MTurker cell plots

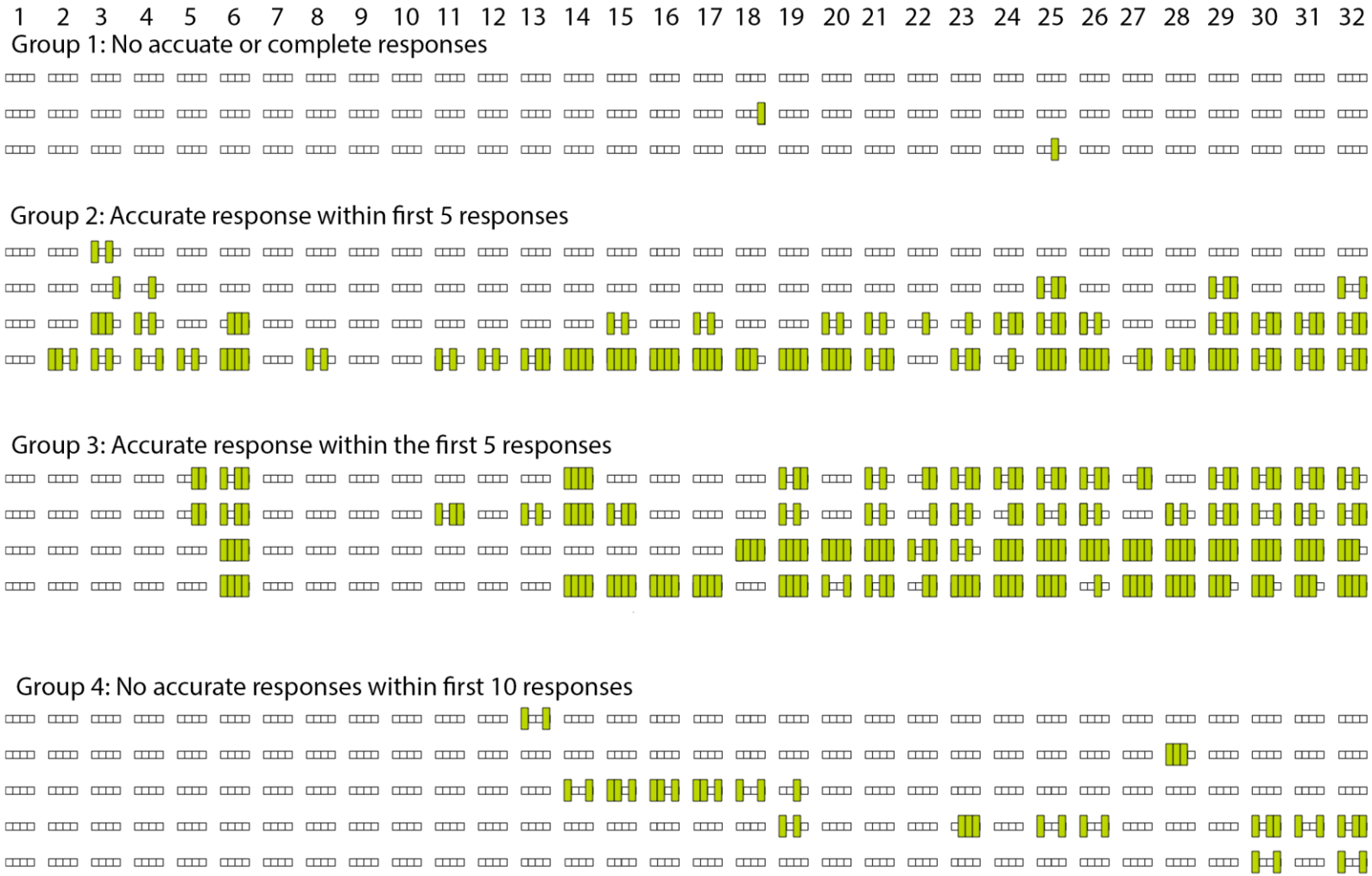

Figure 19. Exp. 3 learner cell plots

7.3.5. Gist accuracy and completion

Across all visualization types, MTurkers provided more accurate responses, and more responses that were both accurate and complete. A significant difference was found between the total number of accurate and complete responses produced by learners in MTurkers, however the distribution of the responses was not normal due to the kurtosis measure of both sets of responses. No significant differences were revealed after performing an ANOVA for each participant group. A chi-square analysis was done to again compare the number of accurate and complete responses according to visualization type for each participant group. In response to RQ1, both learners and MTurkers made the most accurate and complete gist assessments using the mountain visualization.

As hypothesized, learners provided more detailed descriptions of gist than did MTurkers. Learners' responses tended to be longer than MTurkers in terms of both sentence number and word count. The majority of their gist assessments were coded as incomplete, and/or as descriptions. Many participants in both groups simply described what was before them in the visualization, rather than making an assessment on the objects therein. Learners tended to mention axes, heights and counts of objects within the visualizations. The majority of these comments did nothing to further understanding of the highlighted students' performance. The descriptions made by MTurkers tended to note clusters, distributions, increases, decreases, and other patterns found according to the timeline of the discussion activity. These comments were closer to what was expected to be included in the gist assessments and tended to accompany accurate gist assessments. Though not explicitly stated in RQ2, the underlying assumption for the research question – determining which visualization prompted the highest number of recalled features – was that the ability to recall features would have a positive influence on gist. This was not the case. Learners' responses were quite verbose, but this aspect of their responses was not positively related to the provision of accurate or complete gist responses.

With regard to RQ2 and RQ3, we assumed that gist assessments would first be accurate, and only then we would count the number of recalled features or dress levels of descriptiveness. What we found by how gist was described was that the most descriptive

assessments were most often incorrect, because they described the visualization more than gist. As such, these questions did not help us better understand how accurate assessments of gist were conceptualized.

In response to RQ4, accurate and complete responses tended to include a description of performance that was self-oriented and an overview of the performance of the group. These results support H2, that all participants would better attend to the nature-based mountain visualization more than the city visualization as evidenced by the accuracy of the assessments. Cell plots illuminated trends observed in the visual analysis of the data. These trends might speak to the number of exposures to LADs necessary for learners to "get the gist."

At the beginning of the experiment we made three hypotheses. H1, that the accuracy of gist responses would be identical between participants, was not supported. MTurkers generated more accurate and complete responses than learners. The second hypothesis was that both participants groups would better attend to the mountain and tree visualizations than they would the city visualization, as evidenced by the accuracy and descriptiveness of the gist responses. As seen in the contingency tables, both MTurkers and learners performed better-than-expected with the mountain visualization and city visualizations. Performance – in terms of the provision of accurate and complete descriptions of gist – was better with the mountain visualization for both populations but mixed with the tree visualization. While MTurkers performed as expected with the tree visualization, learners' performance was worse. This lends partial support to H2. H3 — that learners' responses would be more detailed — was not supported.

7.4. Discussion

Recruitment for this study presented several challenges. We expected more learners to participate since the recruitment class had an enrollment of 300. To prepare for this study I signed up for an MTurker account, performing several tasks as an MTurker, to understand how studies are found, selected, and completed from the MTurker's perspective. Even with no prior qualifications there were hundreds of studies available to me at any one time. It was easy to understand then how studies can get lost

and why MTurkers use alternate platforms to help them track HITs and optimize their time.

Dropout rates were a concern throughout this study. During the pilot we learned that 20 MTurker participants could be found in a manner of hours, even though the qualification surveys were each set up to run for a full week. Participation in the main study was less immediate. Though many MTurkers qualified, few who qualified proceeded on to the main study. After the third week the study's administration was changed to mitigate any further issues that could be experienced due to high dropout rates. The study was initially planned to be counterbalanced, but if there was a high dropout rate, not enough responses would be collected for the same visualizations to render the data usable. Not varying the presentation order of the conditions had an unintended benefit, as it allowed us to visualize the participants' learning effects over time.

7.4.1. Gist accuracy

In the previous study, learners tended to act on what they perceived at a glance from LAD. By significantly limiting the duration of time participants were allowed to view the LADs in this study, we attempted to discern what could be understood from gist. In asking participants to describe performance from the perspectives of the fictitious students, they were being asked to surmise gist within the context of a learning activity. MTurkers were significantly more apt to produce accurate and complete descriptions of gist than learners. We offer scenarios that could explain the significant disparity between learners and MTurkers in the production of accurate gist assessments.

The first explanation is that learners did not thoroughly read or understand the instructions. MTurkers are motivated to do so since for many, this work is their livelihood. Aside from this, it is possible that MTurkers paid greater attention to the survey. Their comments on the survey suggested this. MTurkers expressed uncertainty in the accuracy of their answers more often than learners. While one MTurker explicitly stated that they thought that they were doing the hit wrong because the question at the top of the screen did not change, several mentioned the instructions not changing. These

expressions of task uncertainty point to a level of task reflexivity not verbalized by learners.

To correctly perform the task, participants had to understand its description. This is why instructions were provided at the beginning of the survey, at the top of every screen where the LADs were displayed, and at the end of every 30 second LAD display. The instructions displayed as part of the LAD GIF summarized the other instructions, saying only to "summarize the image in a 4-6 sentence paragraph," but it is also possible that some participants only followed these instructions. One MTurker noted in their feedback that some of their responses were based on the shorter instructions. They said,

> "If I were to give an excuse for this, it would be because I knew the graph is disappearing and 30 seconds and I wanted to make sure to absorb as much of it as possible. I got a little hyper focused on that. I am very sorry. The moment I noticed the directions above the graph I change my responses to better fit what you are asking. Hopefully, you can understand my confusion and this will not result in a rejection! This was very fun and easy to follow once I realized my mistake."

It is possible that the same thing happened to some of the learners, but they never realized their mistake. The learner-provided feedback – comments like "this hurts my brain," and "it seemed very repetitive" – did not give as much insight into learners' thought processes. The following learner comment came close but lacked the detail necessary to draw fruitful conclusions from it. They said, "similar and meaningless graphs showing, again and again, makes people become more and more doubt about their thought."

How is it possible that learners' responses were so long and detailed, if they did not understand the survey task? Though learners produced more verbose responses, these responses tended to describe what was visually present without attempting to assess performance from the perspective of the fictitious student. In describing what was visually present rather than attempting to analyze the depicted relationships, learners were doing the least amount of cognitive processing possible to participate in the study. In writing 4-6 sentences they did fulfill part of the stated objective – they wrote a paragraph.

Maturity could also be a factor, along with motivation. On average, MTurker participants were a decade older than learners. They may have more exposure to

visualizations and more practice making performance-based estimates than learners. Learners' remittance for participation was course credit, which they still might receive for less than optimal performance. There was no penalty for poor performance, save for not receiving course credit. MTurkers were notified in the consent form that there would be no penalties for poor performance, however one of the core features of the MTurk platform is reputation. HIT rejections – those marked unsatisfactory by their requesters – are part of MTurker reputation, an influential factor in future job qualification.

Results from the within-subjects portion of the analysis suggest that for many participants, their ability to make accurate gist assessments increased with repeated exposure over time. For both MTurkers and learners, this was the largest group of participants. This was also the group of participants that was the most successful with the LADs. While the long-term implications are positive – that repeated exposure and practice with LADs have a positive influence – they do little to improve learners' rapid decision-making with LADs. Considering the practical use of LA, these results are discouraging. In the previous experiment we saw that learners rarely viewed a visualization five times in the progression of a discussion activity, let alone 30. Unlike the current study a real discussion activity carries with real personal consequences, but even with these consequences, learners only briefly attended to the LADs. If learners from the previous study didn't understand the LADs the first time, they rarely attempted to use it a second time. Repeated exposure may increase the accuracy of judgments of learning made with them.

Though MTurkers performed better than learners overall there was one visualization type with which leaners performed better than expected. In comparison to the other visualization types, the mountain visualization representing a medium amount of abstraction. It is possible that this visualization type introduced just enough visual difficulty to require more cognitive effort to understand it and in doing so, made the gist more memorable. This would be consistent with Yue et al. (2012), which demonstrated the addition of visual difficulty positively influencing rapid judgments of learning, and Hullman et al. (2011), who argue that the introduction of some visual difficulty may stimulate learning.

Chapter 8.

Experiment 4 – Proportional estimates of gist

8.1. Introduction

This quasi-experimental study was undertaken to further understand how learners interpret visualized learning performance data using LADs. As in the last study, we compared the gist assessments of learners and MTurkers made after a brief 30 second exposure to three different types of LADS. In this study, the LADS were chosen based on their facilitation of proportional estimates. For the three visualization types selected for this study — bar chart, pie chart, and stacked bar chart visualizations — proportional estimates are made differently. MTurkers again represent laypeople because they represent a wider swath of the general population than university students.

In the previous study, three abstract visualizations were used to see if aspects of the human visual system would prioritize one visualization type over others, resulting in the production of more accurate or descriptive gist assessments. MTurkers produced more accurate and complete responses overall. There was only one visualization type with which learners performed better than expected. Compared to the other LADs, the mountain visualization represented a medium amount of abstraction. It is possible that this LAD type introduced just enough visual difficulty to briefly capture participants' attention, making it more memorable. These results helped to establish a baseline of learners' performance to be used in future studies. This study extends the previous study by addressing one of the primary visual tasks that learners perform with LADs, using 3 new visualization types.

The bar chart is perhaps the most commonly employed visualization type for the comparison of categorical data (Bertin, 1983; Zacks & Tversky, 1999). Both bar charts and stacked bar charts use a Cartesian coordinate system, which facilitates comparisons based on relative length (Yau, 2013). Cleveland and McGill (1985) rated length highly as a visual cue, second only to position. When the bars being compared share an end or anchor point, comparison becomes even easier (Yau, 2013). Upward and downward

trends are more readily identified in Cartesian orientations than in polar coordinate systems (Yau, 2013).

Using a polar coordinate system, pie chart visualizations are also familiar. Length is a perceptual cue for bar and stacked bar chart visualizations, but for the pie chart visualizations perceptual cues are the relative difference between angles, the area of the pie segments, and radius length (Siirtola et al., 2019). The efficiency of pie chart visualizations in comparison-type tasks is contested – proportional estimates tend to be made faster with Cartesian orientations than with polar orientations.

Pie chart comparisons can be complicated, because they require the comparison of the relative number of degrees within each segment of a circle. No matter how many segments there are, all of the angles always add up to 360°. If two angles are being compared then the opposite angle is the conjugate of the first, and the two are quickly compared. Visualization experts like Stephen Few (2007) suggest the use of other, more efficient types of visualizations, especially when comparing multiple part-whole relationships. Criticisms of the pie chart stem from human's inability to accurately estimate angles or the area of each segment (Skau & Kosara, 2016). We tend to underestimate acute angles and overestimate obtuse ones (Robbins, 2012). Proponents of pie charts argue that when choosing between bar and pie charts, task type matters (Hollands & Spence, 1998; Spence & Lewandowsky, 1991). Trends can be seen with pie charts, but this depends on the data and task type. Though Siirtola (2019) found participants to perform faster estimates of proportion with stacked bar charts, there is evidence that pie charts are as effective as bar charts when doing part-whole estimations (Spence & Lewandowsky, 1991), and that it takes longer to make part-whole estimates with bar charts (Hollands & Spence, 1998). Holland and Spence's (1998) summation model hypothesizes that when making part-whole estimates, people must first establish what the "whole" is – with pie chart visualizations this information is readily available.

In terms of complexity, the stacked bar chart lies between the bar and pie chart visualizations. Unlike bar charts, each stacked bar chart represents the whole of the entity it represents. Cleveland and McGill's seminal study (2012) found participants to be more accurate with aligned bar charts. In an attempt to replicate and explain these results

Talbot et al. (2014) performed 4 experiments with different types of bar charts, concluding that distractors decreased the visual saliency of stacked bar charts.

Using these three visualization types selected for their facilitation of proportional estimates, the objectives for this study were to investigate:

- RQ1: Do learners produce more accurate and complete gist assessments than MTurkers with one of the visualizations employing 3 different methods of estimating proportion?
- RQ2: Are more accurate or complete gist responses produced by learners or MTurkers according to the type of visualization?
- RQ3: Do participants with high numeracy produce more accurate or complete gist descriptions?

This experiment carried two hypotheses. First, we hypothesized that for both participant groups, the accuracy of the gist responses made with bar charts would be higher than the other visualization types. Proportional estimations with the stacked bar chart are perhaps the most challenging, however the visual difficulty may incite participants to attend closer to this visualization type. Using the only factor of individual difference carried forward from the previous experiments, the third research question attempts to identify the relationship between numeracy and the production of accurate and complete gist responses.

8.2. Methods

In the present study we again asked learners and MTurkers to make gist assessments with three different types of LADs, positing that differences would be seen in the gist assessments according to visualization type. Using 3 new LAD types in this within- and between-subjects experiment, we compared the accuracy and completion of learners' gist assessments to those made by MTurkers. For each visualization type presented, participants were asked to describe the gist of each visualization after a 30 second exposure from the perspective of a fictitious student. The LADs were created using secondary learning data; they depicted 7-10 day discussions, similar to the real

discussion activity conducted in the previous study. Results were compared between learners and MTurkers, and between the three types of visualizations. This experiment was administered completely online. The final coding scheme from the previous experiment was used to analyze the data in this experiment.

8.2.1. Participants

Recruitment for this and the previous study happened simultaneously. After experiment 3 was filled, learners and MTurkers were then assigned to this study. Participants were not allowed to complete more than one study. Learners were solicited from first- and second-year university courses. Once permission was granted from course instructors, learners were solicited through an in-class presentation, and by email. There was some overlap in the study assignment between this and the previous one, since study solicitation took place over a longer period of time. Learners and MTurkers were assigned to the previous study, and then to this one after the previous one filled. The MTurkers sought for this study were North American residents who had participated in at least one online learning course utilizing LADs, who had not yet achieved a bachelor's degree.

MTurkers were required to pass a qualification survey before being invited to participate in the main study. Both studies were administered to MTurkers through TurkPrime, which then directed participants to the SurveyMonkey website. The HIT was deployed in small groups of 3- 6 surveys at a time to lower costs.

8.2.2. LAD stimuli

Secondary data from previous LA studies was used to create the visualization stimuli. Selected from 8 different discussion activities lasting 7-10 days, the data reflected multiple learning paths. These paths illustrated learners with different levels of participation and success in the learning activity. The timestamp and quality ratings from the secondary data were used along with fictitious student names to create the LADs. Using PlotlyChart Studio and Adobe Photoshop, 32 visualizations were created.

Grayscale and color versions of each visualization were produced, with the student of interest presented in color and the rest of their group and grayscale. Animated GIFs of each of the visualizations were created to be able to display them for 30 seconds on the SurveyMonkey platform. At the end of the visualization a single frame was displayed for 10 seconds that read "summarize the image in a 4 to 6 sentence paragraph." In the legend of each visualization the highlighted student's name was highlighted by framing it with a black line (for example, Emer in Figure 20). Figures 20, 21, and 22 are examples of the visualizations. In all of the visualizations the data of the fictitious student was presented in color. High coherence messages were green, medium coherence messages were yellow, and low coherence messages were red. The messages of their peers were presented at a lower opacity, so they appeared "greyed out." The grouping utilized in all of the visualizations was meant to show the daily totals of the group.

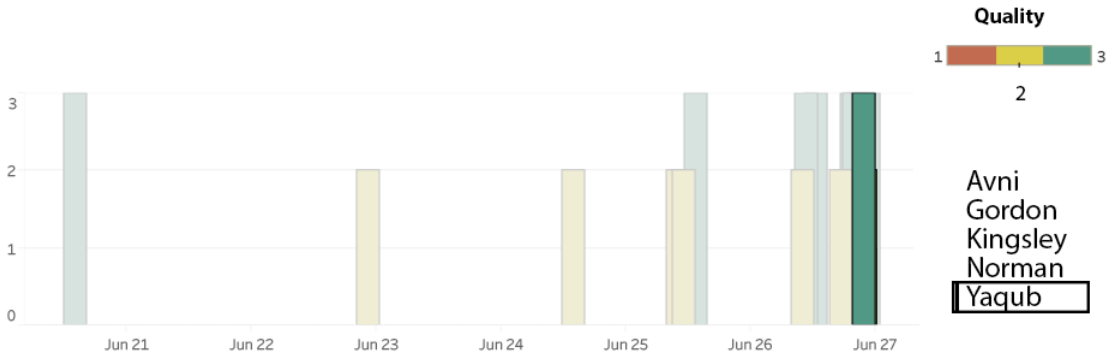

Figure 19. Exp. 4 bar chart visualization of Yaqub's performance

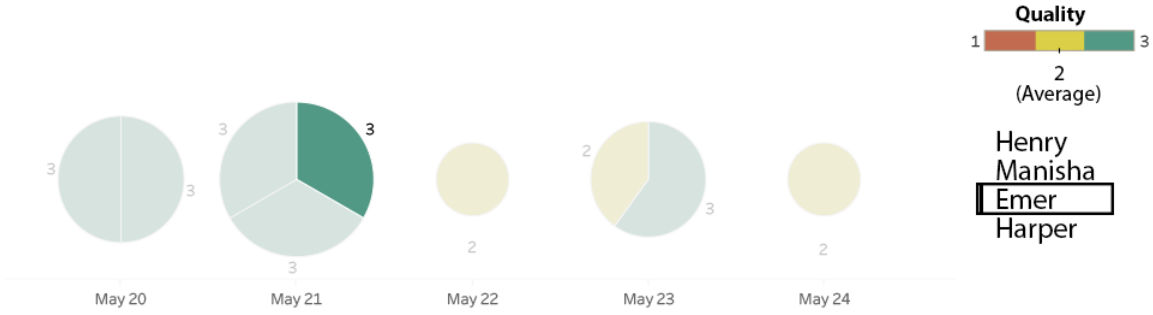

Figure 20. Exp. 4 pie chart visualization of Emer's performance

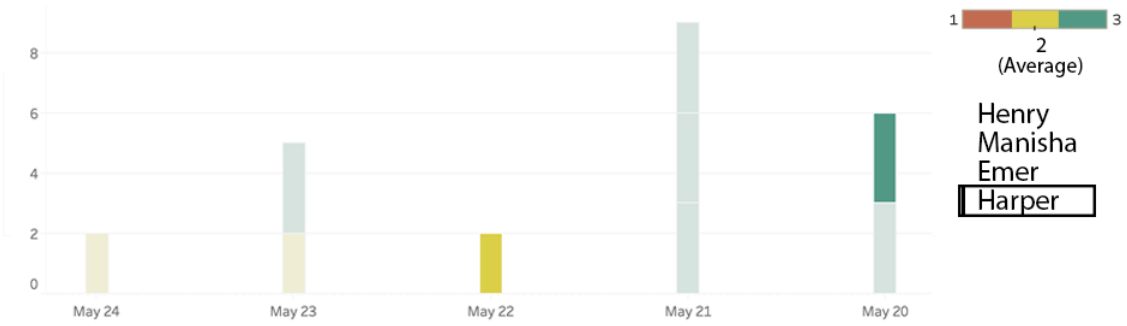

Figure 21. Exp. 4 stacked bar chart visualization of Harper's performance

8.2.3. Survey instrument

The survey was administered on SurveyMonkey. After completion of demographic information and the Subjective Numeracy Scale, participants received the following instructions:

> "You are about to begin the graph assessment part of this survey. All of the graphs represent small group discussions over time, with 3-6 students in each group. This task tests your understanding of abstract graphs. The objects in the graphs represent learning data; each has a high, low, or medium value. The position of the shapes, or data, within the graph could indicate relationships between them. The color image represents your performance, the other images represent the performance of your peers.
>
> Each graph will be displayed immediately when you access each question. When the graph has been displayed for 30 seconds it will automatically disappear. After you review the graph, write a paragraph that describes everything you see and understand from the graph. You then have as much time as you need for this part. If you need to take a break, do so before accessing the next question."

A button appears after this passage labeled "I'm ready to begin;" the visualizations began on the subsequent page.

8.2.4. Additional study instruments

Highly correlated with objective tests of numeracy, the 5-minute Subjective Numeracy Scale (SNS) objectively measures participants' numerical aptitude and preference for numbers over words (Fagerlin et al., 2007; Peters, 2012; Zikmund-Fisher et al., 2007). Here it is used to categorize participants' gist responses according to this

factor of individual difference. The qualification survey was used to assure that MTurkers met the qualifications for the study. In the survey we asked participants about the types of visualizations they were exposed to in everyday life, such as time management, health, utility and bill payment, banking, and educational applications. This additional information helped to obscure the qualifications of interest, while collecting information about participants' previous exposure to visualizations.

8.3. Procedure

This study repeated the procedure from the preceding study. Participants were shown all 32 visualizations in the same order; the visualizations themselves were ordered such that a different type of visualization was shown each time. Presenting the visualizations in the same order allowed us to mitigate the effects of potentially high dropout rates, and to visualize learning effects due to repeated exposure to the visualizations.

8.3.1. Data coding

Data was collected with SurveyMonkey, cleaned, then uploaded to NVivo to be hand-coded and analyzed. The six primary codes used in this study were identified in the previous experiment, using an open, emergent coding scheme (Given, 2008). Employing the same codes allowed for comparison across the two studies. Since the codes were developed in a previous study, this one utilized a selective coding scheme (Given, 2008). The primary codes pertain to gist — accurate, inaccurate, complete, incomplete, details of self, and overview. Responses could be coded as providing both details of self and overview. Responses were either accurate or inaccurate, complete or incomplete, but could not be coded as both. Additional codes used in the study pertained to the experience of the survey itself, uncertainty, or unique aspects of the response. The data were analyzed according to the coding scheme below (Table 17).

Table 17. Exp 3. coding scheme used in exp. 4

Code	Coding scheme examples
Gist – Accurate Gist-Inaccurate	Gist responses were either completely accurate, or inaccurate. Below is an example of an accurate response.
Gist-Complete Gist-Incomplete	These responses were either complete or incomplete. Below is an example of an incomplete response.
Gist-Details of self	These responses described just in terms of the highlighted person, describe the objects of the highlighted person, or remarked upon details of the performance of that person.
Gist-Overview	These responses focused primarily on the group as a whole. They tended to summarize performance in regard to the group, or to describe the performance of the highlighted individual in reference to the group.
Uncertainty	Responses with this code tended to express uncertainty in the instructions, or in the understanding or response for the given visualization.
Feedback	This code was used to separate portions of the response referring to the experience of the survey that did not pertain to the gist response.

8.3.2. Data analysis

The coded gist responses were visually reviewed for trends and patterns, such as which visualization type received the highest number of accurate and complete responses. The results of each individual participant were reviewed over time to see if patterns could be discerned in the frequency of accurate and complete gist responses produced. A t-test was performed to do a between-subjects comparison of the accurate and complete gist responses produced by learners and MTurkers, to see which group produced more accurate and complete responses. A oneway ANOVA was done to perform within-subjects comparisons of the accurate and complete gist responses according to visualization type within both groups. A post-hoc chi-squared analysis was then performed since the results were not normally distributed.

Results from this study were analyzed again with the results from experiment 3. Since the recruitment for these two studies happened on a rolling basis for both

participant groups, the responses may be treated as if both participant groups came from the same participant pools because essentially, they did. As such, it would be appropriate to perform a t-test on the combined results to see if there was a statistical difference between the groups, and ANOVAs to explore the effect of visualization type on the production of accurate and complete gist responses.

8.4. Results

Participants' (learners = 44, MTurkers = 30) demographic information is presented in Table 18. A large number of MTurkers were surveyed; the results of the screened MTurkers are presented first, followed by the subset of MTurkers who participated in this study. The mean age of MTurkers was 13 years older than learners, with a wider range of ages overall. The majority of MTurkers had some college experience but compared to learners, few of the MTurkers were enrolled in any type of higher education.

Table 18. Exp. 4 participant demographic information

	Surveyed Mturkers (N = 599)	**Mturkers (N = 30)**	**Learners (N = 44)**
Demographic Information			
Female	309	16	14
Male	291	14	30
Transgender, two-spirit, agender	7	0	0
Age range in years	18-72	20-72	18-30
Mean Age (SD)	35 (10.7)	34 (12.7)	21 (2.2)
Highest level of education			
High school degree or equivalent (e.g., GED)	76	2	28
Some college but no degree	141	18	11
Associate degree		9	1

	Surveyed Mturkers (N = 599)	Mturkers (N = 30)	Learners (N = 44)
Current enrollment			
Full time at a 4- year undergraduate college/university	54	4	36
Full time at a 2-year undergraduate college/university	12	0	4
Part time at a 4- year undergraduate college/university	24	2	3
Part time at a 2- year undergraduate college/university	14	3	
Not currently enrolled	442	21	

Due to an issue with the survey administration, learners skipped the Subjective Numeracy Scale questions, so we were unable to compare gist responses according to participants' numeracy.

The qualification survey asked participants about the visualizations they used in everyday life. To facilitate comparison of visualizations experienced between groups, results are reported as a percentage of each participant population. A large proportion of both participant groups used visualizations – many types of visualizations – in their everyday lives. Over half of MTurkers used planning, health, loan payment, telephone or internet, utility, educational and banking visualizations, and learners were not far behind. As seen in Table 19, only one MTurker (3%) and 7 learners (15%) had no exposure to any of the visualization types listed below. Banking graphs were the most commonly encountered visualization type with MTurkers, followed by educational graphs and health or exercise tracking graphs. For learners the most frequent visualization types were educational visualizations, followed by telephone or Internet usage, and health or exercise tracking visualizations.

Table 19. Exp. 4 prior visualization experience

Type of visualization	Mturkers	Learners
Banking graphs (ex. a graph of your checking account balance, bill payment graph)	27	18
Educational graphs (ex. grades for an online course, course that posts your participation information online)	22	35
Utility graphs (ex. electricity or gas usage, wood consumption)	21	13
Telephone or internet usage graphs	21	22
Loan payment graphs (ex. mortgage, student loans)	20	3
Time planning or tracking software graphs	16	20
Laboratory result graphs	13	14
Health or exercise tracking graphs	21	20
None of the above	1	7
Other (please specify) stock trading, matlab	1	1

8.4.1. Study completion rates

The completion rate for learners (61%) was lower than that of MTurkers (90%). While 30 MTurkers began the study, 27 completed it. For learners, 44 began the study and 27 completed it. Partial responses were included in the analysis. Three MTurkers and 2 learners dropped out of the study. The completion rate of learners was due to 15 learners attempting to game the system at some point during the survey. Normally this is done by the participant using a single response that might be possible for all the visualizations. A few learners did this, but the majority provided more nuanced ways to cheat the system than seen previously, by typing the same response for each but changing the name of the student of interest or copying and pasting text that was clearly unrelated to the survey into each of the 32 responses. For example, one person submitted a permutation of the following response for every visualization:

> "The data adds up. The name highlighted shows the person the data is about. The color corresponds to the scale that is shown on the top right. The contributions from each person add up. The total can be determined from the y-axis."

Another learner wrote the following sentence for each visualization, changing only the name of the student of interest, "each bar shows the total participation for the day, with each section of it representing one individual score." Betting that there would be at least one day with zero contributions, another learner submitted this for every visualization,

> "[T]he graph at the end has multiple observations for 5 people. For some days, there are zero observations. The maximum that goes is 3 while the lowest is either 1 or 0."

After some deliberation, responses such as the following were removed from the analysis only if a similar response was provided for each visualization:

> "Each block stacks up and adds up to a total value. The total amount can be determined by looking at the y-axis. Each person contributes to the total amount that differs each day. Each person also has a different color to represent them on the graph."

These responses described the visualizations accurately but did nothing to summarize the data within. If the learner previously provided acceptable responses, then only the repeated responses were removed from the analysis.

8.4.2. Accurate gist responses

MTurkers and learners provided gist responses that differed significantly in their descriptiveness. Reviewing the content of the gist descriptions revealed that learners had a greater tendency to describe visual aspects of the visualizations, or to compare the current visualization to the previous ones. Many of these responses read as if learners were asked to assess the visualizations on their aesthetic appeal or relative merits, rather than assessments of gist. Learners also tended to provide longer answers – fulfilling the directive to provide 4-6 sentence responses – without making any summative judgements of gist from the perspective of the highlighted student.

MTurkers and learners had similar rates of accuracy on all three visualizations. The rates of accuracy for the bar chart, pie chart, and stacked bar chart are displayed below according to participant group and visualization type (Table 20). Accuracy is

reported as a percentage of the total respondents for each visualization, since this the number of respondents varied per visualization. Total accurate responses over 25%, the highest seen in this study, are highlighted. Accuracy is just part of a gist response however – an accurate but incomplete response could omit much of the detail required to assess performance, since the accuracy only refers to what was stated in the response. For this reason, we next reviewed the responses that were both accurate and complete.

Table 20. Exp. 4 accuracy by visualization type for learners and MTurkers

Bar chart visualization			**Pie chart visualization**			**Stacked bar chart visualization**		
	Learner	Mturker		Learner	Mturker		Learner	Mturker
01 AlishaB	13%	15%	02 AvniP	13%	15%	03 BainsS	8%	20%
07 ColinB	18%	7%	05 BrittonP	8%	23%	04 BrendaS	26%	20%
11 EmerB	34%	23%	08 DaniaP	20%	23%	06 ChurchS	6%	7%
15 GuyB	24%	23%	12 FranklinP	30%	27%	09 DerekS	24%	30%
18 KayleyB	3%	10%	14 GordonP	24%	23%	10 EamonS	19%	20%
21 LidiaB	0%	7%	16 HarperP	22%	23%	13 GiuliaS	21%	23%
23 MaganaB	4%	7%	17 HenryP	21%	30%	19 KingsleyS	7%	17%
26 PachecoB	27%	17%	20 LevisonP	0%	7%	24 ManishaS	0%	7%
29 RoyB	15%	17%	22 LilithP	3%	10%	27 PattersonS	7%	13%
32 YorkB	11%	23%	25 NormanP	4%	7%	30 SalaS	31%	17%
			28 RhiannonP	14%	27%			
			31 YaqubP	29%	30%			

8.4.3. Accurate and complete gist responses

Accurate and complete gist response results are reported as percentages of the total responses provided for each individual visualization (Table 21); to differentiate the results, the highest percentage accurate and complete responses are highlighted. Across

all visualization types, MTurkers provided more accurate and complete responses than learners.

Table 21. Exp. 4 accurate and complete gist responses learners and MTurkers

Bar Chart Visualizations			Pie Chart Visualizations			Stacked Bar Chart Visualizations		
	Learner	Mturker		Learner	Mturker		Learner	Mturker
01 AlishaB	3%	7%	02 AvniP	5%	11%	03 BainsS	3%	7%
07 ColinB	3%	3%	05 BrittonP	5%	10%	04 BrendaS	3%	20%
11 EmerB	3%	3%	08 DaniaP	3%	3%	06 ChurchS	0%	3%
15 GuyB	3%	7%	12 FranklinP	0%	10%	09 DerekS	0%	13%
18 KayleyB	3%	10%	14 GordonP	3%	13%	10 EamonS	3%	10%
21 LidiaB	0%	7%	16 HarperP	3%	7%	13 GiuliaS	3%	3%
23 MaganaB	4%	7%	17 HenryP	4%	13%	19 KingsleyS	7%	17%
26 PachecoB	7%	10%	20 LevisonP	0%	7%	24 ManishaS	0%	7%
29 RoyB	0%	10%	22 LilithP	3%	10%	27 PattersonS	0%	7%
32 YorkB	4%	10%	25 NormanP	4%	7%	30 SalaS	3%	7%
			28 RhiannonP	0%	10%			
			31 YaqubP	7%	13%			

An a priori statistical power analysis was performed with G*Power software (Faul et al., 2007) to estimate the required sample size. Using a medium effect size of 0.5 according to Cohen's (1988) criteria for a t-test with alpha = .05 and power = 0.90, the projected sample size would need to be approximately N = 34. For power = 0.80, the projected sample size would be N = 26. The participant sample size met this criterion. A 2-sided t-test[16] was conducted to compare the total responses that were both accurate and complete between the two participant groups (Table 22), learners and MTurkers.

[16] The test conducted was the Aspin-Welch-Satterthwaite-Student's t-test using JMP 15. The Student's t-test was adapted to work with nonequal group variances.

There was no significant difference between the responses of MTurkers (M = 2.8, SD = 7.54) and learners (M=0.64, SD= 3.33) t(37) = 1.48, p = 0.15. There was not enough evidence to reject the null hypothesis, that no difference existed between these two populations.

Table 22. Exp. 4 accurate and complete gist means by visualization type

	Learner	***Mturker***
Bar chart visualization	M = 0.21 (SD = 0.17)	M = 0.71 (SD = 0.46)
Pie chart visualization	M = 0.27 (SD = 0.17)	M = 1.10 (SD = 0.46)
Stacked bar chart visualization	M = 0.18 (SD = 0.17)	M = 0.90 (SD = 0.46)

A oneway analysis of variance (ANOVA) was performed to make comparisons of the number of accurate and complete gist responses produced by visualization type within each participant group, to determine if any of the means differed from the others for each of the participant groups. There were no statistically significant differences between group means for MTurkers ($F(2, 92) = 0.18$, $p = 0.83$) or for learners ($F(2, 131) = 0.08$, $p = 0.93$) as determined by one-way ANOVA.

These results were not normally distributed. The distribution of accurate and complete responses for learners had a skewness of 0.04 and kurtosis of -1.49. For MTurkers the skewness was 0.26 and kurtosis was -1.13. The skewness for both participant groups was acceptable, but the kurtosis values for both were less than -1. This meant both distributions were too flat, making the distributions non-normal (Hair et al., 2017, p. 61). Viewing the graph of the results confirmed that they were not normally distributed, making the assumption of normality not viable, so a follow up nonparametric test was conducted.

To determine if the provision of accurate and complete responses was related to visualization type, we performed a chi-square analysis for two or more independent samples. Assuming independence between visualization type and response type, we put forth the following hypotheses for both groups of participants, learners and MTurkers:

H_0: Visualization type has no relationship to the provision of accurate, complete responses

H_1: Visualization type is related to the provision of accurate, complete responses

The two categories of responses used for the analyses were 1) accurate and complete, and 2) inaccurate, incomplete, or inaccurate and incomplete. Using the chi-squared values from the contingency tables below (Tables 23 and 24), 2 degrees of freedom, and the values from 0.05 probability from the chi-squared critical value table:

For learners, $\chi 2(2, n = 512) = 1.9, p < 0.05$

For MTurkers, $\chi 2(2, n = 798) = 4.5, p < 0.05$

Since we could not reject the null hypothesis, that the visualization type would have no relationship to the provision of accurate complete responses by either learners or MTurkers, these results supported our earlier hypothesis that participants would perform better with some types of visualizations.

Upon visual inspection, MTurkers seemed to have performed better with the pie chart than the other visualization types. To test this, a chi-squared analysis for 2 or more independent samples was performed, with the following hypotheses for each participant population:

H_0: visualization type has no relationship to accurate and complete responses

H_1: Visualization type is related to accurate and complete responses

The groups for this test were the visualization type and response. The categories of responses are 1) accurate and complete, and 2) inaccurate, incomplete, or inaccurate and incomplete.

Using the contingency tables (see Tables 23 and 24) for each participant group, an alpha value of 0.05 and 2 degrees of freedom, the critical value is 5.991. For comparison, if alpha = 0.10 the critical value would be 4.605.

For MTurkers, $\chi 2(2, n = 954) = 1.6, p < 0.05$

For learners, $\chi 2(2, n = 1{,}042) = 0.7, p < 0.05$

Table 23. Exp. 4 MTurker contingency table for accurate and complete responses for bar, stacked bar, and pie visualizations

Mturker Contingency Table			
	Accurate and complete responses	Inaccurate, incomplete, or inaccurate and incomplete	Total responses
Bar visualization	22	253	297
expected	20	244	
	16.12528588	0.303565277	
Pie visualization	34	284	357
expected	31	294	
	0.209472468	0.324130436	
Stacked bar visualization	28	248	300
expected	26	247	
	0.09509434	0.005307695	
Totals	84	785	954
so $x^2 = 1.6$			

Table 24. Exp. 4 learner contingency table for accurate and complete responses for bar, stacked bar, and pie visualizations

Learner Contingency Table			
	Accurate and Complete responses	Inaccurate, Incomplete, or Inaccurate and Incomplete	Total responses
Bar viz	9	267	316
expected	8	267	
	0.03046765	6.20E-05	
Pie viz	12	330	392

Learner Contingency Table			
	Accurate and Complete responses	Inaccurate, Incomplete, or Inaccurate and Incomplete	Total responses
expected	11	331	
	0.20414319	0.00336627	
Stacked bar viz	7	283	334
expected	9	282	
	0.43462882	0.00304689	
Totals	28	880	1042
so $x^2 = 0.7$			

Since chi-squared for both learners and MTurkers was less than the critical value we accepted the null hypothesis for both, concluding that for both populations, there was no relationship between visualization type and the provision of accurate and complete responses. Given the brief amount of time the participants had to review gist, we reasoned that the responses may tend to be incomplete, more than if participants had more time to view the visualizations. If completion was excluded, would accuracy be affected by visualization type? We repeated the chi-squared analyses for the provision of accurate responses only, with the following hypotheses:

H_0: visualization type has no relationship to accurate responses

H_1: Visualization type is related to accurate responses

Using the contingency table for only accurate responses in each population (Tables 25 and 26), 2 degrees of freedom, and an alpha value of 0.05:

Again the chi-squared for both learners and MTurkers was less than the critical value so we accept the null hypothesis, concluding that there was no relationship between visualization type and the provision of accurate and complete responses in either population. This left RQ1, which sought to determine if learners produced more accurate or complete descriptions of gist, unsupported. In regard to RQ2, MTurkers produced

more accurate and complete gist responses than learners. RQ3 could not be determined since the numeracy data for learners was not collected.

Table 25. Exp. 4 contingency tables for accurate responses made by MTurkers for bar, stacked bar, and pie visualizations

Mturker Contingency Table			
	Accurate	Inaccurate	Total responses
Bar viz	44	253	297
expected	53	244	
	1.41005173	0.30356528	
Pie viz	73	284	357
expected	63	294	
	1.50557628	0.32413044	
Stacked bar viz	52	248	300
expected	53	247	
	0.02465409	0.0053077	
Totals	169	785	954
so $x^2 = 3.6$			

Table 26. Exp. 4 contingency tables for accurate responses made by learners for bar, stacked bar, and pie visualizations

Learner Contingency Table			
	Accurate	Inaccurate	Total responses
Bar viz	49	267	316
expected	49	267	
	0.00033662	6.20E-05	
Pie viz	62	330	392

Learner Contingency Table			
expected	61	331	
	0.01828591	0.00336627	
Tree viz	51	283	334
expected	52	282	
	0.01655103	0.00304689	512
Totals	162	880	1042
so $x^2 = 0.04$			

8.4.4. Visual analysis of learning progression

To determine if learning effects could be observed in this study, similar to the previous study, we created cell plots for each participant grouped by population (see Figures 23 and 24). Each visualization has four cells – one each for accuracy, completion, details of self, and gist overview. Positively valued cells are green and negatively valued cells are white. The cell plots' order reflects the order the visualizations were presented in the survey. The cell plots were created using JMP 15, then reorganized and presented using Adobe Photoshop. Participants were then ordered according to when their first accurate response was observed.

In Figures 23 and 24, participants in Group 1 provided no accurate responses on the survey before they quit. This group included 3 learners (7%) and 8 MTurkers (27%). Since most participants in this group dropped out by the fifth LAD, there is no way to say if they would have accurately assessed the LADs later. This was the smallest group of participants for both learners and MTurkers.

Participants in Group 2, 27 learners (61%) and 13 MTurkers (43%), provided at least one accurate response within the first 5 visualizations. Though this was the biggest group for both sets of participants, the patterns observed in these groups differed.

MTurkers who provided an inaccurate response within the first 5 tended to provide accurate responses repeatedly throughout the survey. These participants were the most successful with the survey overall, producing the most accurate and complete responses. This was also the case for the 9 learners (20%) who provided more than five accurate responses. The 5 most successful learners (11%) produced more than 10 accurate responses.

Group 3 participants made their first accurate response between the 5th and 10th LAD; the fourth group provided their first accurate response sometime after the 10th LAD. The third group (4, 13%) and fourth group (5, 16%) were composed only of MTurkers. In both groups, the provision of accurate and complete responses was sporadic. Visually the responses of this group matched the learner subset of Group 2 that provided more than five accurate responses.

Figure 22. Exp. 4 learning effects cell plots for MTurker responses (missing 3 people who viewed but did not answer the first response)

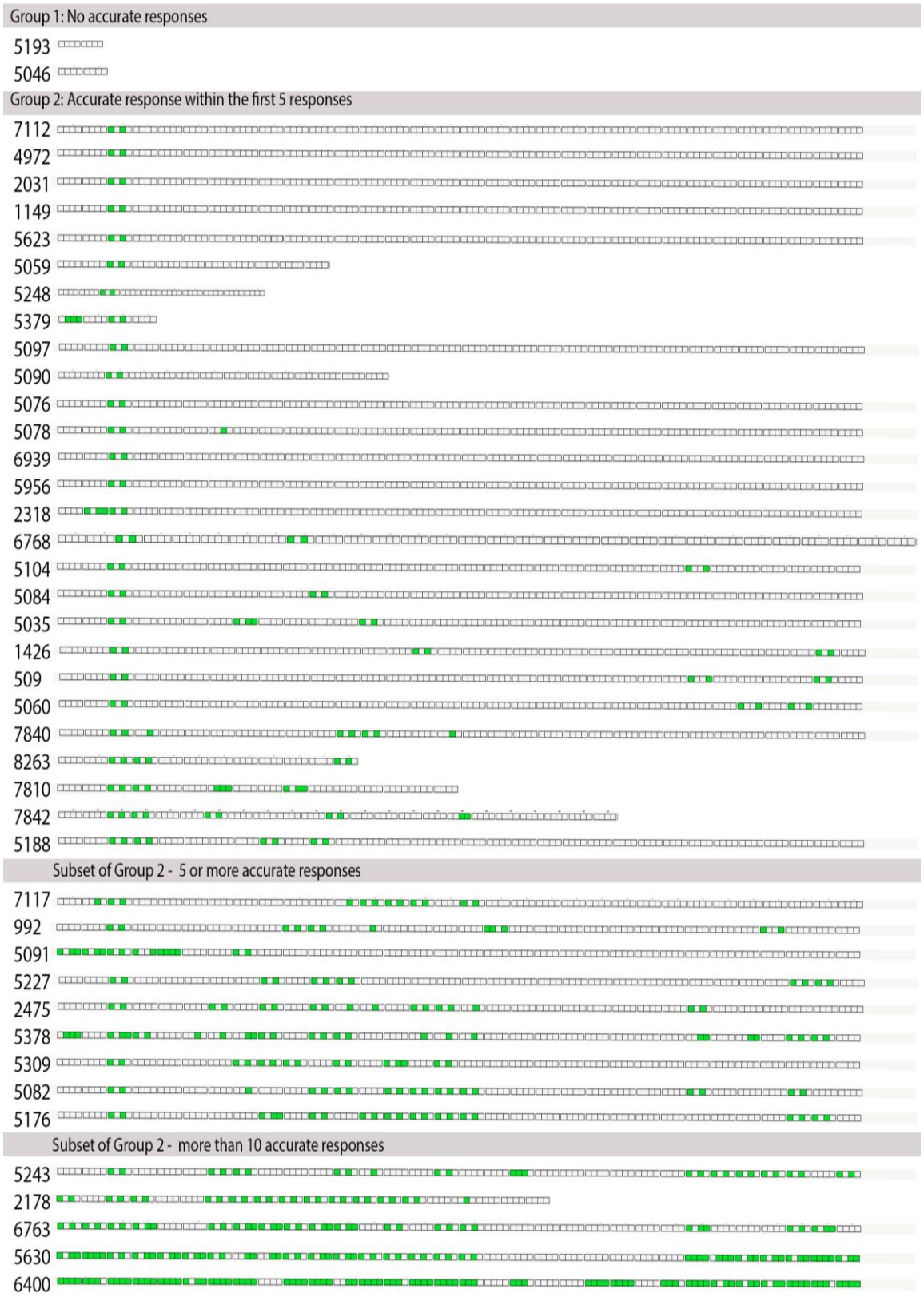

Figure 23. Exp. 4 learning effects cell plots for learner responses

8.4.5. Combined analysis of experiments 3 & 4

Since the responses were produced by participants from the same groups, it could be assumed that the within-group means would not be statistically different. Combining the total number of accurate and complete responses for experiments 3 and 4 would yield learners (N = 60) and MTurkers (N = 54), well over the estimated total sample size of 26 as calculated previously. A two-tailed t-test was performed. There was a statistically significant difference t(77) = 2.94, p = 0.004 between MTurkers (M = 1.70 SD = 2.77) and learners (M = 0.44, SD = 1.39).

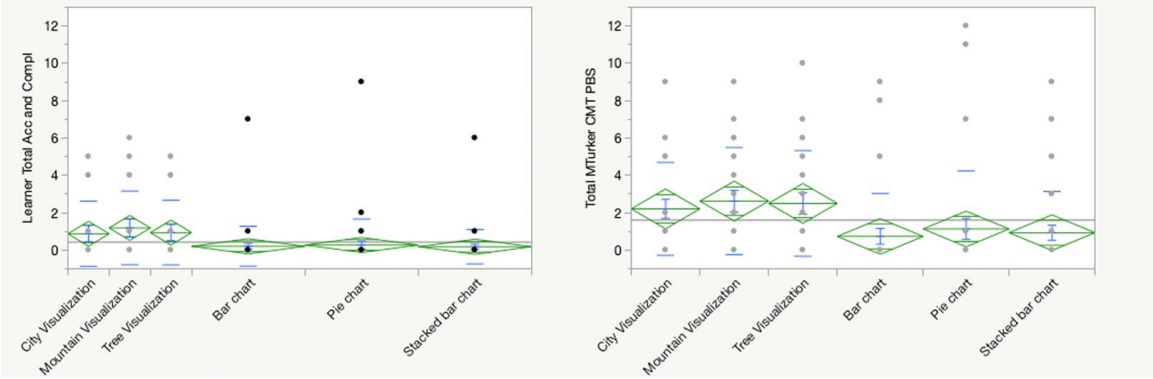

Figure 24. Exp. 4 oneway ANOVA of accurate and complete gist responses by visualization type for learners (left) and MTurkers (right)

A oneway analysis of variance (ANOVA) was performed to compare the accurate and complete responses produced by each participant group. For learners a statistically significant difference, (F(5, 179) = 2.47, p = 0.03), was seen between all of the visualization types listed in Table 28. A post hoc sensitivity power analysis computed using G*Power software (Faul et al., 2007) to estimate the effect size of this result with alpha = .05 and power = 0.90 and learners (N = 60) yields an effect size of 0.55. For an ANOVA, the effect size benchmarks for eta squared are 0.14 small, 0.25 for medium, and 0.4 for a large effect size (Cohen, 1988).

Table 27. Exp. 4 visualization means from combined analysis

	Learner	MTurker
Bar chart visualization	M = 0.20 (SD = 0.20)	M = 0.73 (SD = 0.48)
Pie chart visualization	M = 0.27 (SD = 0.20)	M = 1.10 (SD = 0.46)
Stacked bar chart visualization	M = 0.18 (SD = 0.92)	M = 0.93 (SD = 0.8)
City Visualization	M = 0.88 (SD = 1.75)	M = 2.21 (SD = 0.54)
Mountain Visualization	M = 1.19 (SD = 1.97)	M = 2.63 (SD = 0.54)
Tree Visualization	M = 0.94 (SD = 1.73)	M = 2.5 (SD = 0.54)

For MTurkers a statistically significant difference, ($F(5, 161) = 2.74$, $p = 0.02$), was seen between all of the visualization types listed in Table 27. A post hoc sensitivity power analysis computed using G*Power software (Faul et al., 2007) to estimate the effect size of this result with alpha = .05 and power = 0.90 and MTurker (N = 54) yields an effect size of 0.58. For an ANOVA, the effect size benchmarks for eta squared are 0.14 small, 0.25 for medium, and 0.4 for a large effect size (Cohen, 1988).

Visually reviewing the results from both studies in Table 27 raised another question. Is there a difference between the means of the abstract visualizations from experiment 3 and the more traditional visualizations of experiment 4? For each participant group the city, mountain, and tree visualizations were combined into a group called abstract visualizations. The other visualizations were combined in a group called traditional visualizations. Then a one-tailed t-test was performed for both groups, since a visual inspection of the means indicated that the production of accurate and complete responses was higher with the abstract visualizations. For learners there was a statistically significant difference $t(61) = -2.83$ $p = 0.003$ between the abstract visualizations ($M = 1.79$, $SD = 0.26$) than the traditional visualizations ($M = 0.22$, $SD = 1.13$). For MTurkers there was a statistically significant difference $t(148) = -3.63$ $p = 0.0002$ between the abstract visualizations ($M = 2.44$, $SD = 2.7$) and the traditional visualizations ($M = 0.93$, $SD = 2.55$).

8.5. Discussion

The research questions for this study aimed to determine if, using these visualization types, differences would be seen in the gist responses made by learners and MTurkers. The coding of gist established in the previous study was carried forward to this study, repeating the details of self and overview codes to see if one or both of these components of gist description was more commonly associated with accurate or accurate and complete responses. So few accurate or accurate and complete responses were provided that we could not explore this aspect of gist descriptiveness. Likewise, since learners' numeracy was not captured, we could not explore the relationship between accuracy and numeracy within and between participant populations.

In this study, as with experiment 3, MTurkers provided more accurate and complete gist responses than learners. Learners provided fewer actual assessments of gist in their responses than MTurkers, instead choosing to comment on aesthetic aspects of the visualizations. A few learners even commented on the merits of the individual types of visualizations. One learner had this to say about one of the pie chart visualizations:

> "To me, it [the pie chart visualization] seems like a misleading representation, because it can hide the fact the group wasn't very productive overall each day. By having a single pie with a minimal amount of digits around it, it draw attention away from the fact that this group hardly contributed to this project over the course of the week, but that is nonetheless beneficial when trying to misrepresent negative data."

Though the observation was insightful, this learner skipped the more challenging task of analyzing the LAD from the perspective of the fictitious student. This learner population's tendency to avoid providing complete gist descriptions could reflect a lack of motivation or confidence, the desire to avoid effortful thought, or miserly information processing (Toplak et al., 2013). Learners attempted to game this study more often, and in more nuanced ways. In the provision of nuanced responses— changing the name of the fictitious student in each response or slightly altering each response provided — these learners were not saving time, as this required a greater cognitive expenditure than just copying and pasting the same response repeatedly. These learners were attempting to expend less cognitive energy than it would've taken to actually do what the study required.

It was surprising that neither participant group demonstrated higher levels of proficiency with any of the three new visualization types. Thinking gist assessments would be more accurate with familiar visualization types than those made with the abstract visualizations used in the previous study, we hypothesized that participants would perform better with bar chart visualizations, since they are the most commonly experience visualization type. Both participant groups had a good amount of exposure to everyday visualizations, bolstering this hypothesis. When using LADs to assess their performance during discussion activities, learners commonly make part-whole comparisons while comparing their work to that of their peers. The visualization types used in this study — the bar chart, pie chart, and stacked bar chart — have all been studied for their facilitation of estimations of proportion (Few, 2007; Spence & Lewandowsky, 1991). Familiarity and the facilitation of part-whole comparisons did not result in a higher number of accurate and complete gist responses being produced with bar chart visualizations.

In both participant groups, gist accuracy was lower in this experiment than the previous one. When comparing the results of experiments 3 and 4 we saw the production of more accurate and complete gist responses with the abstract visualizations than the traditional ones. Our findings lend support to the body of information visualization research that claims that aesthetic appeal aids sensemaking (Bateman, 2010; Berlyne, 1970; Lim et al., 2007; Miniukovich & De Angeli, 2015). Similar to the findings of Gillian and Sorensen (2009), it may be that that the embellishment aided the popout effect, making target features easier to locate (Treisman & Gelade, 1980). It is also possible that participants paid more attention to visualizations because of their novelty or aesthetic appeal (Chatterjee & Vartanian, 2014).

Chapter 9.

Experiment 5 – Stability of LAD-based mental models

9.1. Experiment design and procedure

LADs provide feedback on overall learning strategy enactment, including the progress learners have made toward their short- and long-term goals. For LADs to help learners assess their performance-related data however, they must first be understood. As is evident from the preceding studies conducted as part of this dissertation, it cannot be assumed that learners' gist assessments are accurate. In either case, accurate or not, it is important to gain greater insight into the factors that shape learners' conceptualization of gist. Based on the previous studies comparing the gist assessments of learners and MTurkers, it is possible that the inclusion or exclusion of peer comparisons in gist assessments may lead to a tendency towards inaccuracy. It is yet to be seen if human factors such as numeracy or goal orientation also play a part in this tendency. An incorrect gist assessment presents another interesting case. If learners recognize that their perception of their own performance differs from what is depicted in the LAD, which one do they believe? Further, which perspective is reflected in their subsequent learning strategy enactment?

This experiment was conducted to determine the role LADs played in shaping learners' mental models of their performance during an ongoing learning activity, and to see if these mental models persisted. Specifically, we wanted to better understand 1) the mental models formed through interactions with LADs, 2) if these models persisted, 3) if and how they were augmented by repeated exposure, and 4) if these models influenced learners' subsequent learning strategies. The research questions addressed were:

- RQ1: What role did the LADs play in shaping learners' mental models of their participation in the learning activity?
- RQ2: Exemplifying these mental models, did the gist gleaned from the LADs change, or was it persistent over the course of the learning activity?

- RQ3: Were learners' gist estimates accurate? To what extent were these assessments shaped by peer comparison – which is visually prioritized in the LADs – or human factors such as numeracy or goal orientation?
- RQ4: What did learners do if their mental models did not match what was depicted? Did they believe the LADs, or their own perspectives?

Since the LAD's design prioritized social comparison, we hypothesized that learners would more often describe their performance in comparison to their peers. Further we posited that these peer comparison-based gist assessments would be richer and more accurate than those based solely on individual performance.

As learners metacognitively monitor their learning with LAD, their perceptions of themselves as learners may shift toward assessments of performance achieved through social comparison. Taking the position that academic accomplishment is predicated by first, a person's own beliefs about themselves (Bandura, 1977; Zimmerman & Bandura, 1992), we acknowledge that participants' perceptions of themselves as learners may be more persistent than what is depicted by LAD.

9.2. Methods

This empirical inquiry was both in depth and in-situ, in that learners' responses were based on their actual experience using LADs over the course of a week-long learning activity. First, learners participated in small group discussions, using a LAD designed for this experiment. Shortly after the conclusion of the learning activity, learners were invited to participate in semi-structured qualitative interviews. Performing the interviews soon after the learning activity provided a way to interrogate learners' mental models of their performance formed with the LAD without interrupting, and possibly negatively affecting, the learning process. Learners were questioned about their internally held goals, motivations, and self-concept. Along with the duration, format, and subject matter of the discussion activity, all of the aforementioned factors contributed to gist. Trace data —consisting of message counts, their timing, and their coherence ratings— were collected from the discussions. This data was used to re-create the LADs viewed during the learning activity.

Retrospective research methods (Eger et al., 2007; Pätsch et al., 2014; Frith & Harcourt, 2007; Harper, 2010) were employed to help participants recall their learning strategy enactment with LADs. Specifically, learners were asked to verbalize the mental models of learning that proceeded their task accomplishment, as well as the mental models of learning they had before, during (i.e. gist), and after interactions with the LADs. The re-created LADs – identical to what was seen by students during the process of learning – were used to guide the qualitative in-depth interviews, prompting detailed gist descriptions and helping learners to recall their learning strategies. All the interviews were conducted on and recorded using the Zoom[17] online platform. The video recordings were deleted after the interviews were transcribed. Open, axial, and selective coding was performed on each transcribed interview. The multiple sources of data collected in this study were meant to comprehensively address the learning context and the phenomena of interest.

9.2.1. Participants

Participants were undergraduate students enrolled in an art and science-based program of study at a 4-year university. They were solicited from second-year undergraduate courses offered at a Canadian university. A short 5-minute description of the study was presented on Zoom the day the discussion activity started. Participants who used the LAD at least once and completed the study consent form were invited to participate in the interviews.

9.3. LAD stimuli

Given the difficulty observed in the previous studies, the LAD designed for this study was simplified in order to facilitate rapid gist assessments (see Figure 25). The quality of learners' message posts was visualized according to its coherence rating and the date and time of the post. Ratings were high, medium, and low, according to the

[17] Zoom.us

thresholds identified in previous research on cognitive presence (Garrison et al., 2001). The learners' personal data points were displayed in orange. The data points of their small group members were all displayed in gray, making it easier to differentiate the two. Small group members' data points were purposefully not labeled, to focus the learners' attention on their data relative to that of the group as a whole. A line was drawn that connected message data points, according to the time and date that they were posted. This was done to reinforce the idea that the messages were connected, that each message was part of an ongoing conversation. The title of the visualization stated, "Quality of Posted Messages: Message Coherence." Directly beneath this was a definition of coherence that read "coherence is a measure of how well your message is written, based on the requirements of the assignment." The unpopulated visualization had another message written across it that said, "No messages posted yet."

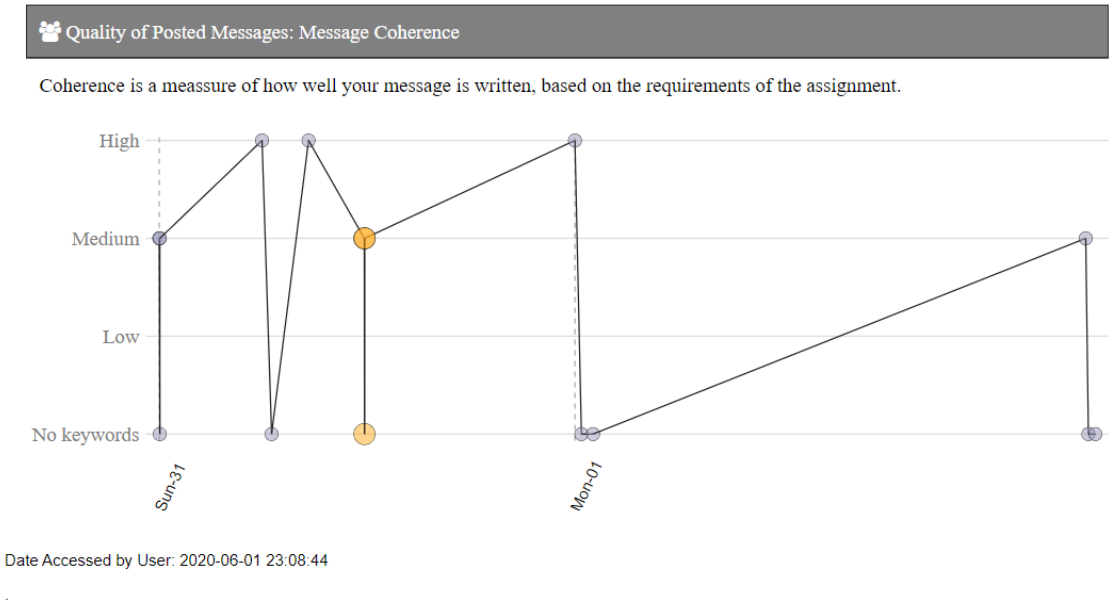

Figure 25. Exp. 5 example of LAD as seen by P4 when the LAD was first accessed

9.4. Interview protocol

The interview protocol acknowledged the learners as the experts of their lived experience and co-creators of the knowledge produced in this experiment. In the development of the interview protocol, the ways in which presentation may influence the

learners' interpretation of the research or the researcher's intentions were carefully considered. Take for example, preparation for the Zoom-based interviews. Zoom meetings tend to take place in the home and are often more casual than face-to-face meetings. When conducting the participant solicitation and interviews, care was taken to dress professionally and ensured the visible background mimicked an office-like set up.

The solicitation script reinforced the idea that learners did not have to achieve a high level of proficiency with the LAD for their feedback to be valuable. Learners were told that if they had difficulty with the LAD, that the issue was likely a fault in the design rather than their own. This choice of language served to make learners more comfortable, and reinforced the value of their feedback, regardless of the positive or negative nature of their experience with the LADs. It hopefully encouraged participants who did not understand the LAD to participate in the study.

In preparing for an interview it is important to address one's own biases as a researcher, including a priori assumptions about participants (Chenail, 2009), to consider the current social positioning of myself and the participants, relative to one another. While formulating solicitation materials and interview questions, I considered the positive and negative "interview identities" that I as a researcher brought to the experience. This was reflected in the choice of language used in both the solicitation and interview scripts. In the solicitation I recognized that they, as learners, had the feedback that I needed for my own research. I did this to recognize their contribution in a way that more evenly balance the perceived power variance between our roles as researcher and potential participants. In the interview I was very careful to avoid language that would prioritize one kind of experience over the other, for example, taking care not to ask participants if their experience was "normal." Doing this would imply that a certain type of experience was expected, and therefore valued, more than another. Neutral language was used as much as possible to validate the participants' subjective experience, and to value feedback shared about positive and negative experiences equally.

To prepare for my online presence when conducting the interview, I observed model interviews and reviewed methods for conducting behavioral research online (Hughes, 2012; O'Connor & Madge, 2017; O'Connor et al., 2008). After drafting the

interview questions, I wrote a personal subjectivity statement. It explored my position as a researcher and as an educator, and how these positions were reflected in the interview language. The majority of the interview questions were revised as a result of writing the subjectivity statement.

Different interview approaches yield rich data about the phenomenon of interest (Bourgeault et al., 2020; Potter & Hepburn, 2012; Kvale, 2020; Houtkoop-Steenstra, 2000; DiCicco-Bloom & Crabtree, 2006; Holstein & Gubrium, 1995). Interview questions also do more than simply aid in data collection – they can help to promote a mutually beneficial interaction between researcher and participant. For many of the participants, this may be the first time that they have been asked to articulate feelings and experiences related to their learning. This may be difficult or overwhelming. For individuals who identify as poor students, or who experienced difficulties within their groups or with the learning activity, or who felt that they were unsuccessful with the LAD — asking them to relive negative experiences may feel particularly vulnerable. To make them feel comfortable, I used nonjudgmental language and gently probing follow-up questions of varying levels of complexity (Chenail, 2009).

Beyond working to develop feelings of trust, comfort, and safety for participants (Oakley, 2015; Smith, 2013), the interview had to meet research goals of reliability, response validity, and construct validity. Matching the interview methods to the research questions increased construct validity and my ability to infer that the interview responses adequately describe participants' experience using the LAD. Careful language choice, proper question sequencing, and pilot testing the interview script helped to ensure that participants understood the questions as intended and answer according to their real feelings and experience. Finally, agreement across participant responses spoke to reliability.

Participants' presentation of self in these interviews may be aspirational, rather than realistic. There are challenges inherent in answering questions about one's perception of self, especially if the questions touch upon identities important to the individual being questioned. Stated another way, if the participant holds dear the perception of themselves as successful learners, no matter what actually transpired during

the learning activity or what they themselves know to be true (i.e. what actually transpired), when questioned the person may adhere to the perception of themselves as a success and answer accordingly. From this it is clear then, that any deception employed during interview may not carry malicious motives, but instead may stem from a desire to leave certain perceptions of self unchallenged. It is also possible that learners simply may not have the knowledge to recognize what they did not know (Kruger & Dunning, 1999; Serra & DeMarree, 2016).

9.4.1. Interview

Though there are no "set in stone" standards for interview methods, there are guidelines for the many types of qualitative interviews (Bourgeault et al., 2020). The choice of interview method as well as the actual interaction that transpires, both influence the knowledge gleaned from the interview. This study blended approaches from conceptual and biographical narrative interviews (Kvale, 2020), within a semi-structured interview format. The level of scripting was due, in part, to the number of anticipated participants. The semi-structured interview format allowed for flexibility; the scripting eased comparison across multiple cases (Bourgeault et al., 2020). The qualitative interview was undertaken as a joint endeavor to explore the phenomenon of interest. In this instance knowledge was co-created. The combination of the conceptual and narrative interview was best to explore the conceptual, i.e. mental models, while seated in the context of the individuals' lived experience. In narrative interviews the researcher may ask questions to elicit detail, but the primary focus is to support participants' storytelling and listening as the story unfolds (Kvale, 2020). The interview itself, the interaction involved, influences the knowledge gleaned from it (Bourgeault et al., 2020).

Participants were asked to tell the story of their experience using LAD, giving detailed accounts of their experience – including their motivations, understandings, and subsequent behaviors – including any aspects they felt contextually relevant (Mishler, 1986).These interviews were biographical narratives in the sense that learners are speaking to their own lived experience (Wengraf, 2020). They created narrative knowledge in that they included, as described by Yin, "[A] sequence of events ('I did this

then she said that …') that allow the person to organize experience in a way that reflects human purpose and intentionality ('… and then I walked out because …'), and also to evaluate it (the 'moral' of the story)." (Yin, 2009, p. 18)

Part of the interview process entailed transcribing and analyzing each interview as soon as it was completed. This helped me improve the techniques employed in subsequent interviews, and is a good reflective practice as a researcher (Schon, 1983). Coding conventions were employed that allowed for note sequencing, researcher comments, and the inclusion of speech production characteristics such as changes in pitch, amplitude, and intonation (ten Have, 1999). The paralinguistic elements – elements such as gestures or body language, as well as changes in tone, pitch, or speed— provided another layer of meaning that could support or refute the verbalized responses. Their inclusion supported the researchers' own inference generation, by lending credence to what was said (Wengraf, 2011). The inclusion of session notes further contextualized the questions and responses received.

9.4.2. Interview questions

In writing the interview script (see Appendix A), efforts were made to establish rapport and put the participant at ease throughout the interview. At the beginning of the interview participants were thanked and reminded that their feedback would be used to improve LA for learners. Participants were informed that some of the questions would be asked multiple times, in multiple ways, in an effort to truly understand their perspective. This set expectations while attempting to mitigate misinterpretation, such as the assumption that the provided responses were "wrong" if a question was asked multiple times.

Going in the order of each LAD interaction, each of the prepared questions stemmed from one or more of the research questions. Initially broad in nature, the questions narrowed to focus on the participant experience during the learning activity. The first part of the interview was a warm-up that addressed numeracy, giving the participants a chance to get comfortable with being interviewed using a low stakes

question. Using their survey results, participants were asked to give context for their responses on their perceived numerical ability and preferences.

The second section of the script contained questions about online discussions. Standardized questions from the Achievement Goal Orientation Framework (Elliot & McGregor, 2001) associated with mastery approach, mastery avoidance, performance approach, and performance avoidance (see line 15 Appendix A) were used to facilitate participant categorization according to achievement goal orientation.

According to the question order effect, a preceding question often shapes the response to subsequent questions, especially if there are successive questions on the same topic (Rasinski et al., 2015). To capitalize upon the question order effect (Rasinski et al., 2015), questions about motivations and goals preceded questions about the LAD, to make learners more apt to readily identify their motivations when they clicked on the LAD for the first time. To prompt longer, more descriptive responses, participants were asked to take their time to share everything that they remembered. Their anticipated responses are identified in Table 29. Loosely categorized, these responses included: getting an overview of personal or group performance, seeing improvement, seeing the effects of the most recent post on the group or LAD, making comparisons, getting feedback on current class standing, and general curiosity.

Table 28. Exp. 5 closed question example

The first time you looked at the LAD, why did you look? (overview, comparison, see how others are doing, to see something in particular, to see progression in time, change as a result of my last post)
To get an overview of self or group.
To see if I have improved since my last post.
To see how the visualization has changed since I last posted.
To see how my last post changed the visualization.
To see the rating of my last post.
To compare my last post to my group.
To compare all of my posts to my group.

The first time you looked at the LAD, why did you look? (overview, comparison, see how others are doing, to see something in particular, to see progression in time, change as a result of my last post)
To compare myself to a particular classmate.
To see if I completed all of the required posts.
To see how many required posts I have left to complete.
To see if I am behind.
To see if my position in the group has improved or decreased.
To see if I have kept my position in my group.
To make sure I am keeping up with my peers.
General curiosity.

Unfortunately, there is no way to provide a list of this length (see Table 28) in an interview of this kind. Closed questions such as this, with answers befitting a checklist, run the risk of interviewer effects because the participant is likely to choose a response before all of the options are presented. Examples like this also communicate what kind of response is desired.

The script was designed to guide the flow of the interview. There were several instances in the script when two questions were intentionally asked simultaneously or in rapid succession. Asking two questions at the same time – for instance, asking why the visualization was viewed and about the information sought by the participant – cued the participant to provide a narrative response. There are several instances in the script where questions are asked immediately one after another (line 26, 33, 36, 44 in Appendix A). This breaks the question-response rhythm of the script on purpose, to encourage a pause for thought and longer responses. Alternate questions were included to fit a range of participant contexts.

The question order changed when addressing the second LAD interaction to break the previously established rhythm, because this is a particularly important interaction. Curiosity assuaged, participants now know what kind of information is provided and

perhaps, their motivations are better defined. Subsequent sections employ questions similar to the previous ones that are posed in a different manner. The final section of the questionnaire asked for summarized statements about the participants' perspective on the LAD, its effects on their perception of their performance, any difficulties experienced, and their perceived utility.

9.5. Procedure

Performed in groups of 4 to 5 learners, the learners participated in a 7-day online learning discussion activity. The LADs were available for learners to use if they wished, accessed through a link in the assignment discussion thread that would open a new page displaying the visualization. Shortly after the conclusion of the discussion activity, participants completed their consent and an online interview. The numeracy questionnaire was completed as part of the consent; an interview was arranged within a week of the discussion conclusion. The semi-structured interviews were facilitated through the Zoom meeting platform due to COVID restrictions. The video recordings were deleted after the interviews were transcribed. Open, axial, and selective coding was performed on each transcribed interview.

9.6. Results

Interviews were conducted with 9 learners (Table 30) from the same class about their experience with the LAD designed for this study. With the university switching to fully online classes partway through the previous term due to COVID restrictions, this previously face-to-face course was delivered 100% online. All of the participants were pursuing a bachelor's degree program at a four-year university and had used some type of visualization in their personal lives (Table 31). The learning analytics-enabled class was a required course for all participants' programs of study.

It came to pass that all of the individuals who chose to participate in this study were international students, and for all of them, English was an additional language. English as a Second Language or (ESL) was not a useful descriptor of participants' spoken English proficiency. While some participants were quite adept in expressing

themselves, others had great difficulty understanding what was being asked of them in the interview. For example, Participant 3 (P3) didn't understand the words online, visualization, or goal. This was frustrating for both interviewer and interviewee and made the interpretation of some of the participants' statements challenging.

The average length of each transcribed interview was 14 pages. Overarching themes present in this learning context were extracted; they are discussed in depth in subsequent sections. Performing the interviews online — with the availability of video — greatly improved the transcription process. One of the participants had a prominent stutter; the majority spoke with accents that combined with audio drops, made it difficult to discern some of their words with audio alone. The video provided context cues and the ability to zoom in on their mouths while they spoke.

Table 29. Exp. 5 interview participants' demographic information

Demographic Information	
Female	4
Male	5
Non-binary, transgender, agender (specify)	0
Age range	19-28
Average age	21
Highest level of education	
High school degree or equivalent (e.g., GED)	6
Some college but no degree	3
Current enrollment	
Yes, full time at a four-year undergraduate college/university	7
Yes, part time at a four-year undergraduate college/university	2

Table 30. Exp. 5 prior visualization experience

Type of visualization	Learners
Banking graphs (ex. account balances)	5
Educational graphs (ex. grades for an online course)	6
Utility graphs (ex. electricity or gas usage)	5
Telephone or internet usage graphs	5
Loan payment graphs (ex. mortgage, student loans)	0
Time planning or tracking software graphs	4
Laboratory result graphs	2
Health or exercise tracking graphs	5

9.6.1. Augmented interview protocol

The semi-structured nature of the interview combined with its open structure allowed for unscripted questions that made the interviews more conversational. Off-script questions were used to make participants feel more at ease. Participant Two (P2), visibly relaxed while talking about his academic interests, speaking slower and giving more eye contact after the exchange. These seemingly unrelated questions also led to more accurate contextual knowledge, and changes in the overall interview protocol.

In response to questions about the experience of taking this course online, P1 stated that the workload was heavier for the current week. Previously the discussion activity entailed answering a single question, coming to agreement as a group using a Google doc as a means of communication, and one person submitting the conclusion. Now the students had to come to their own individual conclusions, read and respond to their peers' posts, and do multiple summaries for multiple questions — in both the discussion thread and the in-lab presentations.

Had P1 not mentioned it, it would not have been known that the discussion activity format changed significantly from the previous week. These changes induced

anxiety in some students, because they had to contend with a number of new procedures to complete the assignment, and because they did not know how their grades would be affected by absentee or low-participation group members. Additionally, this was the first small group discussion for which students were to find their own groups. Previously they had been assigned to groups and placed in the appropriate discussion thread. Many students did not realize that there was an extra step to add themselves to a group in the LMS, or that being in a group was a requirement for the assignment and to participate in this study. Three quarters of the students who signed up to participate in interviews for this study were disqualified because of issues created by this change.

As a result of asking off-script questions about how the group functioned, it was learned that P1's group had ongoing discussions about the learning activity outside of the LMS. Her group heavily strategized their discussion posts and their responses to each other. Following this interview, every participant was asked if their group also followed a similar practice, performing the brunt of the discussion outside of the LMS.

The use of Zoom also allowed participants to share their screens. Early in the interview when there was a question as to what one of the participants saw; the ability to share their screen and essentially drive that portion of the interview gave the participant greater agency. They used the visualization re-creations in a similar way, as a tool to facilitate storytelling. The interview protocol was augmented to include screen sharing for the remainder of the interviews.

A post-activity LAD was also added to the interview script. This LAD was not accessed during the discussion activity; participants saw it for the first time in the interview. Thus participants saw re-creations of the LADs that they viewed during the learning activity, and a new version of the LAD that displayed all of the small group's data after the discussion conclusion. This post-activity LAD depicted what would have been seen had the participant clicked on the LAD after the conclusion of the discussion. The post-activity LAD was particularly helpful for participants who saw an unpopulated or sparsely populated visualization during the learning activity, to be able to interrogate their understanding with the fully populated post-activity LAD. Participants were asked

to share their screen while being questioned about the post-activity LAD, so they could point to salient areas of the visualization if desired.

9.6.2. Interview language

As previously stated, participants demonstrated a range of spoken English proficiency. This required more definition than anticipated and lengthened most of the interviews over their anticipated times. This also meant that there was more frustration to contend with. The phrase “I don’t know” was uttered numerous times, with many different meanings. It was used to communicate humility or hesitation, frustration, forgotten details, lacking confidence, or unwillingness to further extrapolate one’s responses. One of the most difficult and fruitful aspects of the interview was learning to sit in the silence following these utterances. More often than not, participants filled the silence with explanations — explanations that always contributed to understanding the participant experience.

9.6.3. Numeracy

Participants were asked to complete the Subjective Numeracy Scale (SNS) before their interviews, then asked questions pertaining to their numeracy in the interviews, to see if the qualitative responses matched the quantitative scale results. The SNS results are reported in Table 32. There was a good degree of variance amongst participants. the results indicated that they exhibited a preference for numerically presented information higher than their perceived proficiency with numbers. These results were lower than seen in experiment 2.

Aside from the results of P6, participants’ stated numeracy matched their SNS results. What was interesting is the basis each person used to form their opinions. Excerpts are included in Table 32. For several participants their preference for numerically displayed information stemmed from having to communicate in English. P3 said the following:

> "Especially as a ESL, I would say words-- If you are a native speaker you would know how to evaluate the difference between two words, but that is kind of hard for us to evaluate words, then numeric numbers I would say."

For these participants then, their stated preference for numbers could be attributed to difficulty expressing themselves in another language, rather than a genuine preference.

Table 31. Exp. 5 participant SNS results

	Experiment 5 Participants (N = 9)		Experiment 2 Participants (N = 32)	
	Mean	Standard Deviation	Mean	Standard Deviation
SNS ability subscale (1-6)	3.19	0.98	3.89	1.1
SNS preference subscale (1-6)	3.75	0.76	4.67	0.71
Average SNS (1-6)	3.47	0.66	4.28	0.70

Table 32. Exp. 5 participant qualitative numeracy responses

	Avg. SNS (1-6)	Ability with numbers	Preference for numbers
P1	3.75	Average "I guess it's like middle, It's not I like, super."	Yes "If it's not a pie chart or something, like just a number, I sometimes just can't get it."
P2	3.75	Above average "I always considered myself to be good with numbers."	Depends "I find that the numbers are usually useful, but depends on the context."
P3	4.1	Average "I would say its decent [laughs]."	Yes "Because I feel like words for me, especially as a ESL, I would say words… If you are a native speaker you would know how to evaluate the difference between two words, but that is kind of hard for us to evaluate words, then numeric numbers I would say."

	Avg. SNS (1-6)	Ability with numbers	Preference for numbers
P4	4.2	Average	Numbers “It depends on information itself. Some information is easier to understand with numbers…”
P5	5.6	Above average “I am pretty good working with numbers.”	Numbers ” I prefer numbers, it is more obvious to know the percentage, or like… like its more specific”
P6	5	Average “If I could use the calculator, that would help me a lot. Math isn’t really my strong suit…”	Numbers
P7	3.2	Above average	Numbers “Because I'm comfortable and I would rather see that, empirical data…”
P8	4.8	Average	Numbers “Words big problems, and because I don’t like to read.”
P9	3.3	Below average “Probably not that good because I mean, I didn't I didn't do really well in math class.”	Numbers “They are more precise than text.”

9.6.4. The learning activity

This study centered the small group discussion as the learning activity supported by the LAD. The learning activity involved 1) reading a paper introducing several theories, 2) viewing several interactive art pieces, and 3) participating in an online small group discussion. In the online small group discussion students were asked to categorize interactive artworks according to the theory introduced in the paper. At the end of the discussion one team member, called the wrapper, summarized the groups’ position for each of the three artworks and posted these summaries in the thread. In the presenter role, a different team member presented the groups’ conclusion during lab on the last day of

the discussion. Each individual was required to provide a post for each of the three questions during days one through three of the discussion, and a response to at least one peer for each of the questions during days four through six. This would result in a minimum of six posts for each group member, and nine for the discussion wrapper.

Worth a total of 30 points, 50% of the discussion grade was based on the content of each individual's posts. Thirty percent of the grade was based on collaboration, i.e. how well each person engaged in the dialogue with their peers. Ten percent of the grade was allocated to tone and mechanics, message components such as spelling and grammar. Ten percent of the grade was based on the quality of the arguments presented by the conclusion in the thread and the in-lab presentation. The posts were also graded based on quality, in that the criterion above that account for 100% of the grade would be scaled down based on the total number of individual messages posted. If six or more messages were posted per individual, the group would get full marks based on the above 5% for that portion of the grade. For five posts it would be scaled down 4%, and so on. Everyone in the same group received the same mark for 10% of their grade, based on the aforementioned criterion.

9.6.5. Group interactions

Participants were asked how they felt about their group's participation to identify the effects of their group on their participation, and to see how the group influenced the opinions they formed about their learning experience. P5 and P8 were the only interviewees who randomly joined a group; everyone else joined groups with friends previously known to them. A few participants experienced issues due to the differences in time between Vancouver and their Asian countries of residence, but still found their group experience to be positive. P8's group was unaffected by the time difference; it was entirely composed of SFU students residing in Taiwan and China because of COVID. P3 was excited to mention the influence her group had on her understanding of the course materials. One of the group members was a master's degree student auditing the course. This person pushed the group to reflect, to go more in depth than they otherwise would have. According to P3,

> "I do realize that having, um, having a more proactive person [laughs], a professional in my group, made me realize the differences between [long laugh] when you really understand what you are doing than when you are just trying to get a bachelor's degree I guess. It's just that, a lot of the things that she – it feels like she knows the bigger structure of what we are studying right now, let's say for us, we are trying to identify, to put the work into each category, but for her, she can relate those different discipline all together and then try to get us to think about the different possibilities of the artwork I would say."

P5 was the only person who mentioned having difficulty collaborating with his group. He did not understand how to join a group thread and did so late; this contributed to his missing part of the discussion. Though he said there wasn't clear instruction given on how to communicate or exchange personal information, the student also wasn't proactive enough to actually look for this information in the LMS.

Only two participants, P4 and P5, were in groups that did not use additional means of communication such as WeChat or Facebook messenger to conduct group discussions outside of the LMS. The groups of P2 and P3 held group discussions outside of the LMS but didn't strategize this particular assignment in these discussions. As designed the current learning activity was meant to capture the back-and-forth nature of learners' dialogue, and the diversity of thought a group activity should foster. In the groups of P1, P6, P7, P8, and P9 the students strategized the discussion activity outside of class. P1 and P7's groups went so far as to plan who would post what, deleting any duplicate ideas in their responses, and who would respond to each person, if at all. This may explain why many of the posts had low ratings or lacked keywords altogether. In attempting to ensure that each person added something new to the discussion – one of the maxims of the co-operative principle of communication known as the maxim of quantity (Grice, 1975) – participants potentially lowered the coherence of their own messages and those of their peers. Similarly, the lower amount of feedback given in heavily strategized threads could be construed as lack of engagement by an outside observer. It could also deprive students of the learning experience as designed. We saw this take place in P9's group; he was frustrated by the decreased interaction in the LMS, which was a direct result of the groups' Facebook discussions.

Three participants, P4, P3, and P5, had never taken an online course. Though P3 had no previous online learning experience her high school classes were blended, so she was used to having a mixture of in-person lecture and access to online resources via LMS. P3 preferred doing discussions face-to-face because of the amount of time required to commit her thoughts into text online. P2 said he would not have chosen to do the current course online even though he had taken online courses offered on Udemy, and was currently working with an online tutor. P6 had online learning experience but preferred in person courses because online, she had to be extremely organized to stay on top of her work. She went on to complain about the experience, not recognizing that the same could be said of all courses.

> "In the beginning they would send out like these emails and messages, but then near the end it was just like, you had to know what was due – and then they just expected it to be done at that time."

Though she thought face to face courses were easier in this respect, she tended to participate more in online discussions.

> "I feel like because I am more shy, so in person I would be less likely to participate or share my thoughts, whereas online, it's easier to just like, type and then like nobody really knows who you are but then you can share thoughts."

P7, P8, and P9 were the only participants who were enthusiastic about online learning; their enthusiasm stemmed from previous experience. P9 took online English courses in elementary school. He prefers learning online because it carried fewer of the social pressures that may stop people from conversing face-to-face. P7 thoroughly enjoyed his previous experience learning online. He enjoyed asynchronous discussions in particular, and the variety of online tools to support interaction such as polls, breakout rooms, and interactive real-time sketching. P8 previously experienced both flipped classrooms and online-only instruction. He loved the experience of his previous online course because all of the work was performed individually, and he had a week to complete all assignments. Discussions in the previous online courses were graded only for participation, unlike the current discussion activity. He found the group work and online communication method challenging in the present course, because it was more difficult to understand written text than to engage in a face-to-face conversation.

9.6.6. Goals

Early in the interview, participants were asked to describe their goals going into the learning activity. This was followed by a request to select a single goal from a list of four adopted from the 2x2 Achievement Goal Orientation Framework (Elliot & McGregor, 2001). Participants were asked to choose which goal was most accurate for them – to get high marks, to do the minimum amount of work, to learn as much as possible, or to avoid a low mark. This framework (Elliot et al., 2011) is used to investigate students' achievement goal orientations, i.e. why students set about academic tasks. It offers 4 constructs, mastery-approach, mastery-avoidance, performance-approach, and performance-avoidance goals. The avoidance orientation describes the avoidance of failure, while the approach describes the pursuit of success. Mastery is internally oriented and related to developing competence. The performance orientation is normative and speaks to the demonstration of a particular competence. Near the interview's end participants were asked if their goals changed at any point during the learning activity.

Asking an open-ended question first allowed us to identify participants' goals without prompting. Following this with a short-scripted question from the AGO provided a standardized method of goal-based categorization. Interrogating goal orientation at the end of the interview was an opportunity for learners to examine the persistence of their goals after what was, in essence, a long period of prompted reflection on their learning experience. Table 33 combines participants' goal-related responses.

Table 33. Participant goal orientation table

	Stated goal at the beginning of interview	Chosen goal at the beginning of interview	Goal at the end of interview
P1	To finish the assignment	To get a high mark	Changed from high marks to "[it's] not about the mark, is just to get it right."
P2	To do his best and complete what was asked of him.	Avoid a low mark	Remained not getting a low mark

	Stated goal at the beginning of interview	Chosen goal at the beginning of interview	Goal at the end of interview
P3	To see how other people determine different types of interactions.	Learn as much as possible	Changed to avoiding a low mark
P4	"Learn how to produce interactive for my future job career…"	To get a high mark	Remained to get a high mark
P5	"To refine my opinion"	To get a high mark	Changed to completion
P6	"See how other people interpreted interactive art pieces"	Avoid a low mark	Changed to "learn and understand it well"
P7	"Understand the interactive strategies and employ them."	Learn as much as possible	Remained learning as much as possible
P8	Finish it, get the mark, and learn from peers instead of reading.	Learn as much as possible and get a high mark	Remained learning as much as possible
P9	"I don't know."	To get a high mark and to learn as much as possible	Remained high mark and learning as much as possible

P7 was a unique case in that his goal was to learn as much as possible – this, even after stating that this class would have no bearing on his future career – and he never deviated from this goal. The goal was further supported by how he approached the assignment, his strategic group selection, and even the language he used to discuss them, frequently calling them teammates, rather than groupmates. He methodically approached the assignment and believed the group discussion gave him "a different lens of looking at things."

In regard to his stated and chosen goals, P8 was the complete opposite. Though his goals did not deviate, his stated goals were at odds with his chosen goal. P8 wanted to finish the assignment to get the mark, to gain more knowledge from the reading, and to learn from peers since he didn't understand the reading. Recalling that this was the participant who did not like to read in any language, it made sense that he would prioritize learning from the explanations of others over reading. He chose two goals,

learning as much as possible and getting a high mark. Some of P8's actions in the learning activity better matched someone whose goal was simply to finish the assignment. While he was diligent in completing and improving his message posts, he did not read the assignment rubric to ensure that his post met requirements. He was aware of the rubric, but instead chose to ascertain the assignment requirements through analyzing the posts of his peers. At one point in the interview he mentioned not wanting to take this particular course online, because it was harder to find the answers to this type of course using Google.

Two participants had avoidance base goals. For P2, these goals resulted from previous learning experience.

> "I guess I just want to make sure I do my best. Before I was taking [a writing] course and we were supposed to do something similar, like, reading summary stuff. And I always would get this feedback that 'you did a good job, but like not quite what we wanted. You didn't answer the question we required you to answer, we were quite there.' Like, I guess my goal from now on is to actually do, like step-by-step answer all questions, and be sure that actually cover what is being asked of me."

Four of the nine participants' goals changed during the learning activity, most often because time got away from them. Initially eager to get started and learn as much as possible, as time went on, they just wanted to finish the assignment and do enough to not earn a low mark. P3's description of the experience was echoed by several other participants.

> "So it's, it's more like a curve. In the first day you are really passionate about doing this [laughs], and then it kind of dies down until you realize that oh, that the deadlines is coming, and that's when you go back to the readings and trying to complete everything that you have left."

For one participant, P6, the change went in the opposite direction, after seeing that the theories introduced in this activity would be used in subsequent discussions. Her goals changed after she realized that the readings were more important than she previously thought. P1 attributed her changed goals to her use of the LAD. Her primary goal changed from getting a high mark to being right. Being right, as evidence by the visualization, was more satisfying than the mark.

9.6.7. Participant experience

In the following sections, the participant experience is examined according to commonalities of experience with the LAD, such as the number of times the visualization was accessed. Though the participants are grouped, it is still important to understand their individual goals, and the contexts in which they contributed to the learning activity. The provided interview excerpts include both the question and response when all possible, to better seat the participants' responses within the conversation, and to increase the reliability of the interview's interpretation.

Contextual information is provided as necessary to paint a more complete picture of why participants acted in the ways that they did, how they interpreted the LAD, and how this interpretation influenced subsequent action. P8 and P1's experiences represent the extremes. While P1 was successful, using the LAD in expected and unexpected ways, P8 was unsuccessful with the LAD and completely unaware of what he could have done to improve his understanding or overall learning experience. All other participants' experience fell within a spectrum between these two individuals. Though the majority accessed the LAD between 1 and 5 times, P1 accessed it 23 times.

Accessing the LAD the first time

Similar to Experiment 2, participants most often cited curiosity as the reason that they initially accessed the LAD. Only one person, P5, accessed the LAD intentionally to check the quality of their work. As was evident from Experiment 2, the first time the LAD is accessed represents a crucial decision point for learners, because this is when they most often decide if they will continue to use the visualization. This is often when expectations are set or met, when valuation and conceptualization of self-performance meets expectation. In Experiment 2 many participants stopped using the LAD if it was not populated initially, because they thought the LAD did not work. For this study we added an overlay across the LAD explaining why it was not yet populated (see Figure 26). Unfortunately, this was not effective, as few participants read it or any of the other text that accompanied the visualization.

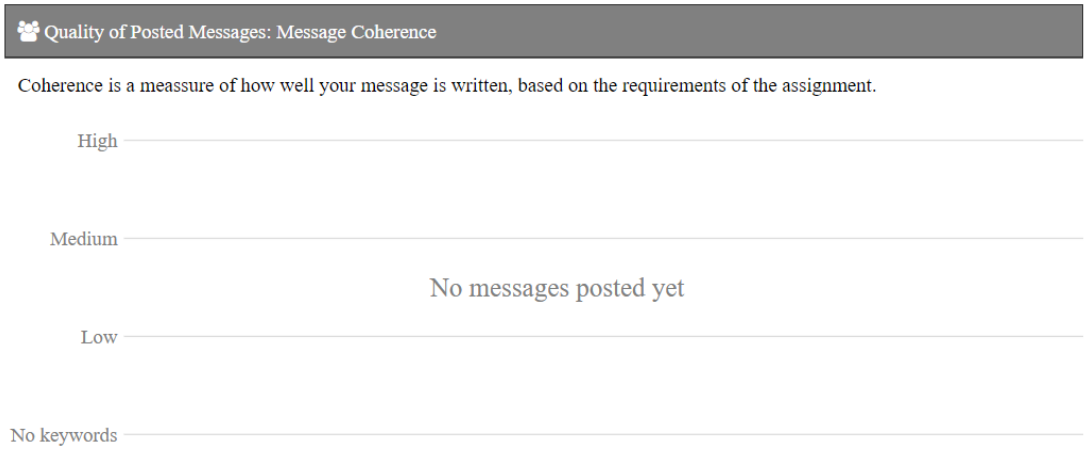

Figure 26. Exp. 5 unpopulated LAD

The visualization was not populated when first accessed by P1, P2, P5, P8, and P9. P2 and P8 first accessed the LAD just to ensure it was available. Two participants, P2 and P1, did not attempt to understand it. P5 and P8 assumed the LAD did not work. P8 was the only participant to mention having formed his opinion of the LAD before he accessed it, expecting the LAD to be populated because he thought it made comparisons between all of the small discussion groups. Only one participant in this group, P9, understood that the visualization required posted messages to populate. This was likely because he was the only participant to read the explanatory text – the title of the LAD, its description, and the definition of coherence. This also included the button text that read "[c]lick here to open the visualization in a new window. It updates every 5 minutes." P9 understood that the LAD compared his data to that of his group members, and that the results were keyword-based.

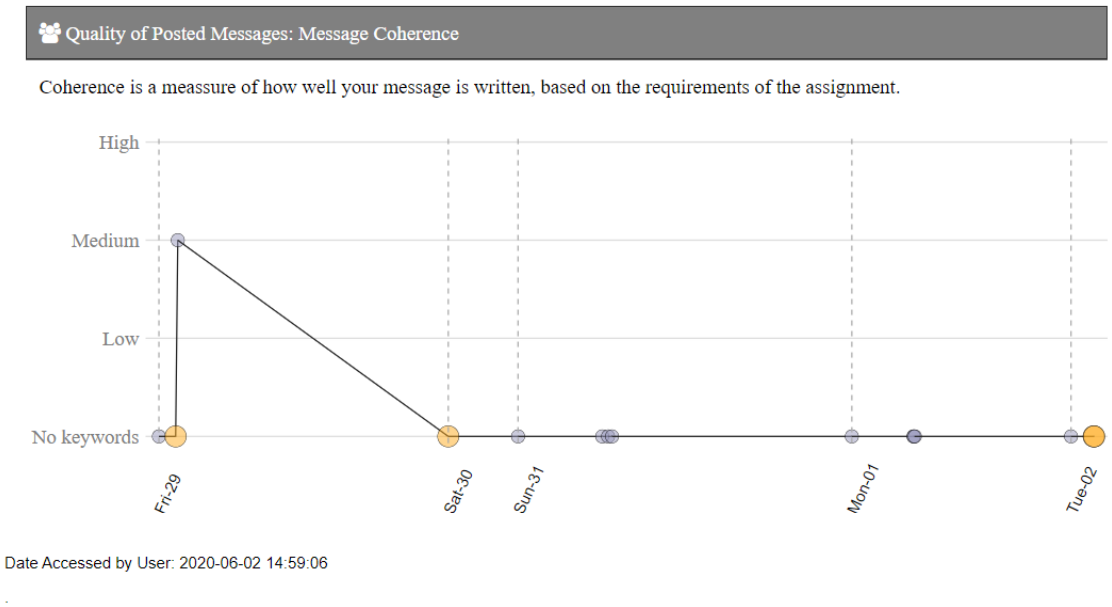

Figure 27. Exp. 5 first LAD viewed by P3

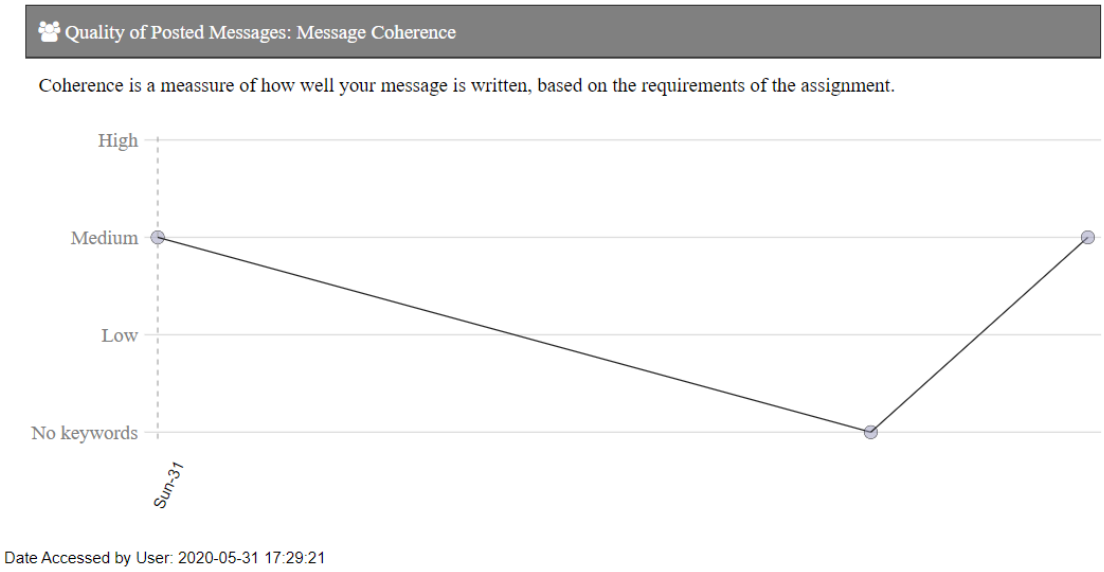

Figure 28. Exp. 5 first LAD viewed by P6

The LAD was populated when first accessed by P3, P4, P6, and P7. Of these participants, P3 and P6 thought the LAD (in Figure 27 and Figure 28) was a piece of interactive art to review for the assignment. P6 recognized that it would change and thought that it was “cool to look at.” Like P6, P7 recognized that the LAD would change or move, but he made no attempt to better understand what it visualized. He did however

understand that it was a visualization, and that the visualization was somehow related to his work. P4 knew that the yellow points in the visualization represented her data (Figure 29). What she did not understand was why one of those points had a low rating, since she though all her posted deserved a medium rating. She reviewed her post again, but then decided that the discrepancy wasn't important enough to ponder further.

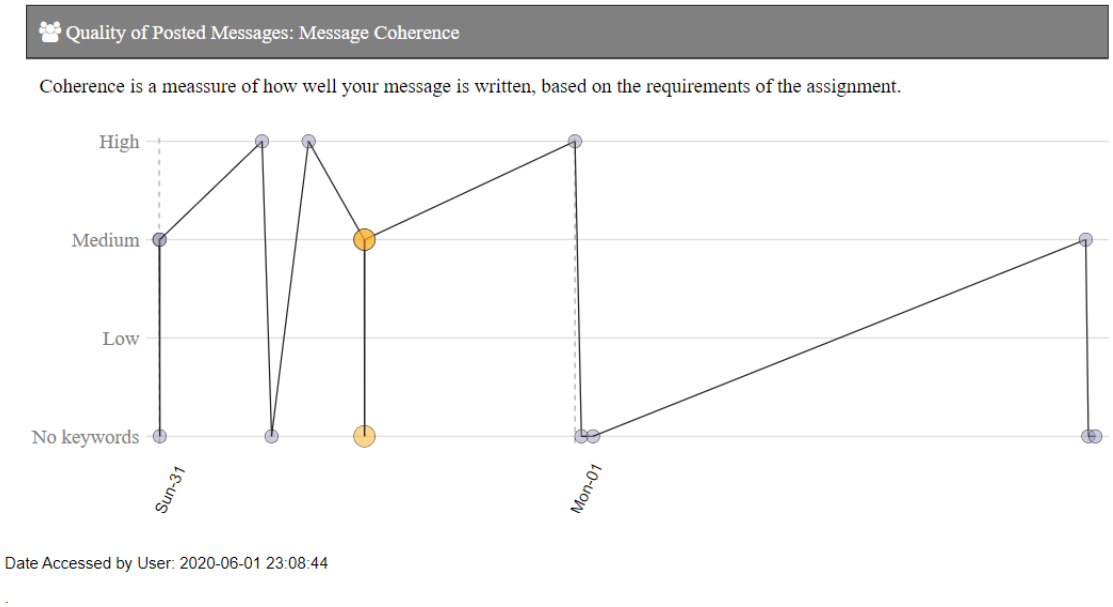

Figure 29. Exp. 5 first LAD viewed by P4

P3 and P8's experience – using the LAD once

We review the experience of two participants, P3 and P8, together because of their similar traits, experience in the learning activity, and interactions with the LAD. Both participants interacted with the LAD only once. They both struggled in the learning activity due to low written English proficiency, mentioning the significant amount of time required to translate their thoughts into English as a significant challenge. This difficulty influenced P3's stated numeracy and likely her goals, as during the learning activity, they changed from learning is much as possible to not earning a low mark. They also both stated a preference for information displayed in numbers rather than words. P3 was direct in saying that her stated preference was due to difficulty evaluating words in English. Likewise, P8's preference stemmed from his aversion to reading or writing in any language.

Having experience blended courses in high school, P3 was used to using an LMS to access course resources. P8 was comfortable learning online, having enjoyed the several online courses he took previously. P3 had a vested interest in both the class and the subject matter, and by all accounts seemed like an eager learner. Aversion aside, P8 still took the time to revise five of his 6 posts to improve their grammar. His stated goals were to get a high mark, to gain more knowledge from the readings, and to gain knowledge of the concepts from the readings by way of the discussion. He chose two goals from the list of four – to learn as much as possible and to get a high mark – because for him the two were synonymous.

The combination of their goal orientations, learning experiences, and approach to the learning activity make them sound like "good" students, primed and ready to understand and benefit from the LAD. Neither of them experienced the LAD as intended, though for different reasons. Neither P3 nor P8 had a strategic goal in mind when they first accessed the LAD. Since neither of them read any of the explanatory text, they did not understand why the LAD was not populated.

At first it was not readily apparent why P3 had difficulty with the LAD. Changes in her speech patterns hinted at her reticence to admit that she experienced any difficulty. Laughter also often accompanied what seemed to be embarrassment or shame. Whenever she mentioned behaviors deemed less acceptable, she switched from "I statements" to "you statements." The laughter and pronoun switching provided cues that aided in the interpretation of her statements. Each time she was asked about her motivation or understanding in reference to the LAD, her answer changed. After three completely different descriptions of her understanding of the LAD, she admitted that she actually had no idea what the LAD depicted, because until that moment, she thought it was one of the pieces of interactive art being critiqued in the learning activity.

In the post-activity portion of the interview P3 was immediately contrite, suddenly focusing her gaze off-camera, when she had previously been giving direct eye contact.

> P3: [Laughs for a long time – looking down and away from camera] Yep. Wow, looking at it now I kind of feel bad for myself and my group [laughs more].

Int: Why would you say that?

P3: Now that you point out the algorithm is actually evaluating the quality of our discussion, I feel like maybe a lot of the discussion could be re—like, elaborated a little bit more.

She went on to say that she didn't realize how connected the visualization was to her work. She excused her misunderstanding in saying that she had only interacted with the LAD once, and by reiterating the idea that online discussions are more difficult than oral discussions because of the need to write out one's responses. It was interesting to note that while P3 did not believe that "algorithms" could signify learning, or even when learning occurred, she believed that the LAD was accurate because everyone in her group was likely trying to just "get the assignment over with." The following reply illustrates what aspects of the LAD she would use to assess the quality of her teammates, and presumably her own, work.

> "I, I don't want to deny -- you can't really determine the quality of those discussion based off how academic those writing were… but its also, is also true that a lot of my group mates, looking at the time of the replies, looking at the link, and even looking at the punctuation that they use, the grammar, the spelling, you can tell that a lot of the times people are just trying to get it over with. So I do feel like this graph represents the overall quality of the discussion."

Using the post-activity LAD, she easily surmised from it that her group did not do well in the discussion.

> "Yeah I do feel like even though, after knowing such algorithm, I would sort of force myself to use keywords in my discussion in order to achieve like a higher-quality [laughs] prose rating on the graph that, but-- I do believe that that's a good thing to do so, because at the end of the day you are you are trying to perceive, like, you're trying to be like, a professional in the field. You really need to understand the terminology that you will be using in the future and I do feel like it would help [pause], help us, like, in the general academic writing or how we -- It sort of reminds us that even small discussion like this also matters and we should not, we should try to practice academic writing during everyday discussions."

P3 was aware of the role that LADs could play in helping her development discipline-specific language.

Like P3, P8 also produced multiple explanations of what he thought the LAD visualized. First, he thought the LAD was where the TA gave feedback. This contradicted a previous statement, that he expected the LAD to be populated when first accessed. If

the LAD was a TA-directed mechanism for feedback, it still would have been empty because he had not yet posted a discussion response. Next P8 said he thought the LAD compared his group to other groups in the class, so that is why it should have been populated. He further reasoned that since it is not customary to publicize individuals' grades in Canada – unlike his own country – that this was the purpose of the LAD. With this reasoning P8 contradicted both of his earlier statements.

It was clear from these convoluted assertions that P8 did not understand the purpose of the LAD or what it visualized. The exchange below exemplifies the level of critical thought P8 applied to understanding the LAD.

> Int: So did you read the sentence that I'm highlighting now about coherence when you looked at this visualization?
>
> P8: I think I did, but I didn't look that closely, yeah. [Begins reading the line about coherence that I highlighted aloud to himself.] Coherence is a measure of how well your message is written, based on the requirements of the assignment.
>
> Int: Okay what do you understand about coherence? What does it mean?
>
> P8: More accurate? Something like that for me. Like if you organize well, like, if you're like, writing the specific point, the point that they want, that's what I think at first. Not sure if I'm right or not!
>
> Int: I'll explain it once we –
>
> P8: Because I saw that one a lot. But I'm not, pretty sure like, they keep writing that word but I don't know, I don't know what they mean [laughs].
>
> Int: Wait, you saw the word coherence a lot?
>
> P8: Yeah, a lot. Not in this [indicating this class] -- I'm not pretty sure what is accurate. From what I know I feel like its accurate, correctly, similar? That's what I say, yeah.

P8 didn't bother to look up the word coherence even though, by his own admission, he had seen the word used multiple times in his classes. Additional behaviors were not aligned with his stated or chosen goals. For example, he contributed to the discussion early – an action that aligned with one of his secondary goals, not being late with any of his assignments – but he did not review the provided rubric. Unprompted he mentioned the rubric — only to say that he never looked at it. Instead, he chose to assume that this assignment would be graded similarly to the online assignments from a previous online course, and to deductively learn this particular assignment's requirements from his peers'

posts. His prior online learning experience aided his comfort levels with asynchronous learning but did him a disservice when setting expectations for the current learning activity.

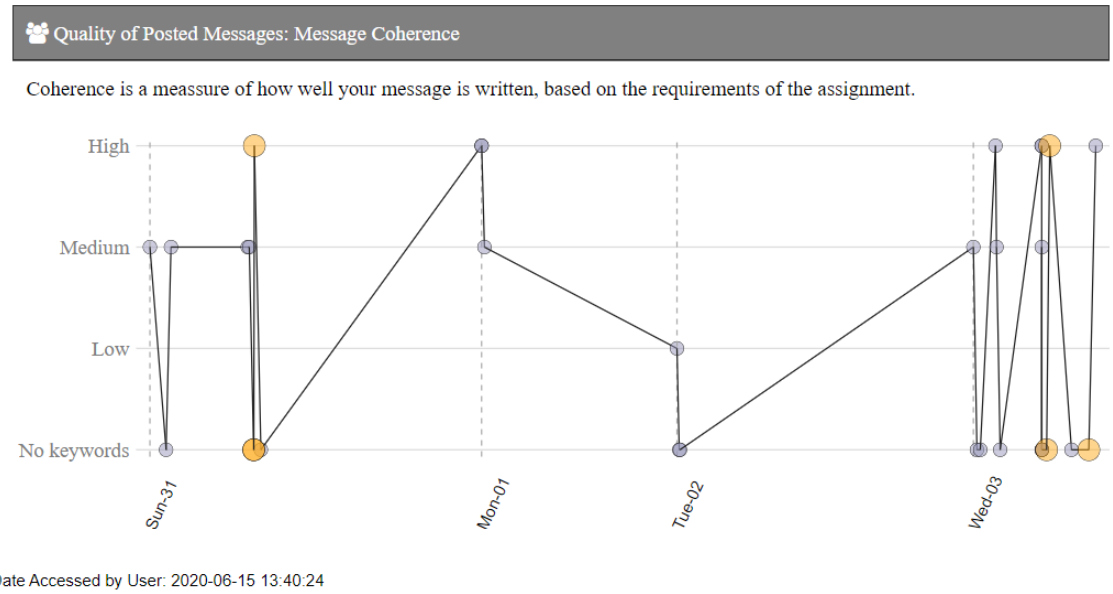

Figure 30. Exp. 5 P8's LAD after discussion conclusion

P8's response to the post-discussion LAD in Figure 30 was enthusiastic. "Oh, there's something there! Oh, I don't know that!" He then read the description above it aloud, but this did little to clarify the visualization, as is evident in the following exchange.

> Int: Can you tell me everything you understand now?
>
> P8: Which one is my own post, for the orange one? I know what they mean, like basically, like, I know what they want to show us but… Is that dot for individual, for group, or for myself? Like orange one is me, or my group? Something like that.
>
> Int: So you're trying to figure out what the dots mean. What do you think they mean?
>
> P8: From what I guess? What I guess [prolonged sigh, as if in annoyance]. I think its about orange one is for my group or for me. Blue one is from other individual, or other group mates, or with our team or other group in the class from what I know. And the orange is for me, I'm the orange one if I – if I just – Because there's no description on the bottom or the sides, I'm not really sure what does that mean from what I guess.

This went on, with him eventually guessing that the orange points represented his data because there were fewer orange data points than gray ones.

P8 went into the assignment expecting his performance to be medium; he was surprised that several of his posts were of high quality. Viewing Figure 30, P8 became aware of the three posts that he made with no keywords. He fixated on them, muttering "no keywords" to himself multiple times. Citing the variance between the high and no keywords data points, he went on to underestimate his performance. P8 described his performance as below average even though quality of his posts would have averaged to medium, alongside those of his peers. He concluded that reviewing the summative LAD made him realize just how little he understood about the assignment.

When asked if he would use a similar LAD in future courses, P8 said that he would, if improvements were made that would aid intelligibility. He then went on to suggest that we add descriptions to the LAD – the descriptions provided with the current visualization that he never bothered to read. He seemed oblivious to the fact that he read these passages aloud just minutes before. Complaining that the provided coherence ratings were too broad, the final improvement P8 suggested was to simplify the LAD to indicate only average performance. Knowing the average would benefit him most because his goal was to be above average in all of his courses.

To conclude the interview P8 was asked if he had anything more he would like to add, or any remaining questions to be answered. He asked what the study was "really for." The question was confounding given that he was present for the description of the study given to the class, subsequently signing up for the study titled "Learning Analytics Visual Cognition Interviews." Since P8 had also completed the consent form, it was perfectly logical to conclude that he knew the study was to provide feedback on the learning analytics he used in his course. Instead, he said he thought the study might be about the theory introduced in the reading, the lab, or the lecture.

Again, it was explained that the interviews were conducted to get feedback on the visualizations. He was told that the LAD was designed to give learners immediate feedback on the quality of their writing during the process of engaging in a discussion activity, to help them adjust their learning strategies in the moment, to contribute to their

ultimate academic success. It was reiterated that the LAD could be used to guide learners as they completed their assignments. Realization finally hit P8 when he was told that learners could "gauge the quality of their messages during the activity." He asked if the visualization updated automatically, or if it was created by the TA who did the grading. He was shocked to learn that the LAD updated automatically, even though the label on the button used to access the LAD said that it "updates every five minutes."

> P8: Automatically? How come Canvas knows what we are learning?
>
> Int: [Explains that the professor identified keywords used in the LAD, then the updated display.]
>
> P8: I see! Wow, I never know this before! Really! That's something new to me, that you can do that in Canvas!
>
> Int: So I'm curious, you clicked on it [the LAD] without reading what it said?
>
> P8: Yeah, I just think because they say, they told me to click on it-- Automatically, and oh, wow --
>
> Int: Now you see why it is useful as you're learning, so it doesn't matter if you are writing and posting in the middle of the night or whenever, you can get feedback pretty quickly.
>
> P8: Yeah, yeah, I like! Woah, that's why we are – oh man, I think that we are doing – woah, I are doing, we have hope, we hope we have this one [puts both hand up in the air as if to high five], it kind of graph.
>
> Int: And that's why we are doing things like this, is because anything that helps students get feedback, particularly in courses where it is easy to get lost and it is very difficult to understand how you're doing in relationship to your peers, or when you are doing a lot of small group work activities and you can't judge where you are in the class. Do you understand?
>
> P8: Yeah, because we always have to like, we always have to wait for grade, like for a long time that we even get our grade and we are not sure what to do next. Without getting the grade for this time, and this one [meaning the visualization] the ultimate update one, can be really helpful like — right now if I know that, I can look at this first and do my [laughs], my second one [meaning his posts].

The discrepancy between P8's enthusiasm for learning analytics and his actual experience, belies the importance of testing learning analytics with intended users, instead of relying solely on self-reports. Though he made a point to access all of the course resources such as the LAD, he did not read any of the descriptive text. When the LAD did not act as expected – when he thought the empty LAD should have been

populated – he did nothing to better understand why. The effort P8 put into grammatically correcting his work and the continued distress he exhibited over the no keywords posts were evidence that he cared about his learning. The difficulty he had understanding what the present interview entailed or how the LAD could have been useful to him may signify a lack of awareness, ability or willingness to think critically. P8 seems unaware of the steps he should take to reach his academic goals. His enthusiasm for all that LA has to offer was heartening but dampened by the actual experience he had when exposed to LAD.

In particular, P8's responses exemplify a lack of awareness or willingness to think critically. His behaviors were misaligned with his stated goals of achieving high marks and learning as much as possible. The LAD had no effect on the learning of either P3 or P8. Neither participant's actions during the learning activity were commensurate with the type of student that they superficially appeared to be.

P6 and P7's experience – Two interactions & peer validation

The LAD was populated when first accessed by P6 and P7, who both accessed the visualization twice. Both viewed the LAD the first time due to curiosity, and the second time to see it change. Neither read the provided text descriptions initially and when they viewed the LAD the second time, both participants saw results that were lower than expected. When this happened, they both looked to their peers for validation. While their understanding of the LAD and approaches to their learning differed, the LAD had the same minimal result for both participants.

P6's approach to the discussion activity was similar to many students interviewed in Experiment 2, in that she waited until other people had posted to contribute her own messages and adjusted her work to align with theirs. She did so even after having previously strategized the discussion with her group. When she initially accessed the LAD, P6 did not attempt to understand what was visualized because she thought it was interactive art. When P6 accessed the LAD a few hours later, she had posted two messages of her own. At this point P6 was fairly sure that the yellow data points represented her posts, but not entirely sure whose posts the gray points represented. P6 was surprised by what was visualized because she thought her posts should have been

rated higher. Though no other posts had been added since she initially accessed the LAD, P6 attributed this improved understanding of the LAD to its being populated. After P6 – whose chosen goal was to avoid a low mark – viewed the LAD the second time she went back to review her post. Not knowing why her peers had higher ratings or how elaborate on the keywords to improve her posts' ratings, she felt stuck. P6 eventually decided that it was not important enough to continue trying.

P7 took pride in his work; he described himself as the type of person who participates the most in group work, is often first to contribute and to coordinate the work of his teammates. Unlike P6, his initial understanding of the LAD was that it assessed his performance and measured engagement of some kind. Reading the definition of coherence during his second LAD interaction enhanced this understanding. He assumed the yellow data points belonged to him, since he believed that this color always signified something of importance in a graph. He was confused by the gray data points because they were not labeled with the names of his group members. When prompted he guessed correctly, but he was adamant that he did not understand the LAD. Looking at the visualization in Figure 31, P7 was surprised that he got medium coherency scores on his posts, because he was expecting a high rating.

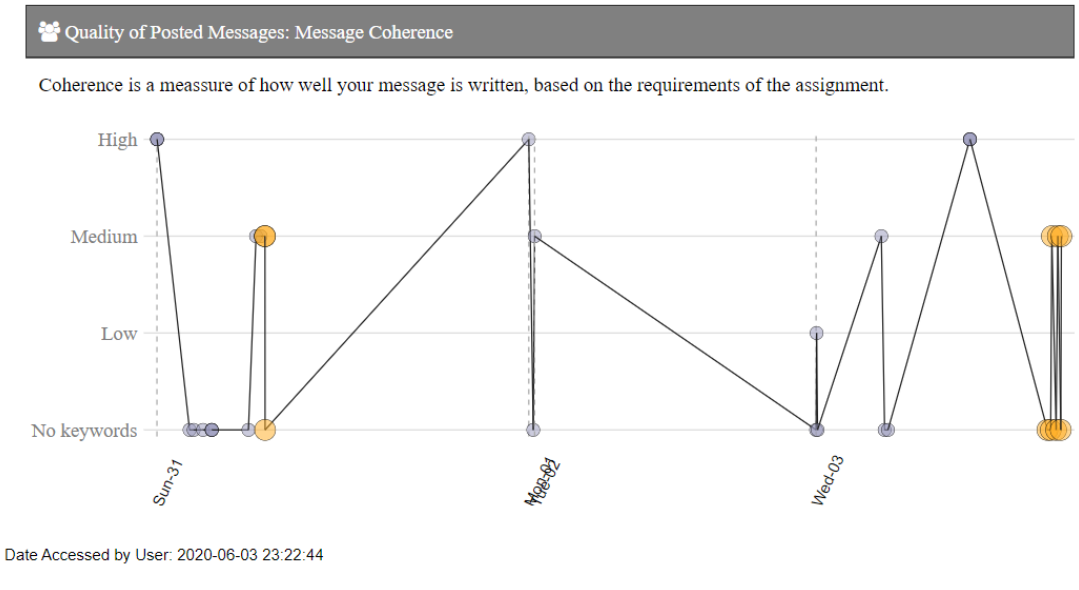

Figure 31. Exp. 5 second time P7 accessed LAD

He immediately read the provided definitions and attempted to make sense of the LAD.

> "It is a quality over quantity, or quantity over quality? Because coherency is how much you stay on the subject right? Well, that's how I understand it. But also it depends on how much you stayed on the subject, and not drifted away as much from the subject. So I just, I just, I don't know."

Explaining his thought process, he elaborated further.

> "I thought it was coherent enough to answer their questions based on what they said and connect it and tie it back to what they said and give them feedback on that. But if I get a medium score on my coherency level, I just, I don't know what the system regards as high. You have to talk a lot? Do you have to not talk a lot? Do you have to be on point with what you say? Do you have to be connected with what they say, and what they don't say — it's like more of a quality versus quantity type of measure. It was a bit confusing to me as to [sighs, then laughs]—
>
> What should I do? Should I talk a lot? Should I not talk a lot? Should I go straight to the point, should I give them feedback on this and not? And most profs are different too – some profs go on word count. If you write a lot, you've written a lot. If you haven't written a lot, but you've written a quality of piece of work then you get – so there's different preferences there, and that's what I thought about with how this graph assessed as well, so I just didn't know."

Like P6 P7 went back to review his work, comparing it against that of his peers. Finding nothing apparently wrong, P7 decided to not entirely trust the visualized ratings. Since the quality of his work was reaffirmed by both the professor and his peers, he trusted this information more than that provided by the LAD.

Both P6 and P7 went back to review the quality of their work, and both looked to their peers for validation of its quality. P6 lacked confidence in her work and saw her peers on the same level of ability. This left her without a gauge. Though she did look to others to anchor the mental model of the quality of her performance, she could not discern what differentiated a medium post from a high-quality one. Had she known what to do to improve her work, she said that she likely would have done so. P7 had an entirely different conceptualization of the quality of his work. With multiple sources of feedback that confirmed his self-conceptualization, he chose to distrust the LAD, since its feedback was not aligned with the other sources.

P7 found the post-activity LAD helpful and even though he did not correctly understand what was visualized, the gist he gleaned from it was accurate.

> "It was definitely helpful to see, a little bit helpful to see, how well the team is doing, and how they should do better, when they should've posted, and when they should engage, and when they should um yeah... So like when the graph is flat, like the dots at the bottom, and gave me questions like "why was the graph flat at that point, and when did it spike up suddenly, and then why did it drop again, and why… So it gave me an idea about why — the engagement process should be active, it should not fall to the ground like that suddenly and it should just try to stay as it is and, yeah, we should share our ideas more often. And uh yeah, stuff like that."

Both participants' mental models of their performance were influenced by the LAD, though these impressions were not persistent for either participant. Once explained, P6 and P7 both thought that the LAD would aid them. P6 said she would use the LAD during the process of writing her discussion posts, even though she did not do this in the current case. Even though he did not trust its assessment of *his* work, P7 thought the LAD was helpful because it allowed him to see how well his team was doing and that "the engagement process should be active."

P2 and P5's experience – Using the LAD 3 times

P5 and P2 may have each interacted with the LAD 3 times, but neither of them thought critically about what was displayed until they were in the interview. P5 joined the class late, just days before the discussion activity began. Like several of the other participants, P5's goal changed over the course of the learning activity – from achieving high marks to just completing the assignment. P5 said he first clicked on the LAD to "see what the tool was for." Since it was not yet populated the first two times P5 accessed the LAD, he assumed that it did not work. Like several of the other participants, he expected it to be populated, though he had not posted.

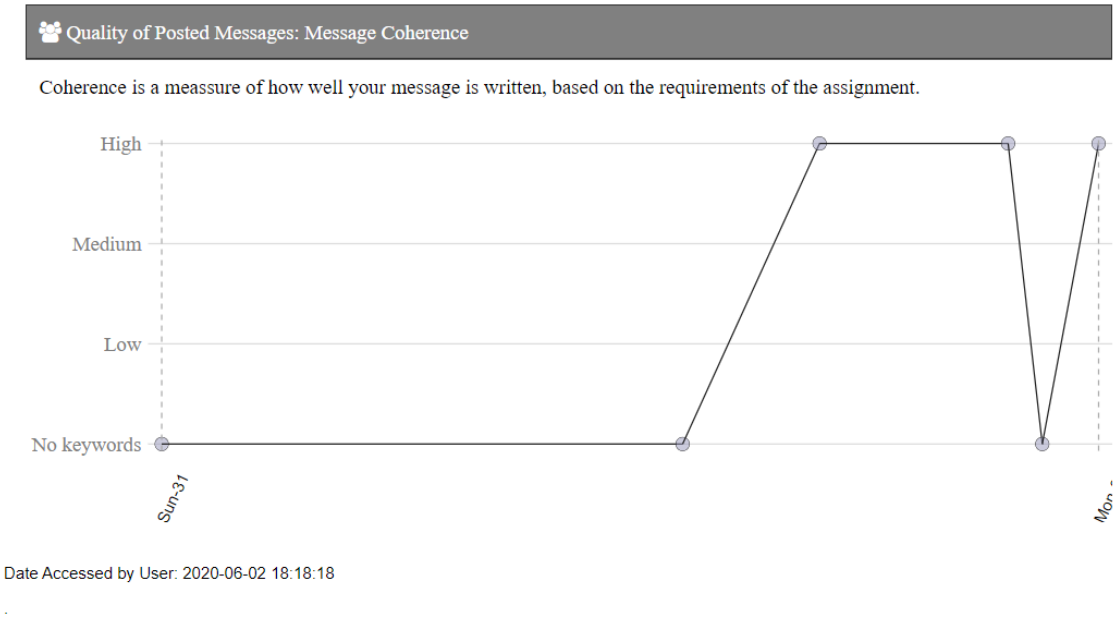

Figure 32. Exp. 5 P5's LAD viewed third time

The third time P5 viewed the LAD it displayed 6 data points, none of them his. Until the interview, he had assumed that the data points represented his work, even though he had not yet contributed a message to the discussion. His understanding of gist at this point was that "it [the LAD] is for how well the group was working." When asked to reiterate, he said, "I mean the graph is only for the context, or like what… the performance of the whole group." Using the LAD in Figure 32, P5 was asked to describe all that he understood from the visualization.

"Um… . I have no idea. Is it good? [Laughs]. The graph says the keywords are in the high level. [Long pause] Well, what I can see from my graph is, it shows from 31st of May to 4 June."

Reminiscent of the student responses from Experiment 3, P5 noted one explanatory detail of the visualization, experienced difficulty, and then switched to descriptive details. To see if he could better understand the LAD if it included his own data, P5 was then asked to review the summative LAD. He easily summarized his work using this LAD, but this was not challenging since all of his data points had the same rating of no keywords.

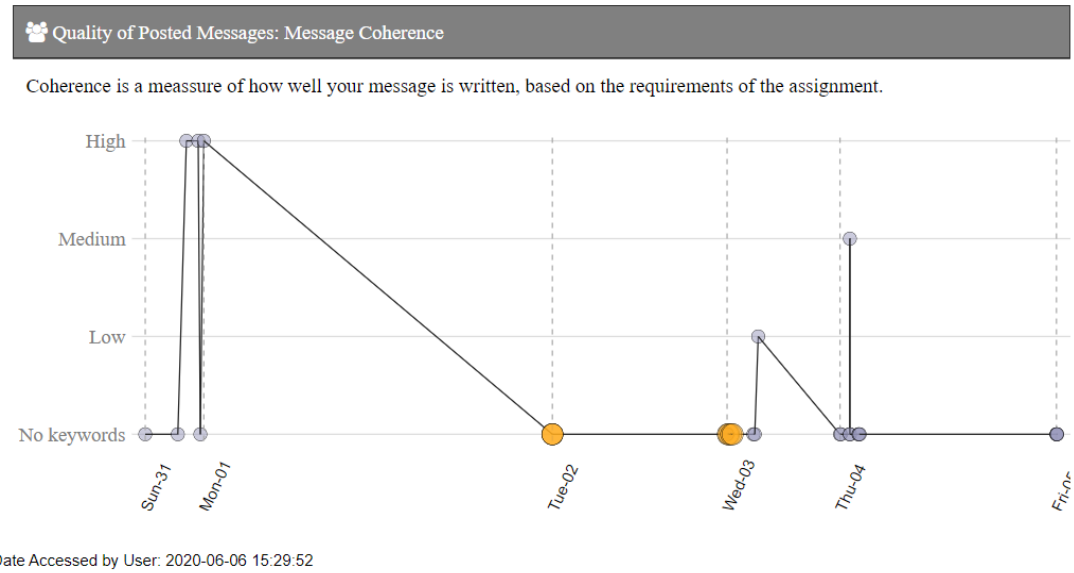

Figure 33. Exp. 5 LAD P5 viewed after discussion conclusion

Using the LAD in Figure 33, P5 initially blamed the lack of keywords in his posts on his earlier mistake of posting in the wrong area, but then changed his mind.

> Int: When you are looking at the visualization, how did you think that you were doing based on what you saw?
>
> P5: It was the relationship between preferments [preference], or what I expected. I didn't see anything, so I was like, "I so disappointed!" I was like, "oh, why are there no keywords in my comments?" [Laughs]
>
> Int: Did you think that it was broken, or did you think that something was wrong with your post?
>
> P5: I thought that probably, something was wrong with my post, that the quality of my post was not that high, yeah I would say it like that.

Once the purpose of the LAD was explained to him P5 said that he would find it useful in future courses.

> "I think the tool is great, like it is part of a reflection about our work. I would say another human-to-human communication or feedback is much more specific or is helpful to point where should we refine or to critique."

Like P6, P5 felt that the LAD should indicate how he could go about improving his work.

From the way he approached interactions with his group to the reason he first accessed the LAD, P2 was motivated by the avoidance of negative consequences. The second time P2 accessed the LAD there were six data points in the visualization. Even

though he had contributed only one message to the discussion, until this was brought to his attention in the interview, he assumed all of these data points were his. Though his gist assessment was not accurate and he did not read the descriptions that would have aided his interpretation, he did act on what was visualized in Figure 34. He could not explain how he interpreted the LAD at this point, only saying that he was surprised that his message contained no keywords. To clarify his perspective, this line of questioning was revisited later in the interview, but P2's response made little sense.

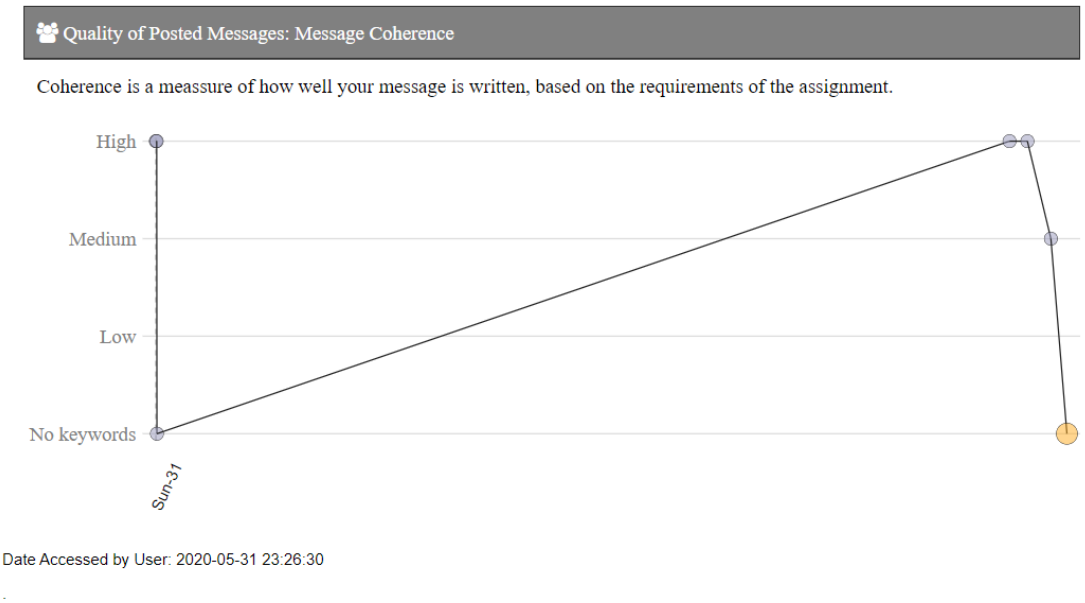

Figure 34. Exp. 5 visualization seen by P2 the second time LAD was accessed

Int: Did you think that all of these dots were yours, or how did you understand what you are seeing here [Figure was on the screen]?

P2: Yes, that is like all my my, like, depend on a question. So like, you know, I thought the first thought is like the first question, the level of my keywords. The second that was my second answer, so I thought it like, only my graph.

Int: So you thought that the graph was showing only your information? So were the dots, did you think the dots themselves were individual keywords?

P2: Well, I was thinking more about like level of statistics, so like, not necessarily keyword but like, just, level of quality I guess.

Int: So what were the dots?

P2: Will be higher, but not quality, but amount of keywords. Yeah, because it says keywords, so amount of keywords used. And I didn't really like, pay attention since there was like, three

questions and how many [counts the number of dots on the screen now] like, 6 dots? Yeah.

Int: So were the dots then, maybe, the questions you were thinking initially?

P2: Yeah, and I didn't – yeah, so like, what I was thinking like, you know, like, was that it takes my time of answering [runs mouse along line up from Sun 31st to later] and like, okay so at this time when you start your answer here's the amount of keywords you have, and as you progress, here's like, more or less how many keywords you use compared to like overall.

If the LMS functioned as he described, then it would have taken at least 30 minutes for him to write his message. Since there is no way to save a message post in the LMS, this would not be possible. It is more plausible that he assumed that all the data points represented his own work, because he failed to think critically about what was depicted.

P2 accessed the LAD multiple times but did not read the two provided sentences of descriptions that would have aided his interpretation. He reviewed his work and compared it with his peers but did not attempt to revisit his post because he did not trust the rating the LAD displayed. Later, when his work shown to have a higher rating he said he did trust the LAD. P2 only trusted the LAD when it displayed positive information that he agreed with. Taken together these actions may seem to exemplify a lack of awareness, but when asked if there were times when he found the LAD helpful for this learning activity he had this to say:

> "Well, thing is that I didn't really like, thought about it, so like, I guess, uh like, like, since it showed overall performance it was interesting information to know. That at the time like, since I didn't really like, thought about it, it wasn't useful for me. If I would actually think about it, and do like, thinking about how I can improve for next time, it would be way more useful."

When the long-term goals of learning analytics visualizations were explained to him, he added to this earlier assertion.

> "Like if I would spend more time on my own to actually understand it, I probably will understand it, but since I actually didn't like pay enough attention to it, like, I guess what I'm trying to say is like, if student would want to understand it he or she would be able to."

By his own admonition P2 said the visualizations would help perfectionists, not students like him, unless they were willing to engage. He admitted that he would have given more

effort to reviewing the LAD properly if it would have had a more significant impact on his grade. It is possible that had P2 felt more confident in his understanding of the visualization or valued the learning activity more, he would have been willing to take more concrete action. Had P2 looked anywhere for additional information the second time he accessed the LAD — especially the description of what the LAD displayed that was written about the visualization – perhaps he could have successfully used the visualization to improve his message posts.

Exposure to the LAD had a positive impact on P2's learning strategy, even though he said it was not useful to him. He reviewed his work again after seeing that his initial post had no keywords. Further, he stated that overall, the LAD indicated that he really did not understand the required reading. As result, he planned to go back to the assignment and reread the theories that the assignment was based on. P2 concluded the interview with this:

> "What I'm trying to say is that like, this graph, it's like, like once somebody will submit his answer or her answer, this graph will show if it's going to be on a low point it will say like, it will let the person know that he or she will like, need to redo assignment, and like it will show like, the level of understanding it's like pretty low, and like you need to like redo it, and like which will make people go back to reading and read again, and like, understand and more deeper context."

P2 clearly understood the benefits of learning analytics visualizations, even if he personally chose not to reap said benefits.

P4 and P9's experience – Repeat users

P4 and P9 used the LAD repeatedly – accessing it 6 and 9 times respectively – but did so in very different ways. While P4 interacted with the LAD on an almost daily basis, P9 interacted with it 8 times within a two-hour time span, then once the next day. Both shared the chosen goal of getting high marks and saw lowered-than-expected ratings in their LADs.

Having begun online courses in elementary school, P9 was comfortable learning online, preferring it to face-to-face learning. His stated goals for the current learning activity were getting high marks and learning as much as possible; his interactions with

the LAD spoke to the strength of these goals. The LAD was not populated until the last 2 of the 9 times it was accessed.

It was unclear if P9 actually read the description of coherence when interacting with the LAD. Though he said he did, his body language and laughter sent conflicting messages. His understanding of coherence was that the LAD searched for keywords. Thinking the keyword rating was based only on the inclusion of the artwork names and key terms, P9 was surprised by a lower-than-expected rating on his second post (see Figure 35).

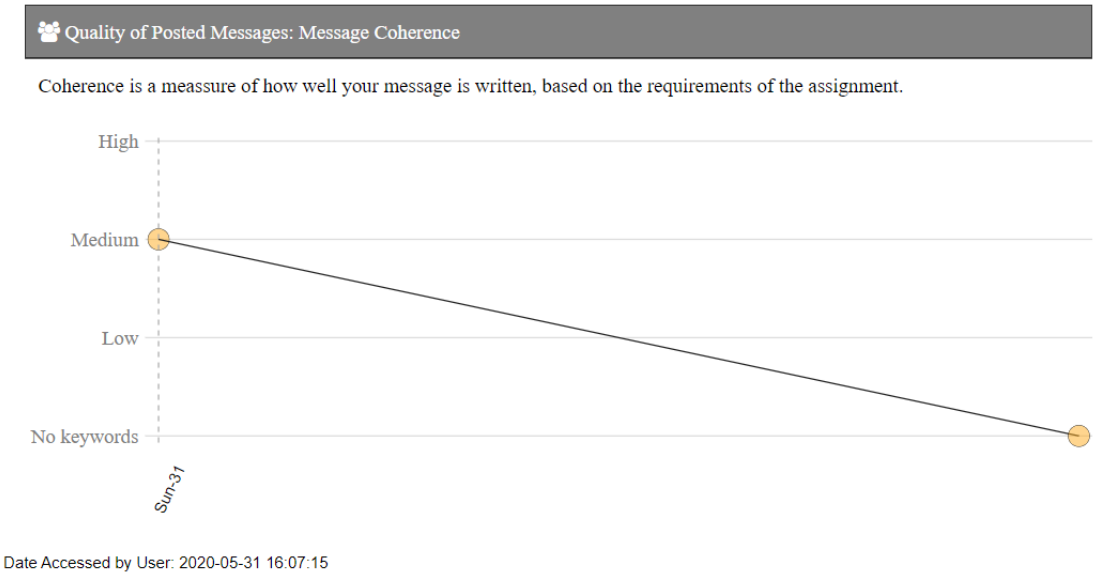

Figure 35. Exp. 5 first populated LAD seen by P9

After reviewing Figure 35, P9 went back and reread the article and the assignment rubric. Since he was unfamiliar with how the ratings were calculated and did not know what was "right or wrong with his post," he questioned both his internally held mental model of his performance and the visualized rating of his work. Peer comparison did not help him assess the quality of his work. When he referenced the LAD again the next day, he found no real difference between his work and that of his peers, nor did he have a sense of whose post received the high ratings depicted.

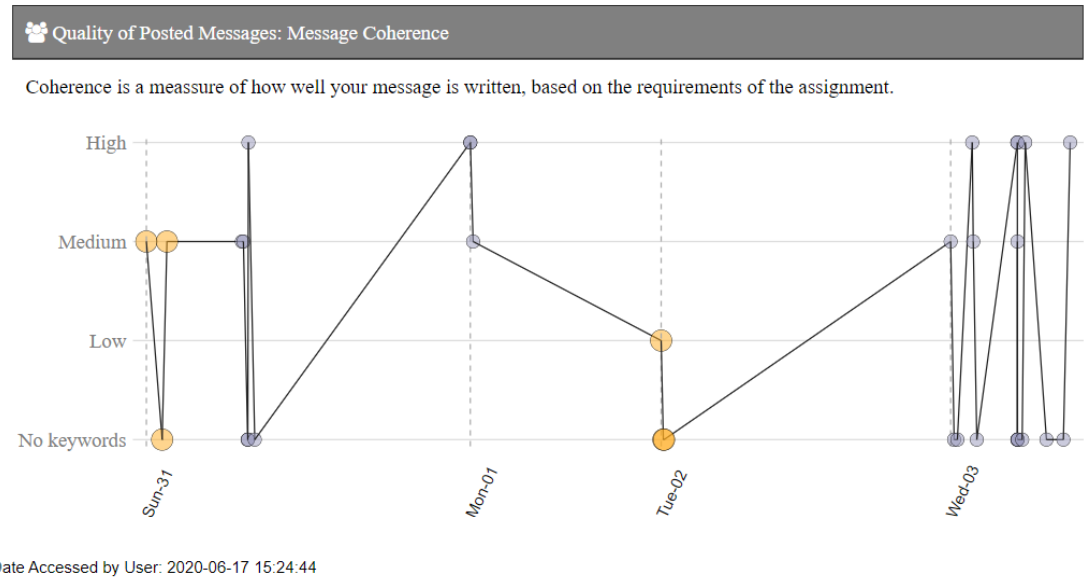

Figure 36. Exp. 5 P9's post-activity LAD group

Using the summative LAD in Figure 36, P9's perception of his performance changed. P9 said it was "kind of accurate," though he could not articulate why. He thought the visualization was helpful in providing perspective on how other people were doing, but not helpful for him personally because he didn't understand how the ratings were calculated. In reflection, P9 said that the visualization changed his opinion of his performance, by telling him that he was not doing enough.

P4 had no online learning experience prior to COVID. Like P9, her goal was to get a high mark. P4 accessed the LAD four days in a row — twice within the first five minutes, and once each subsequent day until the end of the learning activity, when she accessed it twice. The first time P4 accessed the LAD she and several of her classmates had already posted. She understood that the orange points represented her data, and that the gray ones represented that of her classmates, but she was surprised to see her second post at a lower rating than her first post. After reviewing her post, again she decided that the discrepancy was not big enough to warrant further attention. The next day she accessed the LAD by accident. The following day P4 returned to the LAD because her group member had sent a picture of her own visualization. Wanting to see if her own visualization had changed too, P4 accessed the LAD again. Until the interview, she didn't

realize that the LAD updated every five minutes. Similar to other participants, she was surprised to see a lower rating than expected on her most recent post, but was unable to determine why her third post had a low rating.

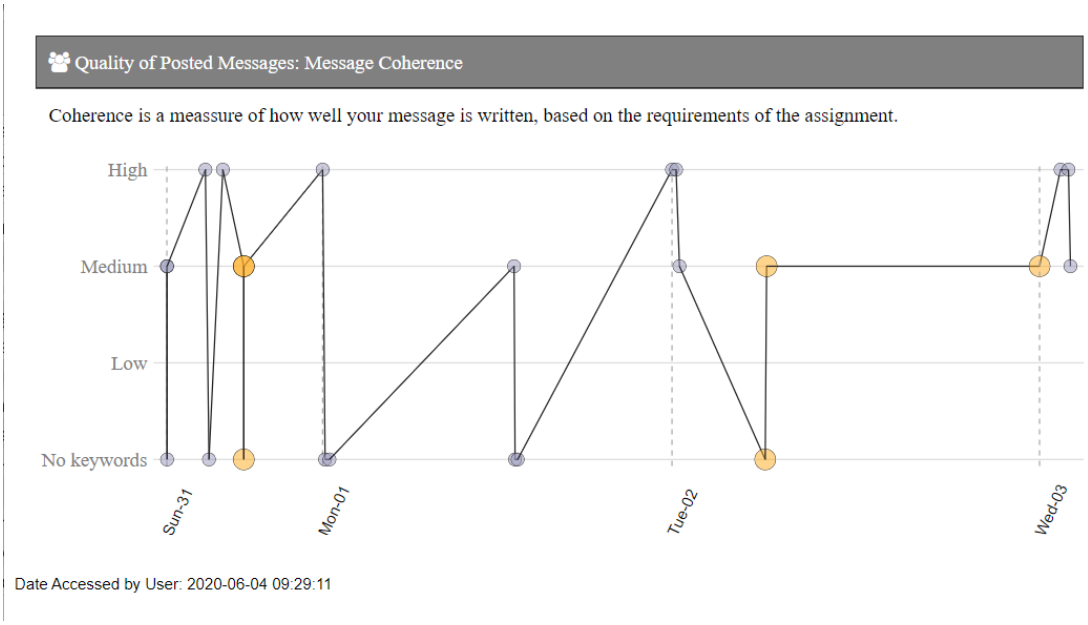

Figure 37. Exp. 5 final time P4 accessed the LAD

The final time she accessed the LAD in Figure 37, P4 saw that her last post, the longest post she had written, had a medium rating. This prompted her to think that ratings were based on the length of the post, in addition to quality. This explained why her previous posts had lower ratings, because her classmate who wrote longer messages had higher ratings than her. She felt like the LAD made two comparisons, each person to a standard, and the group members to each other.

Though she did not agree with the low rating of her posts, she felt like the visualizations were accurate. Her goals did not change, but based on the visualizations, her opinion of her performance did.

> Int: Did this visualization change your opinion of your performance that you had going into the assignment?
>
> P4: Yeah kind of, lower than my expectations. Like I don't give up. I know that I'm doing well in my other assignments, which are like more higher graded.
>
> Int: Could you say that another way? I heard that you mentioned your other assignments but I couldn't quite discern what you said earlier on.

P4: Oh I said that I am pretty well in my other assignments, so yeah, I don't think that this discussion will affect my grade so much.

Int: Okay so still the same goal, but what it reflected to you was that your performance wasn't as good in this discussion as you had hoped?

P4: Nnn, yeah, I'm saying a high grade because I believe that correlation between good knowledge and high-grade. So you can't get a good grade without high knowledge. So accordingly I will get a lot of knowledge from this course, and also I will get a high-grade.

In this excerpt P4 started talking about how the LAD lowered her opinion of her performance, but by the end, the statement morphed into how she would get a good grade in the course because she learned a good deal. In this passage P4 seemed to resolve her mistrust of the LAD by saying that it didn't matter, since the discussion did not represent a major portion of her grade.

In the interview P4 seemed overly eager to provide the responses that she thought were desired, even though she had difficulty understanding what was asked. Whenever she was asked to repeat herself it seemed as though she changed her answer, rather than reiterating what she stated earlier. In the excerpt above this tendency was less noticeable, but it may still be seen how this tendency could alter the interpretation of her statements.

P4 and P9 both got the gist of the LAD and acted on what they saw visualized, even thought their LAD usage patterns were quite different. The LAD changed both participants' mental models of their performance. For P4 this change was persistent; for P9 it was not. When asked, P9 had no answer for this change in behavior. Perhaps his distrust of the LAD was what led him to stop using it. It is also probable that he stopped using the LAD because, like P4, he did not know how to improve his work. It is notable that while P4 did not feel that the LAD affected her learning, she found it helpful. She explained that given the competitive nature of design, P4 thought that the LAD could motivate other students to work more, to compete with their peers.

P1's experience – Model user

P1 was unique in that she used the LAD the most of any participants, and she did so in numerous ways. Having previous online learning experience, her primary stated goal going into the discussion was to finish the assignment. Her chosen goal was to get a

high mark. The LAD was not intentionally part of P1's learning strategy – she first accessed it out of curiosity and did not initially understand it. Just having glanced at the LAD, P1 had not yet read the instructions.

> "Like I even don't know that that means the quality of the post, like those are already posted. I was just like, 'oh there's something there so this thing is like working,' but I actually don't know what this is supposed to mean for."

At this point in the discussion, she thought that the LAD measured relative differences in coherence between group members. It wasn't until she saw her own posts visualized that she understood coherence to measure all posts against the assignment requirements. This happened the third time P1 accessed the LAD.

> Int: So do you think it was with repeated use of the visualization that you figured out it was being compared against what was considered a good post, as opposed to your classmates? How did you come to that conclusion?
>
> P1: Yeah because I was posting something different from my group members, and then it's not like they show it as a medium or low quality, but it's showing my post as high-quality so I thought that it wouldn't be coherence between me and my group members, but the coherence between me and the correct answer.

This is when P1 realized that she could refer to the LAD to judge the quality of her posts. She soon developed a daily system of using the LAD. P1 would review her peer's posts, assign a rating in her mind, then check the LAD to see if her perceived quality matched the visualization. Considering this behavior, it was interesting that P1 preferred the LAD not show participants' names. She reasoned that her own feelings would be hurt if the visualization displayed posts that were not high-quality with her name on them.

P1 used the LAD to change the direction of her groups' discussion when she thought they were off-track, and to summarize her groups' work in preparation for the in-lab activity. In one instance when she noticed that no one who answered using a particular keyword had a highly rated post, so she "tried a different direction" and wrote about a different concept. As the wrapper, the person tasked with summarizing the groups' perspective, P1 used the LAD to identify high quality posts to use in the summary. P1 said that even if she wasn't the wrapper, she would have used the LAD to judge how successfully the wrapper performed their role. P1 even used the LAD after the

conclusion of the group discussion. In preparation for the in-lab presentation, P1 sent a screenshot of her visualization to her group members as evidence of the quality of her contributions.

Using the LAD influenced P1's learning strategies, goals, and internally held mental model of her performance. When asked if her goals changed from completing the activity or getting a high mark during the discussion activity, P1 had this to say.

> "Yeah probably I was just to like, get things right, it is not like the mark because I feel like the mark will not be too dissatisfying, I think, so is like, it will be okay as long as I posted. [Pause] But sometimes I would like to get it right, get the answer right. Its not about the mark, is just to get it right."

The LAD began to change her opinion of her performance early on in the learning activity – by her second post.

> P1: [Long pause] Yeah, it's quite interesting when I find that most of us are not doing well with the second question, and then when I post my second post to the second question and then it showed that my answer was of high quality. That was quite interesting. And I feel like it was, like, kind of fulfilling because I use this one, this technique, the system – how should I call it?
>
> Int: The visualization?
>
> P1: Yeah, the visualization, to get it right. Yeah. So it's like more fun, I have more motivation because I want to see like the other yellow got on the highlight. Yeah.
>
> Int: Normally would you be comparing your posts to your classmates? Would you have such a clear sense of how you were doing in comparison to them?
>
> P1: No, it's like normally if I'm not using this one [the visualization] and I might not check each other, everyone's posts. I might just whatever, [laughs] just post my response to whoever.

Fortunately, P1's gist estimates were accurate, as they shaped her mental models of her performance going forward and persisted after the conclusion of the learning activity. She was aware of the role of the LAD's feedback in her changed perspective.

> "But it's really helpful I think, because I'm not sure about the quality of my reply before, and also like, I wouldn't consider if my reply is good or not that much. It's like, and I wouldn't care about it that much, because we are just doing, we used to just do it in one doc and then submit it. That's not, I wouldn't, like, be getting feedback from this graph and then change my answer, so it is really helpful."

She mentioned being more confident in future interactions with her group and anticipated taking a more active role in planning future team projects because “I feel like I'm the one who knows more about this topic.” As she sat up straighter, smiling directly into the camera, P1’s newfound confidence was evident. She used words such as fulfilling, motivating, more confident, and fun to describe her experience using the LAD.

P1 used the visualization more often, and in more ways than any other participant. Reciprocally, exposure to the visualization affected her in more demonstrable ways than the other participants. It directly affected her learning strategy, goals, motivation, and confidence. In her attempts to steer the direction of the discussion and the quality of her posts, use of the LAD also indirectly impacted P1’s group. The third time P1 accessed the LAD was the turning point, when she realized what it measured and how it could help her accomplish her academic goals. From then on P1 used the LAD on an almost daily basis to update and improve upon her contributions to the group activity, to monitor the quality of her peers’ posts, to change the conversational direction of the group, and to quickly see if any new posts had been added – functionality that was not present in the LMS.

Comments and suggestions on the LAD design

To see how changes in the LAD designs could address their varied needs and perspectives, participants were asked to suggest changes that would improve the LAD’s intelligibility. Responses varied. Several participants suggested explanatory text be added, even though they did not read it when it was provided. P7 wanted to see a detailed explanation of how the LAD’s ratings were achieved. While transparency is particularly important when reflecting the caliber of learners’ work back to them, P7’s perspective was certainly a minority one. Most of the participants did not read three lines of provided text, so they likely would not read an additional explanatory paragraph.

There were also suggestions that would change the functionality of the LAD. P1 found the spaces between the data points distracting and suggested that time be omitted from the visualization altogether. This would downplay the significance that is visualized when there are long stretches of time between posts. P6 wanted to see the visualization during the process of composing her post, to be able to see the changes in the quality of

her writing. P2, P3, P5, and P6 asked that the LAD list specific information telling them how to improve their work.

Who got the gist and when

If participants got the gist of the LAD, they tended to do so by the second or third interaction (see Figure 38). As in experiment 2, if the LAD was populated when first viewed, this contributed to participants' understanding of what was visualized. P9 was the only person who understood the LAD from the start because he read and understood the provided descriptions. P4 partially understood the LAD when she initially accessed it; this was due in part to it being populated when she saw it. She was not sure what the grey data points represented until the third or fourth time she accessed the LAD. Seeing their own posts visualized also helped P6, P9, and P1 get the gist. P2 partially understood the LAD – though he thought both the grey and the orange data points were his – but got enough of the gist to be able to understand that his posts were rated lower than expected. P7 understood the LAD when he read the descriptions, but because he did not know how the ratings were achieved, he did not fully trust what was visualized.

The participant who was most successful with the LAD, P1, initially understood the LAD to measure differences in coherence between group members. It was not until the third time that she accessed it, when it included visualizations of her own posts, that she understood it correctly. P1 was the only person who knew how to improve upon the quality of her work and once she started using the LAD in this way, she kept doing so.

Interestingly, most participants acted on what they saw. They attempted to improve the quality of their posts but not knowing what to do, stopped acting on what was visualized even if they kept viewing the visualizations.

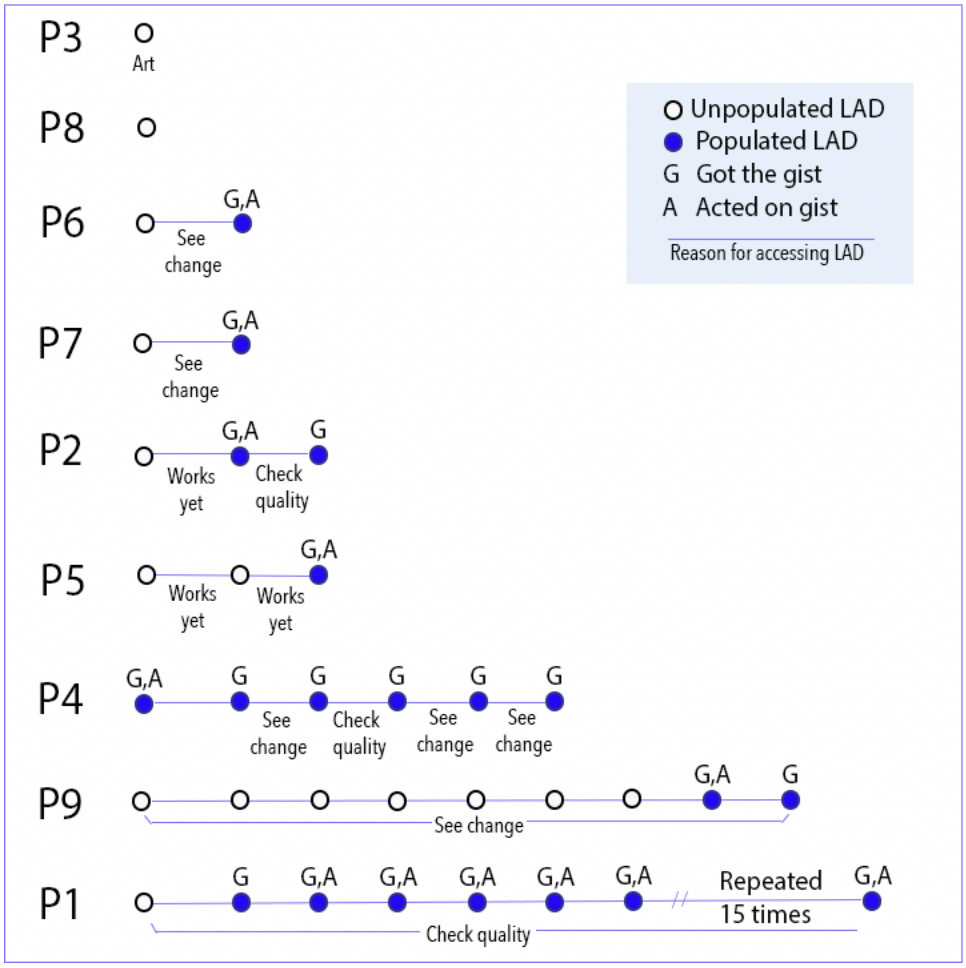

Figure 38. Exp. 5 LAD views demarcating when participants got the gist

9.7. Discussion

Participant responses established the context of the learning activity, their emotional states and metacognitions about the learning experience. The semi-structured interview format allowed unscripted questions, which in turn provided information about the learning activity that the interviewer would otherwise not have been privy to. For example, asking P1 if her group members used the LAD led to her revealing that she shared screenshots of her LAD to convince her group members of the quality of her work. Adjustments were made in the interview, such as the post-activity LAD, to better capture aspects of the participants' experience. With this summative LAD, we were able to see if participants were better able to make gist assessments with fully populated visualizations. It was particularly useful for interrogating the understanding of gist for

participants who viewed the LAD 3 times or less, and those who saw unpopulated or sparsely populated visualizations during the learning activity.

Language difficulties made conducting and interpreting the interviews far more challenging than anticipated. It is certainly plausible that the received responses were less verbose or expressive than they would have been if participants had greater English proficiency. That said, no further presumptions can be made – it is possible that even with improved English proficiency, they would not have been successful with the LAD.

With one exception, where sequences of stuttering from the participant with a pronounced speech impediment were omitted for legibility, the interviews were transcribed verbatim. The inclusion of false starts, hesitations, and filler words such as "um" and "like," aided in the interpretation of participants' intentions and underlying emotional states. Entire turns of talk were reported to frame responses as answers to specific questions (Potter & Hepburn, 2012). This identified the researcher's active role in the interaction, and further aided interpretation. We recognize that interview data amplify contextual meaning, meaning produced in that moment, that may not reflect stable perceptions or attitudes (Bourgeault et al., 2020). In this sense interview data are always incomplete; they also reflect the researcher's worldviews at that point in time (Andrews, 2020).

The role of prior visualization or distance learning exposure

Neither previous online learning experience nor previous exposure to visualizations effected LAD use or gist production, though prior experience did seem to effect how learners approached the discussion. All of the participants had some sort of distance learning experience due to COVID; only three had not participated in a course designed to be administered online. There were certain practices that those with prior experience associated with online learning, such as timely replies and the idea that all posts must be read, even if a reply was not written. In the example of P8, who assumed the grading policies would be the same as those in his previous online course, prior experience actually did him a disservice.

Group interactions

Knowing how the learning activity was completed helped us better understand the influence of the group on each individual participant's understanding of the learning objectives, the assignment requirements, and most importantly, their perception of their performance relative to the group. The difficulties experienced by the participants who joined a group of strangers speaks to the need to structure group learning activities in a way that is more supportive. By the time they figured out how to work together the learning activity had ended. Eddy et al. suggest staying in groups for a few weeks; this encourages group bonding. It also helps to explain why the activity is being performed in groups, so students understand what they are meant to learn from the experience (Eddy et al., 2015).

The design of the LMS hindered the performance of the group discussions and ultimately, what was visualized for learners. Since it was not possible for learners to save their posts, they often composed their messages using external software, then copied and pasted their work into the LMS. Group discussion activities were not intended to be conducted this way, outside of the LMS, and the effect this had cannot be overstated. As designed, the LAD was meant to visualize each learner's contributions to an ongoing learning activity, providing formative feedback to learners during the process of learning. When discussions were performed outside of the LMS, the LAD could not capture the true nature of the groups' interactions as the learning activity progressed. The true nature of the conversational exchange and learners' individual demonstrations of knowledge went unseen; what was visualized instead reflected a staged exchange. What's more, in the groups that used these outside discussions to heavily strategize the learning activity – achieving consensus by omitting divergent or redundant opinions – the visualized post coherence could seem artificially low because portions of the naturally occurring discussion were absent. In doing so these learners adhered to the maxims of quantity and manner in Grice's cooperative principle of conversational communication (Cole & Morgan, 1975), but altered the feedback they could have received from the LAD.

Numeracy & goal orientation

The interrogations of goal orientation and numeracy happened in much the same way, in that qualitative and quantitative measurements of each construct were collected. The results speak to the benefits of obtaining qualitative feedback on quantitative measures if at all possible.

Four of the five participants' goal orientations were not persistent, and for the participants whose goals remained unchanged, their stated and chosen goals were not aligned. In this study, initial goals were not as indicative of behavior as the change in participants' goals. The goal orientation results demonstrate the difficulty in attempting to capture academic goal orientation using a questionnaire administered once. This difficulty may be attributed at least in part to differences in short versus long-term goals. While learning is much as possible may be a long-term goal, short-term needs such as completing the current assignment on time to avoid a low mark, may take precedent.

The average Subjective Numeracy Scale composite rating was 3.89. Comparatively, the score of participants used to validate the scale was 4.2 (Zikmund-Fisher et al., 2007). Though participants' SNS scores matched their descriptions of their numeracy, there were not enough participants to see if a correlation existed between the numeracy and the number of times the LAD was accessed, or accurate and complete gist assessments of the LADs. This could be partially attributed to inflated numerical preference results that indicated difficulty conversing English, rather than an actual preference. The qualitative numeracy feedback illuminated the thought processes of participants that formed the foundation of their selections on the quantitative scale. The qualitative descriptions indicated the social and experiential bases for decisions about two seemingly simple questions – how participants perceived their quantitative ability and if they preferred information displayed in numbers or in words. This highlighted the thought that must go into the qualitative or quantitative measurement of a phenomenon.

Initial access – the zero-start problem

The majority of participants initially accessed the LAD because of curiosity; only one participant did so to intentionally examine the quality of his work. Five participants saw an unpopulated LAD when they first accessed it. Of those five, the single participant

who read the provided text understood why it was not populated at the beginning of the learning activity. Similar to what happened in Experiment 2, two participants thought that the unpopulated LAD did not work and did not attempt to use it again. The other two participants did not devote even that much thought – they ensured that something appeared when the link was clicked and went on to other tasks. Four participants saw a populated LAD when they first accessed it; two of them thought it was a piece of interactive art after looking at it, and the two others only noted that it changed.

This is not what was expected – we anticipated more students intentionally accessing the LAD to use it as designed. We thought the unpopulated LAD was the critical point for understanding, not realizing that just as many students would have difficulty understanding the populated visualization. Two participants who saw populated visualizations still thought they were pieces of interactive art – this shows a lack of critical thought that we did not anticipate. We thought that the provision of explanatory text would positively contribute to participants' understanding of the LAD, helping students recognize the value of the LAD as a learning tool earlier on. In this sense, the explanatory text was not as effective as intended because it only helped one person. When asked, the participants who experienced difficulty were unable to verbalize their initial expectations, aside from the LAD being populated. More commonly, the participants who read the provided text descriptions only did so after repeated exposure to the visualization. This leaves the zero-start problem unsolved – if learners will not read explanatory text when they need it, how else will they understand that the LAD is working as intended when unpopulated? As designers, what can be done to maintain students' curiosity, while simultaneously communicating the value of LAD as a learning resource, to support repeated use? Two options are the provision of tutorial tool tips that must be watched before access to the LAD is allowed, or a video introduction that explains the tool.

Expectations and task valuation

When participants began the discussion activity they did so with few expectations. Aside from the expectation of the LAD being populated, they tended to expect marks higher than what they saw in the LAD, even though the keywords were never explicitly

stated or revealed during the discussion. Of the 7 participants who understood that the visualization rated their posts during the discussion, 6 saw ratings that were lower than expected. For those who understood the visualization, most commonly their mental models of their performance changed but did not persist – either because of mistrust or participants not knowing how to improve their work. If they got the gist they acted on it, going back to review their work against that of their peers. Only P1 changed their behavior, adding to her messages or altering the trajectory of the discussion.

The participants whose action stopped at reviewing their work either did not know how to improve their work or decided that the learning task did not warrant further effort. From the goal orientation of achieving high marks the viewpoint of participants in the latter group was correct – changing the rating of a single message would not do much to change their overall score. This learning activity was chosen for this study specifically for the relatively low stakes task. One participant, P2, alluded to the fact that he did not spend enough time viewing the LAD to properly understand it, because the learning activity would not have much impact on his overall grade. Further, he said he would only access learning analytics in future courses if he thought it would have a major impact on his grade – he would not use learning analytics if they were solely provided for informational purposes.

Trust

Mistrust was a prevalent theme throughout the interviews in this study, although to a lesser degree than in previous experiments. P9 distrusted the quality ratings as a matter of course, even when his post was rated as expected. Other participants who mistrusted the LAD did so because they saw lower than expected ratings.

P1's experience was different in that when she saw a rating that did not match her expectations – in this case a higher than expected rating – she mistrusted her understanding, not the LAD. Early on in her interactions with the LAD, her gist understanding was that the ratings were relative, comparing the members of the group to each other. This would make the ratings variable; they could potentially change based on the most recent posts. Even though her understanding of gist changed with repeated use, that initial gist interpretation persisted as a *feeling*.

P1 described it as a feeling that she could not explain, one that she had even though she knew that the ratings were not relative because of her repeated interactions with the LAD. Had the feeling been merely uncertainty in the quality of her answers, it would have gone away as her confidence increased with repeated LAD interactions, but this was not the case. Since she stated that the visualizations increased not only her confidence but also her motivation, we categorize the persistence of the incorrect gist interpretation as mistrust.

Skill and awareness

Overconfident estimations of performance may stem from a lack of awareness; they could also be attributed to the better-than-average heuristic (Krueger & Mueller, 2002). Simply put, the better-than-average heuristic describes the tendency to perceive oneself superior to others, to skew toward the ideal rather than realistic portrayal of one's skills or abilities (Alicke & Govorun, 2005). This is different than an overconfidence borne of a lack of awareness, awareness that most often results from performance-related metacognition. The Dunning-Kruger effect describes the tendency to overestimate performance as seen specifically in the lowest performers (Kruger & Dunning, 1999); the least competent individuals overestimate their performance because they are completely unaware of their lack of skill. The more competent the person is, the smaller the discrepancy is between their perceived and actual skill.

Overconfident performance estimates are common not only with students (Dunning et al., 2003; Grimes, 2010) but across age groups and professional ability (R. K. Edwards, Kellner, Sistrom, & Magyari, 2003). This effect was evident even when monetarily incentivized (Ehrlinger et al., 2008). As students are developing discipline-specific expertise, they are also building their capacity to correctly assess their performance and gaps in their knowledge. It is expected that they may have deficiencies assessing their performance – this is the reason for many student-facing learning analytics, including the LADs used in this study.

In this study we saw two types of participants act on visualized deficiencies in their performance – those who were unaware of their inability to correctly assess their performance, and those who were aware of a gap between their desired and actual

performance but did not know how to close it. Participants like P6 who could not discern what made a message high quality had no anchor to use as the foundation for the mental models of their performance. Had they known what to do to improve their work, it is possible that they would have done so.

Chapter 10.

Limitations

Each of these experiments garnered feedback from a different subset of the learning population; the feedback was authentic but not necessarily generalizable to the entire learner population. As stated earlier, the wording of the interview solicitation and the compensation offered may have increased the likelihood that these studies attracted lower performers. Had final grade data been collected, we would have more evidence to support or refute this claim.

A major limitation of this work was our reliance on self-report measures, which are subject to selective recall and perception, memory bias, and change. Self-reports are static, and therefore cannot measure changing phenomena. This was particularly evident in experiment 5, observing the differences in goal orientation before, during, and after the learning activity. Goal orientation changed most often due to contextual factors unrelated to the subject matter of the learning activity. Additionally, at least two participants consciously inflated their numeracy results because of language difficulties, rather than a true preference for numeric information. The self-reported information was accurate but needed to be contextualized to truly understand the participant perspective.

Another limitation is the reductionist nature of the data being used to signify learning; this is part of a larger social criticism of learning analytics (Selwyn & Gašević, 2020). The LADs visualized post quality according to the coherence of the keywords utilized in each discussion post. It captured observable artifacts of learning; passive or observational learning from peers within the discussion was omitted. As we saw in experiment 5 with the groups who performed the discussion outside of the learning management system, there is a danger in reducing the definition of learning to only that which may be observed.

Though designed aspects of the LADs highlighted a personalized perspective – the color of the data points meant to differentiate between the individuals and their peers and the omission of individual names – the LADs did prioritize peer, over self-

referencing. This could have caused emotional distress for some learners (Lim et al., 2019), especially those who preferred a self-referenced visualization or were doing poorly in relationship to their peers, causing them to avoid further interaction with the LADs. There is also a danger of peer-referencing leading to a surface, rather than mastery, approach to learning (Lim et al., 2019).

Chapter 11.

Conclusions

Learners, particularly those learning online, often experience difficulty when trying to accurately assess their learning. Having already formed an idea of the difficulties experienced by online learners in evaluating their work in my role as an online instructor, this research was undertaken to design LADs to support these learners. This work ultimately contributes to the design of LADs that learners will be able to successfully utilize as part of their self-regulatory learning strategy.

The brief attention learners devoted to their LADs fostered a desire to better understand how learners interpret what is visualized in LADs and what they do with this information, Gist was operationalized to describe what learners understood from brief LAD interactions. A mix of quantitative and qualitative methods were used to identify why, when, and how learners interacted with LADs to guide their learning.

Though participants for each experiment were a different subset of learners each time, several recurrent themes were observed across the studies. Specifically, learners tended to:

- Allocate scant attention to reviewing the LADs
- Experience difficulty interpreting the LADs
- Not read explanatory text provided with the LADs
- Not think critically about the LADs, especially when lower-than-expected ratings were displayed
- Mistrust ratings displayed in the LADs, particularly when the LADs displayed lower-than-expected ratings
- Demonstrate gist accuracy that improved over time and with repeated exposure

In the following sections we briefly revisit the results of each study to contextualize the conclusions.

11.1. Pilot & exploratory studies

The pilot study results showed task accuracy to be functionally equivalent across all three LAD types – the bar chart, heat map and landscape visualization – with a mean accuracy of 61% (SD = 24%). Though the study did not have enough participants to achieve significance, lessons learnt led to structural improvements in subsequent experimental designs. The high dropout rates were attributed to the objective measurement of numeracy, supporting the results of Fagerlin et al., who found that participants preferred subjective tests of numeracy (Fagerlin et al., 2007). With more participants and task trials, perhaps it would have been possible to observe a relationship between time on task and accuracy, based on visualization type. Since the future direction of the research changed, it was no longer necessary to pursue time-based measurement of LAD interpretation.

Participants were solicited through convenience sampling (Lewis-Beck et al., 2003) from the researcher's peer network for the pilot study, resulting in a greater number of graduate student participants than desired. Nevertheless, the results were informative. It was observed that education level had little bearing on either task accuracy or gist assessments made; both student populations had similar results.

We anticipated that participants would exhibit greater task accuracy with their most preferred visualization type, but saw the opposite transpire. Participants attended to the most familiar visualization, the bar chart, the longest. They selected the bar chart for most useful and aesthetically appealing, but their task accuracy did not reflect these preferences. Though they attended to the landscape visualization for the least amount of time, participants demonstrated the highest gist recall with this LAD type. This performance could not be attributed to time on task, as participants chose to attend to the landscape visualization for the least amount of time.

The relatively low task accuracy — especially in combination with such simplified task estimates and visualizations — prompted further questions. How was it that when presented with a bar chart and four multiple-choice questions, graduate students' performance assessments were accurate only slightly better than half of the time? As expected, participants attended to the bar chart visualization longer. Why didn't

this prolonged attention result in increased accuracy? As the simplest and most familiar visualization type, it was even more surprising that there was not a pronounced difference in accuracy with the bar chart. The most abstract visualization, the landscape, was the one that they attended to for the shortest amount of time. Why then was the gist of this particular visualization type more memorable? None of these questions could be answered within this experiment, because of the experimental design. What was gained in convenience in administering the survey online was lost in the ability to interrogate the responses garnered. To better understand learners' interpretation of LAD, and the resultant behaviors seen with them, we needed to be able to directly ask learners what motivated their decisions. We recognized that we needed to know how learners *actually* made decisions with LADs, not what they would do from the perspective of fictitious learners using generated data to make learning decisions.

The exploratory study allowed us to interrogate how learners performed visual interrogations with LADs. This sets this study apart from those that test usability with generated data, removed from the process of learning. The employment of retrospective cued recall methods aided participant recall, serving as a learning artifact that learners could use to guide their reflection upon their LAD interactions. Commonalities were seen in how learners initially approached the discussion activity, their initial impressions of the LADs used during the activity, and of the LAD prototypes. Almost half of the participants (N=17) had no reason for why they accessed the LAD; many returned to use it a second time because of curiosity. Social influence motivated several participants to initially access the LADs, with 4 participants seeking out the LADs based on a friend's recommendation and 6 accessing it to compare their work to their peers. Surprisingly, only 2 learners access the LADs to assess their own work.

The majority of the interviewed learners tended to 1) wait for others in their group to post messages before contributing to the discussion themselves, 2) to describe their performance from a peer-referenced perspective, and 3) to have difficulty interpreting the LADs or the data visualized within. Feedback from the interviews revealed that learners – even those who successfully utilized the LADs – only briefly attended to them. This prompted the subsequent study of gist, to determine what learners understood from these brief LAD interactions. Thus far many student-facing LADs have been designed with the

assumption that learners understood the data depicted for them, however we found this to not be the case. This pointed to a need to measure how learners *interpret* LAD. In the second phase of the study learners' perceptions of aesthetic appeal and usefulness changed with use as anticipated. After learners were exposed to the visualizations, they were more apt to choose a visually stimulating or affective visualization type.

From this experiment on, all subsequent studies used homogeneous learner samples – learners were all undergraduates in the first or second year of university study, enrolled in classes from the same department. This was done purposely, to be able to illustrate characteristic behavioral patterns of learners' LAD interactions and to make comparisons between them. From there however, participants were self-selected. This self-selection could have been influenced by the compensation offered for study participation. The pilot and exploratory studies offered monetary compensation; for the rest of the experiments course credit was offered. This may have resulted in a sample that was less representative of the student body overall, as the study might have attracted participants who needed extra credit in their courses. This was deemed acceptable, as perhaps they are the best served by learning supports such as LADs.

11.2. Experiments 3 & 4

MTurkers produced more accurate and complete gist responses than learners in experiment 3 using the abstract visualizations. Learners' responses were detailed in their descriptiveness but were incomplete, frequently lacking an assessment of their own performance from the perspective of the fictitious student. Accuracy and descriptiveness results according to visualization type were mixed. Performance in terms of accurate and complete descriptions was better with the mountain visualization for both populations. Cell plots were constructed for both populations to see if the provision of accurate gist responses was the result of learning effects. The cell plots indicated that in both populations, participants with the highest number of accurate responses provided their first accurate response within the first 10 responses. These results suggest that repeated exposure and practice with LADs may have a positive influence on learners' ability to produce accurate and complete gist assessments with them.

In experiment 4 MTurkers provided more accurate and complete responses than learners with traditional visualizations, though the difference was not as pronounced as it was in experiment 3. Again learners produced responses that were highly descriptive, but lacked summative judgements of gist from the perspective of the highlighted student. There was no statistically significant relationship between visualization type and the provision of accurate and complete responses for either population. Cell plots of the provision of accurate and complete responses indicated that the most successful participants in both populations produced their first accurate response within the first five responses. These results support the premise that repeated exposure and practice with LADs may positively affect learners' accurate assessment of them.

When the results for experiment 3 and 4 were combined, a statistically significant difference was seen in the production of accurate and complete responses for MTurkers and learners, with MTurkers performance exceeding that of learners. There was also a statistically significant difference between the abstract visualizations of experiment 3 and the traditional visualizations of experiment 4 for both groups. This lends credence to the argument that "chart junk" aids sensemaking (Bateman et al., 2010). Though we suspect it is due to the popout effect (Treisman & Gelade, 1980), novelty or aesthetic appeal (Chatterjee & Vartanian, 2014), identifying why it happened is left to future study.

11.3. Experiment 5

When prompted to describe their goals at the end of the interview, learners examined the persistence of their goals after what was, essentially, a period of guided reflection on their learning experience. As an awareness tool, the LAD in experiment 5 functioned as designed. The majority of learners in that experiment were aware of discrepancies between the actual and expected coherence of their message posts. As a regulation or reflection tool, the LAD failed all but P1. P1's experience with the LAD was the exception. In using the LAD to regulate her learning, she had an experience akin to that described by transformational learning theory (Ally, 2008). Her interactions with the LAD not only created a persistent change in her mental model of her own work; it imbued her with confidence and increased her perceived efficacy.

Learners tended to act on the gist that they understood, even though their actions had little effect on their learning. Multiple learners made attempts to improve their work but stopped just shy of changing their message posts. Similar to experiment 2, several of them stated that they either wanted or expected the LAD to give them specific instructions on how to improve their discussion posts. It is possible that these learners would take further steps toward enhancing their work if they knew what to do. An *inability* to think critically and an *unwillingness* to think critically are quite different, however they both result in inaccurate or incomplete gist assessments.

11.4. Factors of individual difference

At the beginning of this series of experiments there were four factors of individual difference identified. For the sake of timeliness and participants comfort, the Berlin Numeracy Test was dropped from the study instruments representing factors of individual difference after experiment 2. Even though it was short, we found that participants did not enjoy taking the test, similar to Fagerlin et al.'s findings (2007). Further, we attribute the high dropout rate seen in the pilot study to this measure.

Though relationships were seen between subjective numeracy and spatial acuity, as measured with the PSVT-R, this instrument was dropped as a measure of individual difference because of the amount of time required to complete it. Relationships were also observed between objective numeracy and cognitive reflexivity, and objective numeracy and spatial acuity in experiment 2. The well-known nature of the Cognitive Reflection Test – several participants in experiment 2 reported having taken the test before and there was a good chance that MTurkers had also been exposed to it (Chandler 2013; Haigh, 2016, Thomson & Oppenheimer, 2016) – raised concerns that that prior exposure might skew our results. This led to only the SNS being used as a factor of individual difference in the subsequent studies.

Had we been able to collect numeracy information from learners across studies, we would have been able to use it to categorize learners with this measurement. Although the data collected from experiments 2 and 5 indicates that learners' preference for numerically presented information may be influenced by their ability to express

themselves in English, this nevertheless reflected learners' subjective numeracy. In future studies it would be interesting to look at the relationship between accurate assessments of gist and numeracy.

11.5. Gist

In the pilot study the assessment of gist was measured by correctly or incorrectly answering a reflective question, without being able to revisit the visualization, after the passage of approximately 28 minutes. After a qualitative examination of gist in experiment 2 through the interrogation of learners' experience with LADs, we quickly realized that quantitative assessment of gist did not capture all of the information necessary to truly understand how learners make sense of their learning performance with LADs. This new definition of gist was also influenced by contextual factors, i.e. the brief amount of attention that learners devoted to reviewing their LADs in experiment 2. The 30 second time allotment we used for gist assessments was similar to that used with the Visualization Literacy Assessment Test (Lee et al., 2016), which allows 25 seconds per item. This time frame was extrapolated from the amount of time MTurkers spent answering items in their pilot study. There are additional similarities between our gist assessments and the VLAT, such as the visualization and task types used. Their test use 12 visualization types — including a bar chart, stacked bar chart, and pie chart — and asked participants to perform tasks such as identifying trends or anomalies, determining range, and making comparisons. In operationalizing gist as the understanding learners obtain in their first moments reviewing the LADs, we hoped to better understand the mental models learners formed of their performance through interactions with LADs. We acknowledge that in the brevity of our gist assessments there was likely a trade-off between accuracy and speed, however learners are making judgments in situ with LADs in a similar amount of time.

The perceptual and gist-related research that informed this work all took place in controlled environments, and employed tightly controlled, elementary perceptual tasks (Cleveland & McGill, 1985; Correll & Gleicher, 2014; Elzer et al., 2006; Heer et al., 2010; Kosara & Ziemkiewicz, 2010; Quispel & Maes, 2014; Skau et al., 2015; Skau &

Kosara, 2016; St-Cyr & Hollands, 2003; Talbot et al., 2014; Clarke & Mack, 2014; Epstein, 2005; Josephs et al., 2016; Loschky & Larson, 2010; Mack & Clarke, 2012; Oliva & Torralba, 2006; Sampanes et al., 2008; Wu et al., 2014). The LADs we designed to test gist in experiments 3 and 4 used complex, realistic data that varied in the number of data points and learning paths visualized. This introduced some variability in the difficulty of the gist assessments, but it also increased the ecological validity of our results.

11.6. Critical thinking, the missing component

In experiments 2 and 5, the in situ experiments, we anticipated learners utilizing the LADs in a variety of ways. As part of their self- or social regulatory strategy in a group learning activity, we anticipated that learners would use the LADs to set and regulate their goals. The LADs could also be used to manage communication between group members, with learners using this information to augment how, when, how often, or with whom they chose to communicate. Learners could also use the LADs to make self or peer-referenced assessments of their academic performance. Information gleaned from the LADs could foster time management, information or help seeking behaviors, or lead to motivational or emotional regulation. This all would require a level of critical thinking rarely demonstrated by our university subjects.

On rare occasion, we did see learners interacting with the LADs as anticipated. In experiment 5 we saw P1 using the feedback from the LAD in all the ways that we expected, and in some ways that we did not. Most often, she used the feedback from the LAD to test hypotheses about the quality of her work and how to improve it. We assumed, wrongly, that if learners saw a visualization displaying a lower-than-expected rating of the quality their work, that they would attempt to raise this rating in one or more ways. In a discussion activity, this could be accomplished by multiple means. Learners could search for new or alternative keywords to use in their discussion posts. They might choose to engage with the learning material in new ways, revisiting their previous reading, or searching for alternative sources for the learning material. Within the discussion they could choose to post more often, to change the length of their posts, or to

engage in discussion with a different peer. They could choose to discuss the visualized results with peers, or ask their peers or instructor for help. Given how readily learners compared their work to that of their peers within the discussion — 7 out of 9 went so far as to reread their peers' discussion posts — it is surprising that only P1 discussed the visualized feedback with their group.

Similar behavior was observed in experiment 2, though that experiment featured two different LADs. A variety of interaction patterns emerged in both experiments, ranging from those who interacted with the LAD once to those who accessed it several times to help them regulate their learning. In both experiments, multiple individuals misidentified the LADs and did nothing to improve their understanding, such as reading the title or provided tooltips. We saw learners who understood the LADs only after repeated interaction in experiments 2 and 5, and MTurkers repeatedly providing more complete and accurate gist assessments than learners in experiments 3 and 4. Taken together, our observations prompted us to revisit our conceptualization of engagement.

We postulated that prior to interacting with the LADs, learners would have their own internalized perceptions of their learning and themselves as learners. As demonstrated in the second phase of experiment 2, learners may be motivated to initially interact with the LAD due to its perceived usefulness or aesthetic appeal. Once the learner interacted with the LAD their mental models of their performance may or may not change. This could be dependent upon multiple factors such as the strength of their initial perceptions, external feedback such as that received from an instructor or peers, or the result of comparatively referencing the quality of their own posts to peers using the LADs. These mental models would in turn influence learners' judgments of learning, resulting in changed engagement patterns. Especially upon seeing a lower-than-expected rating, we thought that learners would choose to engage differently with the learning materials, their peers, or the discussion activity. While some learners' mental models changed, most often, interaction with the LADs did not trigger changes in engagement that would positively affect learning.

What was omitted in the behaviors that we observed and in our expectations of engagement was the element of critical thinking. We utilize the five levels of feedback described by Gibson et al. to describe where the learners in our studies faltered (Gibson et

al., 2017). The five levels were impression, interpretation, internalization, integration, and intention. Forming an impression, the learner is able to determine what is happening around them, and what is important to them. The interpretation level is where learners make sense of their current situation. Learners determine how what is happening relates to them, their goals, learning, emotions, or knowledge during the internalization level. During integration learners determine how this fit with other knowledge, experience, or differing perspectives. In the final level, intention, the learner determines their new perspective and possible action based on what they have come to understand through reflection. In our studies even if learners did not understand the gist of their own performance from the LADs, they were generally able to form an impression of what was happening, i.e. how active the small-group discussion was. Where they were lost was interpretation, internalization, integration, and intention.

In experiments 3 and 4, MTurkers consistently provided more accurate and complete descriptions of gist than learners. Learners seemed unwilling or unable to think critically about what the data represented. In verbose descriptions of the features present in the LADs they provided their impressions, stopping short of interpretation and expending a minimum amount of cognitive effort. In the gist descriptions of experiment 5, learners tended to interpret few aspects of the visualizations in the description of their performance. Had they done so, perhaps they would have been able to better understand the lower-than-expected ratings seen. Only one learner, P1, vocalized behaviors or thought processes that belied internalization, integration, and changed intention.

The visualizations utilized in experiment 5 were far simpler than any of the LADs used in the preceding experiments, yet learners still experienced difficulty interpreting them. The gist results from experiment 5 are better considered with respect to the results from the previous studies. Overall, learners in experiment 5 paid little attention to the LADs. This was evident in the lack of attention given to the descriptive text, and in the disregard for rooting out the cause of discrepancies when a less-than-expected rating was received. Further, this was in line with the results seen in experiment 2 in situ, and in experiment 3 and 4, in learners' provision of descriptive but not interpretive summative gist assessments.

This is not to say that overconfidence did not contribute to learners' behaviors – it was observed in the experiment 5 participants who held idealized, unwavering opinions of the quality of their work. Overconfidence is a common occurrence (Edwards et al., 2003; Ehrlinger et al., 2008), particularly within student populations (Dunning et al., 2003; Grimes, 2010), and is remedied with improved judgements of learning. Only then will learners be able to better identify their knowledge gaps. In these experiments the provision of formative feedback from the LADs often was not enough to sway the opinions of overconfident learners. They simply needed more help than the LADs could provide.

Learners had difficulty understanding the LADs in all of the experiments. Even after they understood the information depicted — here referring to experiments 2 and 5 — learners produced inaccurate gist assessments even with highly simplified visualizations. Though possibly attributed to a lack of critical thinking or engagement, the number of responses that mentioned problems understanding the LADs might also be attributed to how the questions were posed in our studies. When we asked learners about their experience of the LADs, we used language that equally valued positive and negative feedback. As such, it is possible that we received more feedback than we otherwise would have if a different instrument had been used to gather this usability feedback.

LADs are commonly assessed by learners on their perceived usability, i.e. a perception of usability not based in actual use. When this feedback is solicited it is often assessed with Likert-like scales, in response to questions such as "was the LAD easy to use," (Mouri et al., 2017) and did they "experience any major problems" (Ruiz, 2016). The experiments detailed herein would have been quite different if usability had been measured this way. As we saw in experiment 5, learners wanting to give positive feedback or who want to avoid embarrassment are not going to say that they experienced issues. Further, what if the learner *thought* they understood the visualization, but they did not? If the learner is not aware of a problem's existence, how are they to answer such questions? We saw this transpire in experiment 5. The answer to "did you understand this," would be quite different from the response to "tell me what you understood from this visualization." If asked if they "found the LAD useful" (Castro et al., 2007), how is a learner who sees the potential but not immediate usefulness of the LAD to respond? The

Evaluation Framework for Learning Analytics (Scheffel et al, 2017b) instrument addresses this most clearly, and the closest it comes to evaluating usability are items that ask if "it is clear what data is being collected" or if it was clear "why the data is being collected." This speaks to the need to approach LAD usability evaluations differently, beginning with the assumption that the learners can correctly interpret them.

11.7. Learners and the learning context

The learning activity was not designed with the instructors of the courses invited to participate in the studies reported in this dissertation, however it was designed with the input of an instructor from that department. The content of the learning activity was then adapted by the individual instructors teaching the sections of the courses studied in experiments 2 and 5. The activity remained the same – it remained a graded small group discussion ending with a summative conclusion. The discussions utilized in this research were graded for both quality and quantity, but learners tended to focus more on the aspect of the assessment that was more easily understood, the quantity of discussion posts provided.

Instructors assumed that learners knew how to correctly navigate the LMS – likely since they had been forced online due to COVID the previous semester – but this was not the case. Many potential interviewees for experiment 5 who signed up for an interview were excluded from the study because they did not correctly enroll themselves in the appropriate group discussion area. Had they been given instruction on how to join an online group thread, the number of interviews that could have been obtained would have tripled, based on the number of interview sign-ups received.

This dissertation centred the learners' experience in every experiment to better understand how learners interpreted LADs, and what they did with this information. While not employing co-design strategies outright, these experiments foregrounded the opinions and experiences of learners, employing iterative design methods akin to design-based research (Wang & Hannafin, 2005). The learners who participated in these experiments were all from the same university, and largely, the same department. Their varied experiences illustrated the range of variation within the learner population, while

the homogeneity of the sample allowed us to describe the participant experience as a group. This sample population may or may not represent this student body, however similar results obtained from repeated sampling do speak to saturation, and the representational nature of our results. That we did not see new insights in experiment 5 supported the idea that our sample size was adequate.

It may be possible to contextualize the experiments according to the caliber of the university, especially with respect to the lack of critical thinking. The university is the highest ranked comprehensive research university in Canada (i.e. without a medical school), and places between 300-400 in international rankings. Admission requirements to the department include a GPA of between 82-85%; international students must have a GPA above 90%. Academically, the student body represents a profile that is similar to many universities, except the highly selective ones. Nevertheless, it would be interesting to investigate if similar results would be observed in universities with more competitive admission, or conversely, if successful LAD studies originating in highly regarded universities would be replicated in a context such as ours, or even at lower ranked community colleges.

11.8. Recommendations

Blended learners need help navigating online learning environments and scenarios. Had we had more input on the design of the learning activity, we would have provided clear instruction on how to participate in the discussion. We also would have provided instructions for how to conduct a good online discussion and examples of substantive discussion contributions. Changing the weighting of the assessment rubric to prioritize quality over quantity would better align it with a mastery goal orientation. Seeing that the duration of the short group discussion did not lend itself to the repeated exposures that experiments 3 and 4 suggested may improve learners' gist accuracy, we would extend the duration of the small group activity in future LAD studies with this type of learning activity if at all possible.

In the interviews the learners – even the ones who thought the LADs were interactive art – were able to deeply reflect on their learning. The key to that reflection

was guidance. Embedding student-facing dashboards in the learning activity similar to that described in the experiments of de Quincey et al. (2019), could serve a dual purpose, teaching the learner how to use the LADs to reflect on their learning, and providing concrete evidence of the tool's value toward the enactment of learning strategies.

Dollinger and Lodge (2018) conceptualize learning analytics as a service; they maintain that the value of LA is co-created through stakeholder interaction. This is the perspective that should be adopted going forward in the design of student-facing LADs. If as educational designers and researchers we ask ourselves how learners are being served by LADs, then we will be better able to evaluate their "value in use" (Vargo & Lusch, 2012), how learners use and experience the service LADs provide. In doing so, perhaps we will produce student-facing LADs that are better aligned with learners' capacities for critical thought, and the self-regulatory behaviors we hope to help them develop. Our findings contribute to the field by putting learners' perceptions and experiences at the forefront of the design process. These experiments shed light on how LADS influenced the learning of the participants— in both their disciplines and as learners being exposed to educational technologies. In the qualitative and quantitative exploration of learners' interactions with LADs during the process of learning and in the operationalization of gist as a means of evaluation, we contribute to the ongoing investigations of learners' sensemaking processes with learning analytics dashboards.

References

Aghababyan, A., Lewkow, N., & Baker, R. (2017). Exploring the asymmetry of metacognition. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference*, 115-119. http://doi.org/10.1145/3027385.3027388

Ahn, J., Gubbels, M., Yip, J., Bonsignore, E. & Clegg. T. (2013). Using social media and learning analytics to understand how children engage in scientific inquiry. In *Proceedings of the 12th International Conference on Interaction Design and Children*, 427–430. https://doiorg.ezproxy.library.ubc.ca/10.1145/2485760.2485805

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. A. Dunning, & J. I. Krueger (Eds.), *The Self in Social Judgment* (pp. 85–106).

Ally, M. (2008). Foundations of educational theory for online learning. In T. Anderson & F. Elloumi (Eds.), *The Theory and Practice of Online Learning* (1st ed., pp. 15–41). Edmonton AB.

American Library Association. (n.d.). Digital literacy. Retrieved January 26, 2021, from https://literacy.ala.org/digital-literacy/

Anaya, A. R., Luque, M., & Peinado, M. (2016). A visual recommender tool in a collaborative learning experience. *Expert Systems with Applications: An International Journal*, *45*(C), 248–259. http://doi.org/10.1016/j.eswa.2015.01.071

Anderson, L. W., Krathwohl, D., & Bloom, B. S. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman, 1–352. http://doi.org/10.2307/42926529

Andrews, M. (2008). Never the last word: Revisiting data. In Andrews, M., Squire, C., & Tamboukou, M. (Eds.), *Doing narrative research*. SAGE Publications, 87-101. https://www-doi-org.ezproxy.library.ubc.ca/10.4135/97808570

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In *Evolution of Teaching and Learning Paradigms in Intelligent Environment*. Springer Berlin Heidelberg, 183–221. https://doi.org/10.1007/978-3-540-71974-8_8

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 267-270. http://doi.org/10.1145/2330601.2330666

Arnold, K. E., Karcher, B., Wright, C. V., & McKay, J. (2017). Student empowerment, awareness, and self-regulation through a quantified-self student tool. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference*, 526–527. http://doi.org/10.1145/3027385.3029434

Axial Coding. (2008). Axial Coding. In *SAGE Encyclopedia of Qualitative Research Methods*. http://doi.org/10.4135/9781412963909

Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, *4*(2).

Bamford, A. (2003). *The visual literacy white paper* (pp. 1–8). Adobe Systems Pty, Australia.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*(2), 248–287. http://doi.org/10.1016/0749-5978(91)90022-l

Baneres, D., Rodriguez, M. E., & Serra, M. (2019). An early feedback prediction system for learners at-risk within a first-year higher education course. *IEEE Transactions on Learning Technologies*, *12*(2), 249–263. http://doi.org/10.1109/TLT.2019.2912167

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594. http://doi.org/10.1080/10447310802205776

Bateman, S., Mandryk, R. L., Gutwin, C., et al. (2010). Useful junk?: The effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2573-2582. http://doi.org/10.1145/1753326.1753716

Baxter, P., & Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. *The Qualitative Report*, *13*(4), 544–559.

Beheshitha, S. S., Gašević, D., & Hatala, M. (2015a). A process mining approach to linking the study of aptitude and event facets of self-regulated learning. In *Proceedings of the 5th International Learning Analytics & Knowledge Conference*, 265–269. http://doi.org/10.1145/2723576.2723628

Beheshitha, S. S., Hatala, M., Gašević, D., & Joksimović, S. (2016). The role of achievement goal orientations when studying effect of learning analytics visualizations. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, 54–63. http://doi.org/10.1145/2883851.2883904

Beheshtiha, S. (2015). Varying effects of learning analytics visualizations for students with different achievement goal orientations (Master's Thesis). Simon Fraser University, Burnaby, British Columbia, Canada.

Bembenutty, H. (2009). Test anxiety and academic delay of gratification. *College Student Journal*, *43*, 10–21.

Bennett, E., & Folley, S. (2019). Four design principles for learner dashboards that support student agency and empowerment. *Journal of Applied Research in Higher Education* (Vol. 12, pp. 15–26).

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Perception & Psychophysics*, *8*(5), 279–286. http://doi.org/10.3758/BF03212593

Black, P., & Wiliam, D. (1998, October). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 139–144, 146–148.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. http://doi.org/10.1007/s11092-008-9068-5

Bodily, R., & Verbert, K. (2017a). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*, *10*(4), 405–418. http://doi.org/10.1109/TLT.2017.2740172

Bodily, R., & Verbert, K. (2017b). Trends and issues in student-facing learning analytics reporting systems research. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference*, 309-318. http://doi.org/10.1145/3027385.3027403

Bodner, G. M., & Guay, R. B. (1997). The Purdue visualization of rotations test. *The Chemical Educator*, *2*(4), 1–17. http://doi.org/10.1007/s00897970138a

Bourgeault, I., Dingwall, R., & De Vries, R. (2020). Qualitative interviewing techniques and styles. In *The SAGE Handbook of Qualitative Methods in Health Research*. SAGE Publications, 1-20. http://doi.org/10.4135/9781446268247

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review*, *10*(1), 3–47. http://doi.org/10.1016/0273-2297(90)90003-M

Branoff, T. J. (2009). Spatial visualization measurement: A modification of the Purdue spatial visualization test—Visualization of Rotations. *Engineering Design Graphics Journal*.

Broos, T., Verbert, K., Langie, G., Van Soom, C., & De Laet, T. (2018). Multi-institutional positioning test feedback dashboard for aspiring students. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 51–55. http://doi.org/10.1145/3170358.3170419

Bull, S., Ginon, B., Boscolo, C., & Johnson, M. (2016). Introduction of learning visualisations and metacognitive support in a persuadable open learner model. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 30-39. http://doi.org/10.1145/2883851.2883853

Cacioppo, J. T., Petty, R. E., & Feinstein, J. A. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253. http://doi.org/10.1037/0033-2909.119.2.197

Canadian Consortium for Self-Regulated Learning. (n.d.) What is SRL? Retrieved January 14th, 2019 from http://srlcanada.ca/what-is-srl/

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. http://doi.org/10.1016/j.visres.2011.04.012

Carroll, J. B. (2009). Abilities in the domain of visual perception. In *Human Cognitive Abilities*. Cambridge University Press, 1-60. http://doi.org/10.1017/CBO9780511571312.009

Castro F., Vellido A., Nebot À., Mugica F. (2007). Applying data mining techniques to e-learning problems. In: Jain L.C., Tedman R.A., Tedman D.K. (eds.) *Evolution of Teaching and Learning Paradigms in Intelligent Environment. Studies in Computational Intelligence*. Springer, Berlin, Heidelberg. https://doi-org.ezproxy.library.ubc.ca/10.1007/978-3-540-71974-8_8

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46, 112– 130. doi:10.3758/s13428-013-0365-7

Chatterjee, A., & Vartanian, O. (2014). Neuroaesthetics. *Trends in Cognitive Sciences*, *18*(7), 370–375. http://doi.org/10.1016/j.tics.2014.03.003

Check, J., & Schutt, R. K. (2018). Single-subject design. In *Research Methods in Education*. SAGE Publications, 1-30. http://doi.org/10.4135/9781544307725

Chenail, R. J. (2009). Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research. *The Qualitative Report*, *13*(1), 14–21.

Chickering, A. W., & Ehrmann, S. C. (1996). Implementing the seven principles: Technology as lever. *American Association for Higher Education Bulletin*, *49*(2), 3–6.

Claessen, M. H. G., van der Ham, I. J. M., & van Zandvoort, M. J. E. (2014). Computerization of the standard Corsi block-tapping task affects its underlying cognitive concepts: A pilot study. *Adult Applied Neuropsychology*, *22*(3), 180–188.

Clarke, J., & Mack, A. (2014). Iconic memory for the gist of natural scenes. *Consciousness and Cognition*, *30*(C), 256–265. http://doi.org/10.1016/j.concog.2014.09.015

Cleary, T. J., & Zimmerman, B. J. (2004). Self-regulation empowerment program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools*, *41*(5), 537–550. http://doi.org/10.1002/pits.10177

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*(4716), 828–833. http://doi.org/10.1126/science.229.4716.828

Cleveland, W. S., & McGill, R. (2012). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*(387), 531–554. http://doi.org/10.1080/01621459.1984.10478080

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, 1, 120–131.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). L. Erlbaum Associates.

Cohen, V. (1985). A reexamination of feedback in computer-based instruction: Implications for instructional design. *Educational Technology*, *25*, 33–37.

Cokely, E. T., Galesic, M., Schulz, E., & Ghazal, S. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*, 25–47.

Cokely, E. T., Ghazal, S., & Garcia-Retamero, R. (2014). Measuring numeracy. In B. L. Anderson & J. Schulkin (Eds.), *Numerical Reasoning in Judgments and Decision Making about Health*. Cambridge: Cambridge University Press, 11-38. http://doi.org/10.1017/CBO9781139644358.002

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21.

Corbin, J., & Strauss, A. (2020). Techniques and procedures for developing grounded theory. In *Basics of Qualitative Research* (3rd ed.). Thousand Oaks, California: SAGE Publications, 195–228. http://doi.org/10.4135/9781452230153

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2142–2151. http://doi.org/10.1109/TVCG.2014.2346298

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, 8, e57410. doi:10.1371/journal.pone.0057410

Culén, A. L. (2014). Visual immediacy for sense-making in HCI. In *Proceedings of the IADIS International Conference on Interfaces and Human-Computer Interaction*, 265–270.

Dabbagh, N., & Kitsantas, A. (2004). Supporting self-regulation in student-centered web-based learning environments. *International Journal on E-Learning*, *3*, 40–47.

De Angeli, A., Sutcliffe, A., & Hartmann, J. (2006). Interaction, usability and aesthetics. In *Proceedings of the 6th conference on Designing Interactive systems (DIS '06)*, 271-280. http://doi.org/10.1145/1142405.1142446

de Quincey, E., Briggs, C., Kyriacou, T., & Waller, R. (2019). Student centred design of a learning analytics system. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 353–362. http://doi.org/10.1145/3303772.3303793

Deller, M., Ebert, A., Bender, M., Agne, S., & Barthel, H. (2007). Preattentive visualization of information relevance. In *Proceedings of the International Workshop on Human-Centered Multimedia*, 47–56. http://doi.org/10.1145/1290128.1290137

Demiralp, C., Scheidegger, C. E., Kindlmann, G. L., Laidlaw, D. H., & Heer, J. (2014). Visual embedding: A model for visualization. *Computer Graphics and Applications, IEEE*, *34*(1), 10–15. http://doi.org/10.1109/MCG.2014.18

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification." In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 9–15. http://doi.org/10.1145/2181037.2181040

Devaney, L. (2010, October 12). Purdue's student achievement technology goes national. *eCampus News*. Retrieved from www.ecampusnews.com/-technologies/purdues-student-achievement-technology-goes-national/

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical Education*, 40(4), 314–321. http://doi.org/10.1111/j.1365-2929.2006.02418.x

Dick, A. O. (1974). Iconic memory and its relation to perceptual processing and other memory mechanisms. *Perception & Psychophysics*, *16*(3), 575–596. http://doi.org/10.3758/BF03198590

Dollinger, M., & Lodge, J. M. (2018). Co-creation strategies for learning analytics (Vol. 15, pp. 97–101). In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 97–101. http://doi.org/10.1145/3170358.3170372

Domínguez, A., Saenz-de-Navarrete, J., de-Marcos, L., Fernández-Sanz, L., Pagés, C., & Martínez-Herráiz, J.-J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education*, 63, 380–392. http://doi.org/https://doi.org/10.1016/j.compedu.2012.12.020

Dringus, L. (2012). Learning analytics considered harmful. *Journal of Asynchronous Learning Networks*, *16*(3), 87–100.

Duncan, T. G., & Mckeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educational Psychologist*, *40*(2), 117–128. http://doi.org/10.1207/s15326985ep4002_6

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*, 83–87.

Duval, E. (2011). Attention please! Learning analytics for visualization and recommendation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, 9–17. http://doi.org/10.1145/2090116.2090118

Eddy, S. L., Brownell, S. E., Thummaphan, P., Lan, M.-C., & Wenderoth, M. P. (2015). Caution, student experience may vary: Social identities impact a student's experience in peer discussions. *CBE Life Sciences Education*, *14*(4), ar45. http://doi.org/10.1187/cbe.15-05-0108

Edwards, A., Thomas, R., Williams, R., Ellner, A. L., Brown, P., & Elwyn, G. (2006). Presenting risk information to people with diabetes: Evaluating effects and preferences for different formats by a web-based randomised controlled trial. *Patient Education and Counseling*, *63*(3), 336–349. http://doi.org/10.1016/j.pec.2005.12.016

Edwards, R. K., Kellner, K. R., Sistrom, C. L., & Magyari, E. J. (2003). Medical student self-assessment of performance on an obstetrics and gynecology clerkship. *American Journal of Obstetrics and Gynecology*, *188*(4), 1078–1082. http://doi.org/10.1067/mob.2003.249

Eger, N., Ball, L. J., Stevens, R., & Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through eye-movement replay. *Electronic Workshops in Computing (Online)*.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121. http://doi.org/10.1016/j.obhdp.2007.05.002

Elliot, A.J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*. 34, 3, 169-189. http://doi.org/10.1207/s15326985ep3403_3

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*. 72, 218-232. https://doi.org/10.1037/0022-3514.72.1.218

Elliot, A.J. and Dweck, C.S. (2005). A conceptual history of the achievement goal construct. *Handbook of Competence and Motivation*. 16, 52–72.

Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*, *80*(3), 501–519. http://doi.org/10.1037/0022-3514.80.3.501

Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3 × 2 achievement goal model. *Journal of Educational Psychology*, *103*(3), 632–648. http://doi.org/10.1037/a0023952

ElSayed, A. A., Caeiro-Rodríguez, M., MikicFonte, F. A., & Llamas-Nistal, M. (2019). Research in learning analytics and educational data mining to measure self-regulated learning: A systematic review. In *Proceedings of the World Conference on Mobile and Contextual Learning*, 46–53.

Elzer, S., Green, N., Carberry, S., & Hoffman, J. (2006). A model of perceptual task effort for bar charts and its role in recognizing intention. *User Modeling and User-Adapted Interaction*, *16*(1), 1–30. http://doi.org/10.1007/s11257-006-9002-9

Emily R Lai, M. V. (2012). Assessing 21st century skills: Integrating research findings. *National Council on Measurement in Education*, 1–67.

Epstein, R. (2005). The cortical basis of visual scene processing. *Visual Cognition*, *12*(6), 954–978. http://doi.org/10.1080/13506280444000607

Esterman, M. (2000). Preattentive and attentive visual search in individuals with hemispatial neglect. *Neuropsychology*, *14*(4), 599–611. http://doi.org/10.1037/0894-4105.14.4.599

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672–680. http://doi.org/10.1177/0272989X07304449

Falk, J. H., & Balling, J. D. (2010). Evolutionary influence on human landscape preference. *Environment and Behavior*, *42*(4), 479–493. http://doi.org/10.1177/0013916509341244

Fan, J. E., Turk-Browne, N. B., & Taylor, J. A. (2016). Error-driven learning in statistical summary perception. *Journal of Experimental Psychology: Human Perception and Performance*, 42(2), 266–280. https://doi.org/10.1037/xhp0000132

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304–317. http://doi.org/10.1504/IJTEL.2012.051816

Ferguson, R., & Clow, D. (2017). Where is the evidence? In *Proceedings of the 7th International Learning Analytics & Knowledge Conference*, 56-65. http://doi.org/10.1145/3027385.3027396

Few, S. (2007). Save the Pies for Dessert *Perceptual Edge*. 1–14. http://www.perceptualedge.com/articles/08-21-07.pdf

Follmer, D. J., Sperling, R. A., & Suen, H. K. (2017). The role of Mturk in education research: Advantages, issues, and future directions. *Educational Researcher*, 46(6), 329–334. http://doi.org/10.3102/0013189X17725519

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2-3), 285–307. http://doi.org/10.1080/01638539809545029

Fortin, J. M., Hirota, L. K., Bond, B. E., O'Connor, A. M., & Col, N. F. (2001). Identifying patient preferences for communicating risk estimates: A descriptive pilot study. *BMC Medical Informatics and Decision Making*, *1*(1), 2. http://doi.org/10.1186/1472-6947-1-2

Forum, W. E. (2016). The future of jobs: Employment, skills and workforce strategy for the fourth industrial revolution (Executive Summary), *World Economic Forum, 1-12*.

Foundation for Young Australians. (2017). The new work order: Ensuring young Australians have skills and experience for the jobs of the future, not the past. 1–50.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*. http://doi.org/10.2307/4134953

Frith, H., & Harcourt, D. (2007). Using photographs to capture women's experiences of chemotherapy: Reflecting on the method. *Qualitative Health Research*, *17*(10), 1340–1350. http://doi.org/10.1177/1049732307308949

Fritz, J. (2011). Classroom walls that talk: Using online course activity data of successful students to raise self-awareness of underperforming peers. *The Internet and Higher Education*, *14*(2), 89–97. http://doi.org/10.1016/j.iheduc.2010.07.007

Fukukura, J., Ferguson, M. J., & Fujita, K. (2013). Psychological distance can improve decision making under information overload via gist memory. *Journal of Experimental Psychology. General*, *142*(3), 658–665. http://doi.org/10.1037/a0030730

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, *15*(1), 7–23. http://doi.org/10.1080/08923640109527071

Gauvrit, N., Soler-Toscano, F., & Zenil, H. (2014). Natural scene statistics mediate the perception of image complexity. *Visual Cognition*, *22*(8), 1084–1091. http://doi.org/10.1080/13506285.2014.950365

Ghazal, S., & Cokely, E. T. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making*, *9*, 15–34.

Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 18, 153-162. http://doi.org/10.1145/3027385.3027436

Gillan, D. J., & Sorensen, D. (2009). Minimalism and the Syntax of Graphs: II. Effects of Graph Backgrounds on Visual Search. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 53(17), 1096–1100. http://doi.org/10.1518/107118109X12524443344998

Given, L. M. (2008). The SAGE encyclopedia of qualitative research methods. Thousand Oaks, CA: SAGE Publications, 10.4135/9781412963909

Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, *47*(2), 183–210. http://doi.org/10.1080/03057267.2011.605307

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, *77*(3), 334–372. http://doi.org/10.3102/003465430303953

Grice, P. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Studies in the way of words*, 3, 41–58.

Grimes, P. W. (2010). The overconfident principles of economics student: An examination of a metacognitive skill. *The Journal of Economic Education*, *33*(1), 15–30. http://doi.org/10.1080/00220480209596121

Guay, R. B. (1976). Purdue spatial visualization test-visualization of rotations. West Lafayette, IN: Purdue Research Foundation.

Hagerhall, C. M., Purcell, T., & Taylor, R. (2004). Fractal dimension of landscape silhouette outlines as a predictor of landscape preference. *Journal of Environmental Psychology*, *24*(2), 247–255. http://doi.org/10.1016/j.jenvp.2003.12.004

Hagh-Shenas, H., Kim, S., Interrante, V., & Healey, C. (2007). Weaving versus blending: A quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1270–1277. http://doi.org/10.1109/TVCG.2007.70623

Haigh M. (2016). Has the standard cognitive reflection test become a victim of its own success? *Advances in cognitive psychology*, 12(3), 145–149. https://doi.org/10.5709/acp-0193-5

Hair. F., J., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. (2014). Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research. *European Business Review*, 26(2), 106-121. https://doi.org/10.1108/EBR-10-2013-0128

Hamborg, K.-C., Hülsmann, J., & Kaspar, K. (2014). The interplay between usability and aesthetics: more evidence for the "what is usable is beautiful" notion. *Advances in Human-Computer Interaction*, N1–13. https://doi.org/10.1155/2014/946239

Harper, D. (2010). Talking about pictures: A case for photo elicitation. *Visual Studies, 17*(1), 13–26. http://doi.org/10.1080/14725860220137345

Hartley, K. (2001). Educational research in the internet age: Examining the role of individual characteristics. *Educational Researcher, 30*, 22–26.

Hauser, D. J., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. http://doi.org/10.3758/s13428-015-0578-z

Hawes, Z., LeFevr, J.-A., Xu, C., & Bruce, C. D. (2015). Mental rotation with tangible three-dimensional objects: A new measure sensitive to developmental differences in 4- to 8-year-old children. *Mind, Brain, and Education, 9*, 10–19.

Healey, C. G., & Enns, J. T. (2012). Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, *18*(7), 1170–1188. http://doi.org/10.1109/TVCG.2011.127

Healey, C. G., Booth, K. S., & Enns, J. T. (1996). High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction, 3*(2), 107–135. https://doi.org/10.1145/230562.230563

Heer, J., Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception (p. 203). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–212. http://doi.org/10.1145/1753326.1753357

Heil, M., Bajrić, J., Rösler, F., & Hennighausen, E. (1996). Event-related potentials during mental rotation: Disentangling the contributions of character classification and image transformation. *Journal of Psychophysiology*, *10*(4), 326–335.

Hillaire, G., Rappolt-Schlichtmann, G., & Ducharme, K. (2016). Prototyping visual learning analytics guided by an educational theory informed goal. *Journal of Learning Analytics*, *3*(3), 115–142. http://doi.org/10.18608/jla.2016.33.7

Hinkle, D. N. (1966). The change of personal constructs from the viewpoint of a theory of construct implications. *Dissertation Abstracts*, 26(11), 6837.

Hirumi, A. (2002). A framework for analyzing, designing, and sequencing planned elearning interactions. *Quarterly Review of Distance Education*, 3(2), 141.

Ho, T. H. Y., Nguyen, T. T., & Song, I. (2016). Visualizing Learning Activities in Social Network. *Intelligent Information and Database Systems*, 97–105. https://doi.org/10.1007/978-3-662-49381-6_10

Hollands, J. G., & Spence, I. (1998). Judging proportion with graphs: the summation model. *Applied Cognitive Psychology*, 12(2), 173–190. https://doi.org/10.1002/(SICI)1099-0720(199804)12:23.0.CO;2-K

Holstein, J. A., & Gubrium, J. F. (1995). The active interview. SAGE Publications. *Qualitative research methods*. https://doi-org.ezproxy.library.ubc.ca/10.4135/9781412986120

Houtkoop-Steenstra, H. (2000). Interaction and the standardized survey interview: The living questionnaire. Cambridge University Press. Retrieved from http://ebookcentral.proquest.com/lib/sfu-ebooks/detail.action?docID=147324

Howley, P. (2011). Landscape aesthetics: Assessing the general publics' preferences towards rural landscapes. *Ecological Economics*, *72*(C), 161–169. http://doi.org/10.1016/j.ecolecon.2011.09.026

Hu, X., Yang, C., Qiao, C., Lu, X., & Chu, S. K. W. (2017). New features in Wikiglass, a learning analytic tool for visualizing collaborative work on wikis. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference*, 616-617. http://doi.org/10.1145/3027385.3029489

Hughes, J. (2012). SAGE internet research methods. SAGE Publications. https://www-doi-org.proxy.lib.sfu.ca/10.4135/9781446268513

Hullman, J., Adar, E., & Shah, P. (2011). Benefitting InfoVis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2213–2222. http://doi.org/10.1109/TVCG.2011.175

Paul Black, & Dylan Wiliam. (1998). Inside the Black Box: Raising Standards through Classroom Assessment. *Phi Delta Kappan*, 80(2), 139–148.

Isenberg, P., Zuk, T., Collins, C., & Carpendale, S. (2008). Grounded evaluation of information visualizations. In *Proceedings of the 2008 conference on BEyond time and errors novel evaLuation methods for Information Visualization*, 1-8. http://doi.org/10.1145/1377966.1377974

Ishizu, T., & Zeki, S. (2013). The brain's specialized systems for aesthetic and perceptual judgment. *European Journal of Neuroscience*, *37*(9), 1413–1420. http://doi.org/10.1111/ejn.12135

Jaakonmäki, R., Brocke, vom, J., Dietze, S., Drachsler, H., Fortenbacher, A., Helbig, R., et al. (2020). Learning analytics in a primary school: Lea's box recipe. In *Learning Analytics Cookbook*. Springer International Publishing, 45–50. http://doi.org/10.1007/978-3-030-43377-2_5

Jacobsen, T., Schubotz, R. I., Höfel, L., & Cramon, D. Y. V. (2006). Brain correlates of aesthetic judgment of beauty. *NeuroImage* (Orlando, Fla.), 29(1), 276–285. https://doi.org/10.1016/j.neuroimage.2005.07.010

Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). Awareness is not enough. Pitfalls of learning analytics dashboards in the educational practice

Jivet, Ioana, Scheffel, Maren, chsler, Henik, & Specht, Marcus. (2017). Awareness is not enough. Pitfalls of learning analytics dashboards in the educational practice. *Lecture Notes in Computer Science*, 82–96. https://doi.org/10.1007/978-3-319-66610-5_7

Jivet, I., Scheffel, M., Specht, M., & Drachsler, H. (2018). License to evaluate. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 31–40. https://doi.org/10.1145/3170358.3170421

Josephs, E. L., Draschkow, D., Wolfe, J. M., & Võ, M. L. H. (2016). Gist in time: Scene semantics and structure enhance recall of searched objects. *Acta Psychologica*, *169*(C), 100–108. http://doi.org/10.1016/j.actpsy.2016.05.013

Kahneman, D. (2011). *Thinking, Fast and Slow*. 1st. New York: Farrar.

Kalat, J. W. (2021). Introduction to Psychology. Pacific Grove, CA: Wadsworth-Thompson Learning, 8, 1–643. http://doi.org/049510289X

Karabeg, A., & Akkøk, M. N. (2004). Towards Language for Talking about Visual and Spatial Reasoning.

Kemps, E. (2001). Complexity effects in visuo-spatial working memory: Implications for the role of long-term memory. *Memory (Hove)*, *9*(1), 13–27.

Khosravi, H., Gyamfi, G., Hanna, B. E., & Lodge, J. (2020). Fostering and supporting empirical research on evaluative judgement via a crowdsourced adaptive learning system. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 83–88. https://doi.org/10.1145/3375462.3375532

Kirk, U., Skov, M., Christensen, M. S., & Nygaard, N. (2009). Brain correlates of aesthetic expertise: A parametric fMRI study. *Brain and Cognition*, *69*(2), 306–315. http://doi.org/10.1016/j.bandc.2008.08.004

Knight, S., Shibani, A., Abel, S., Gibson, A., & Ryan, P. (2020). AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research, 12*(vol. 12 issue 1), 141–186. http://doi.org/10.17239/jowr-2020.12.01.06

Koenig, J. A. (2011). Assessing 21st century skills: Summary of a workshop. *National Research Council, Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education*, 1-24.

Kosara, R., & Ziemkiewicz, C. (2010). Do Mechanical Turks dream of square pie charts? *The 3rd BELIV'10 Workshop*, 63–70. http://doi.org/10.1145/2110192.2110202

Kristjánsson, Á. (2015). Reconsidering visual search. *I-Perception*, *6*(6), 2041669515614670–2041669515614670. http://doi.org/10.1177/2041669515614670

Kristjánsson, Á., & Egeth, H. (2020). How feature integration theory integrated cognitive psychology, neurophysiology, and psychophysics. *Attention, Perception, & Psychophysics*, *82*(1), 1–17. http://doi.org/10.3758/s13414-019-01803-7

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*(2), 180–188. http://doi.org/10.1037//0022-3514.82.2.180

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. http://doi.org/10.1037/0022-3514.77.6.1121

Kurosu, M., & Kashimura, K. (1995). Apparent usability vs. inherent usability. *Conference Companion on Human Factors in Computing Systems*, 292–293. https://doi.org/10.1145/223355.223680

Kurzhals, K., Fisher, B., Burch, M., & Weiskopf, D. (2014). Evaluating visual analytics with eye tracking. In *Proceedings of the Fifth Workshop on Beyond Time and Errors*, 61–69. https://doi.org/10.1145/2669557.2669560

Kvale, S. (2020). Enhancing interview quality. In *Doing interviews* (p. 136). SAGE Publications. https://doi.org/10.4135/9781849208963.n12

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*(1), 65–100. http://doi.org/10.1016/S0364-0213(87)80026-5

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work* (pp. 63–76). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-89350-9_6

Lee, J., Luchini, K., Michael, B., Norris, C., & Soloway, E. (2004). More than just fun and games: Assessing the value of educational video games in the classroom. In *Proceedings of the CHI '04 Extended Abstracts on Human Factors in Computing Systems*, 1375–1378. http://doi.org/10.1145/985921.986068

Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., & Henry, N. (2006). Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors*, 1–5. https://doi.org/10.1145/1168149.1168168

Lee, S., Kim, S.-H., & Kwon, B. C. (2016). VLAT: Development of a visualization literacy assessment test. *Visualization and Computer Graphics, IEEE Transactions on*, *23*(1), 551–560. http://doi.org/10.1109/TVCG.2016.2598920

Lewis-Beck, M. S., Bryman, A., & Liao, T. F. F. (2003). SAGE Encyclopedia of Social Science Research Methods. Thousand Oaks: SAGE Publications.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*(4), 361–381. http://doi.org/10.1002/bdm.752

Liiv, I. (2010). Towards information-theoretic visualization evaluation measure: A practical example for Bertin's matrices. In *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization*, 24-28. http://doi.org/10.1145/2110192.2110196

Lim, L., Dawson, S., Joksimović, S., & Gašević, D. (2019). Exploring students' sensemaking of learning analytics dashboards. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 250–259. https://doi.org/10.1145/3303772.3303804

Lim, Y.-K., Lim, Y., Stolterman, E., Jung, H., Stolterman, E., Jung, H., et al. (2007). Interaction gestalt and the design of aesthetic interactions. In *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces*, 239–254. http://doi.org/10.1145/1314161.1314183

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavioral Research*, 49(2), 433–442. http://doi.org/10.3758/s13428-016-0727-z

Liu, S., Cui, W., Wu, Y., & Liu, M. (2014). A survey on information visualization: Recent advances and challenges. *The Visual Computer: International Journal of Computer Graphics*, *30*(12), 1373–1393. http://doi.org/10.1007/s00371-013-0892-3

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, *18*(4), 513–536. http://doi.org/10.1080/13506280902937606

Macé, M. J. M., Thorpe, S. J., & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: How robust at very low contrasts? *European Journal of Neuroscience*, *21*(7), 2007–2018. http://doi.org/10.1111/j.1460-9568.2005.04029.x

Mack, A., & Clarke, J. (2012). Gist perception requires attention. *Visual Cognition*, *20*(3), 300–327. http://doi.org/10.1080/13506285.2012.666578

Mahatody, T., Sagar, M., & Kolski, C. (2007). Cognitive Walkthrough for HCI evaluation: Basic concepts, evolutions and variants, research issues. *European Annual Conference on Human-Decision Making and Manual Control*.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Marshall, T., Mohammed, M. A., & Rouse, A. (2004). A randomized controlled trial of league tables and control charts as aids to health service decision-making. *International Journal for Quality in Health Care*, 16(4), 309–315. https://doi.org/10.1093/intqhc/mzh054

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 1–23. http://doi.org/10.3758/s13428-011-0124-6

Matcha, W., Ahmad Uzir, N., Gasevic, D., & Pardo, A. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, 1382(c), 1–1. https://doi.org/10.1109/tlt.2019.2916802

McNaughton, C. D., Cavanaugh, K. L., Kripalani, S., Rothman, R. L., & Wallston, K. A. (2015). Validation of a short, 3-item version of the subjective numeracy scale. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *35*(8), 932–936. http://doi.org/10.1177/0272989X15581800

McNaughton, C., Wallston, K. A., Rothman, R. L., Marcovitz, D. E., & Storrow, A. B. (2011). Short, subjective measures of numeracy and general health literacy in an adult emergency department. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *18*(11), 1148–1155. http://doi.org/10.1111/j.1553-2712.2011.01210.x

Mealey, L., & Theis, P. (1995). The relationship between mood and preferences among natural landscapes: An evolutionary perspective. *Ethology and Sociobiology*, 16(3), 247–256. https://doi.org/10.1016/0162-3095(95)00035-J\

Mejia, C., Florian, B., Vatrapu, R., Bull, S., Gomez, S., & Fabregat, R. (2017). A novel web-based approach for visualization and inspection of reading difficulties on university students. *IEEE Transactions on Learning Technologies*, *10*(1), 53–67. http://doi.org/10.1109/TLT.2016.2626292

Mendiburo, M., Sulcer, B., & Hasselbring, T. (2014). Interaction design for improved analytics. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, 78–82. https://doi.org/10.1145/2567574.2567628

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, 15(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Meyer, K. A. (2014). Student engagement in online learning: What works and why. *ASHE Higher Education Report*, 40(6), 1–114. http://doi.org/10.1002/aehe.20018

Mezirow, J. (1981). A critical theory of adult learning and education. *Adult Education Quarterly*, 32(1), 3–24. https://doi.org/10.1177/074171368103200101

Mezirow, J. (2000). Learning as transformation: Critical perspectives on a theory in progress. *Jack Mezirow and Associates* (pp. 1–371). San Francisco: Jossey-Bass.

Mezirow, J. (1991). Transformative dimensions of adult learning. San Francisco: Jossey-Bass.

Miller, G. A. (1995). The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information. *Psychological Review*, *101*, 343–352.

Miniukovich, A., & De Angeli, A. (2015). Computation of Interface Aesthetics. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1163–1172. https://doi.org/10.1145/2702123.2702575

Mishler, E. G. (1986). The analysis of interview-narratives. In *Narrative psychology: The storied nature of human conduct.* (pp. 233–255). Westport, CT, US: Praeger Publishers/Greenwood Publishing Group.

Molenaar, I., Horvers, A., Dijkstra, R., & Baker, R. (2019). Designing dashboards to support learners' self-regulated learning. In *Companion Proceedings 9th International Conference on Learning Analytics & Knowledge (LAK19*), 1-12.

Montebello, M., Pinheiro, P., Cope, B., Kalantzis, M., Haniya, S., Tzirides, A. O., et al. (2018). Enriching Online Education through Differentiated Learning. Presented at the *Fourth International Conference on Higher Education Advance*s, Valencia: Universitat Politècnica València, 1-8. http://doi.org/10.4995/HEAD18.2018.8019

Mouri, K., Ogata, H., & Uosaki, N. (2017). Learning analytics in a seamless learning environment. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 1, 348-357. http://doi.org/10.1145/3027385.3027408

Noudoost, B., Chang, M. H., Steinmetz, N. A., & Moore, T. (2010). Top-down control of visual attention. *Current Opinion in Neurobiology*, *20*(2), 183–190. http://doi.org/10.1016/j.conb.2010.02.003

Nowakowska, A., Clarke, A., & Hunt, A. (2017). Human visual search behaviour is far from ideal. In *Proceedings of the Royal Society. B, Biological Sciences*, 284(1849), 20162767. https://doi.org/10.1098/rspb.2016.2767

Oakley, A. (2015). Interviewing Women Again: Power, Time and the Gift. *Sociology*, *50*(1), 195–213. http://doi.org/10.1177/0038038515580253

O'Connor, H. & Madge, C. (2017). Online interviewing. In *SAGE Handbook of Online Research Methods*. SAGE Publications, 416-434. https://www-doi-org.proxy.lib.sfu.ca/10.4135/9781473957992

O'Connor, H., Madge, C., Shaw, R. & Wellens, J. (2008). Internet-based interviewing. In *SAGE Handbook of Online Research Methods*. SAGE Publications, 271-289. https://www-doi-org.proxy.lib.sfu.ca/10.4135/9780857020055

OECD. (2018). Connected Minds: Technology and Today's Learners. *Educational Research and Innovation.* OECD Publishing, 1-20.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36. http://doi.org/10.1016/S0079-6123(06)55002-2

Orians, G. H., & Heerwagen, J. H. (1992). Evolved responses to landscapes. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*, xii, 555–579.

Owsley, C. (2013). Visual processing speed. *Vision Research*, *90*(C), 52–56. http://doi.org/10.1016/j.visres.2012.11.014

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science: A Journal of the American Psychological Society*, *23*(3), 184–188. https://doi.org/10.1177/0963721414531598

Paris, S. G., & Winograd, P. (2001). The role of self-regulated learning in contextual teaching: Principles and practices for teacher preparation. *European Journal of Psychology of Education*, *24*(1), 1–22.

Rösler, F., Heil, M., Bajric, J., Pauls, A., & Hennighausen, E. (1995.) Patterns of cerebral activation while mental images are rotated and changed in size. *Psychophysiology*, 32(2), 135–149. https://doi.org/10.1111/j.1469-8986.1995.tb03305.x

Pätsch, G., Mandl, T., & Womser-Hacker, C. (2014). Using sensor graphs to stimulate recall in retrospective think-aloud protocols. In *Proceedings of the 5th Information Interaction in Context Symposium*, 303–307. https://doi.org/10.1145/2637002.2637048

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2015). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, 48(1), 341–348. http://doi.org/10.3758/s13428-015-0576-1

Perels, F., Dignath, C., & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in regular math classes. *European Journal of Psychology of Education*, 24(1), 17–31. http://doi.org/10.1007/BF03173472

Perels, F., Gürtler, T., & Schmitz, B. (2005). Training of self-regulatory and problem-solving competence. *Learning and Instruction*, *15*(2), 123–139. http://doi.org/10.1016/j.learninstruc.2005.04.010

Peters, E. (2012). Beyond comprehension the role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*(1), 31–35. http://doi.org/10.1177/0963721411429960

Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40. http://doi.org/10.1037//0022-0663.82.1.33

Pomerantz, J. R., & Portillo, M. C. (2018). Emergent features, gestalts, and feature integration theory. In *From Perception to Consciousness*. Oxford University Press, 187–192. http://doi.org/10.1093/acprof:osobl/9780199734337.003.0016

Powell, K., & Kalina, C.J. (2009). Cognitive and Social Constructivism: Developing Tools for an Effective Classroom. *Education (Chula Vista)*, 130(2), 241-250.

Potter, J. & Hepburn, A. (2012). Eight challenges for interview researchers. In *SAGE Handbook Of Interview Research: The Complexity Of The Craft*, SAGE Publications, 555-570. https://www-doi-org.ezproxy.library.ubc.ca/10.4135/9781452218403

Quispel, A., & Maes, A. (2014). Would you prefer pie or cupcakes? Preferences for data visualization designs of professionals and laypeople in graphic design. *Journal of Visual Languages & Computing*, *25*(2), 107–116. http://doi.org/10.1016/j.jvlc.2013.11.007

Rasinski, K. A., Lee, L., & Krishnamurty, P. (2015). Question order effects. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.* American Psychological Association, 229-248. http://doi.org/10.1037/13619-014

Rensink, R. A. (2000). The Dynamic Representation of Scenes. *Visual Cognition*, *7*, 17–42.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*(1), 1–75. http://doi.org/10.1016/1041-6080(95)90031-4

Reyna, V. F., & Kiernan, B. (1994). Development of gist versus verbatim memory in sentence recognition: Effects of lexical familiarity, semantic content, encoding instructions, and retention interval. *Developmental Psychology*, *30*(2), 178–191. http://doi.org/10.1037/0012-1649.30.2.178

Richter, G., Raban, D. R., & Rafaeli, S. (2015). Studying gamification: The effect of rewards and incentives on motivation. In T. Reiners & L. C. Wood (Eds.), *Gamification in Education and Business*. Cham: Springer International Publishing, 21-46. http://doi.org/10.1007/978-3-319-10208-5_2

Robbins, N. B. (2004). *Creating more effective graphs*. Wiley.

Rodríguez-Triana, M. J., Martínez-Monés, A., & Villagrá-Sobrino, S. (2015). Applying Learning Analytics to a Primary School Classroom: Benefits and Barriers, 1–5.

Rösler, F., Heil, M., Bajrić, J., Pauls, A. C., & Hennighausen, E. (1995). Patterns of cerebral activation while mental images are rotated and changed in size. *Psychophysiology*, 32(2), 135–149. http://doi.org/10.1111/j.1469-8986.1995.tb03305.x

Rousselet, G. A., Macé, M. J. M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision*, *3*(6), 5–5. http://doi.org/10.1167/3.6.5

Ruiz, S., Charleer, S., Urretavizcaya, M., Klerkx, J., Fernández-Castro, I., & Duval, E. (2016). Supporting learning by considering emotions. In *Proceedings of the Sixth International Learning Analytics & Knowledge Conference*, 254–263. http://doi.org/10.1145/2883851.2883888

Ryan, T. A., & Schwartz, C. B. (1956). Speed of perception as a function of mode of representation. *The American Journal of Psychology*, *69*(1), 60. http://doi.org/10.2307/1418115

Sadler, D. R. (1989). Formative assessment and the design of instructional systems, *Instructional Science*, *18*(2), 119–144. http://doi.org/10.1007/BF00117714

Sampanes, A. C., Tseng, P., & Bridgeman, B. (2008). The role of gist in scene recognition. *Vision Research*, *48*(21), 2275–2283. http://doi.org/10.1016/j.visres.2008.07.011

Scheffel, M. (2017). The Evaluation Framework for Learning Analytics.

Scheffel, M., Drachsler, H., Kreijns, K., de Kraker, J., & Specht, M. (2017a). Widget, widget as you lead, I am performing well indeed! In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 17, 289–298. http://doi.org/10.1145/3027385.3027428

Scheffel, M., Drachsler, H., Toisoul, C., Ternier, S., & Specht, M. (2017b). The proof of the pudding: Examining validity and reliability of the evaluation framework for learning analytics. In E. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data Driven Approaches in Digital Education*, 10474, 1–15. Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-66610-5

Schön, D. A. (2016). *The reflective practitioner: How professionals think in action.* Routledge.

Schunk, D. H., & Zimmerman, B. J. (1998). *Self-regulated learning: from teaching to self-reflective practice*. New York: Guilford Press.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review*, *1*(3), 357–375. http://doi.org/10.3758/BF03213977

Schwendimann, B. A., Rodríguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., et al. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, *10*(1), 30–41. http://doi.org/10.1109/TLT.2016.2599522

Sclater, N., Peasgood, A., & Mullan, J. (2016). Case Study A: Traffic lights and interventions: Signals at Purdue University. *JISC*, 1–7.

Sedlmair, M., Meyer, M., & Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics, 18*(12), 2431–2440. http://doi.org/10.1109/TVCG.2012.213

Selwyn, N. (2009). The digital native – myth and reality. In *Proceedings of the ASLIB, 61*(4), 364–379. http://doi.org/10.1108/00012530910973776

Selwyn, N., & Gašević, D. (2020). The datafication of higher education: Discussing the promises and problems. *Teaching in Higher Education*, *25*(4), 527–540. http://doi.org/10.1080/13562517.2019.1689388

Serra, M. J., & DeMarree, K. G. (2016). Unskilled and unaware in the classroom: College students' desired grades predict their biased grade predictions. *Memory & Cognition, 44*(7), 1127–1137. http://doi.org/10.3758/s13421-016-0624-9

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*, *3*(3), 560–578. http://doi.org/10.1111/j.1756-8765.2009.01066.x

Shah, P., Freedman, E. G., & Vekiri, I. (2009). The comprehension of quantitative information in graphical displays. In P. Shah & A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking*. Cambridge University Press, 426-476. http://doi.org/10.1017/CBO9780511610448.012

Shneiderman, B. (2002). Understanding human reactivites and relationships: An excerpt from Leonardo's laptop. *Interactions*, *9*(5). http://doi.org/10.1145/566981.566982

Shneiderman, B., & Plaisant, C. (2010). *Designing the user interface: Strategies for effective human-computer interaction*. Addison-Wesley.

Siemens, G. (2011). LAK 2011: 1st International Conference on Learning Analytics and Knowledge, 1–1.

Siemens, G., & Baker, R. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 252–254. https://doi.org/10.1145/2330601.2330661

Siirtola, H. (2019). The cost of pie charts. *23rd International Conference Information Visualisation (IV)*, 151–156. http://doi.org/10.1109/IV.2019.00034

Siirtola, H., Räihä, K.-J., Istance, H. O., & Spakov, O. (2019). Dissecting Pie Charts. *Interact*, *11747*(387), 688–698.

Simkin, D., & Hastie, R. (2012). An information-processing analysis of graph perception. *Journal of the American Statistical Association*, *82*(398), 454. http://doi.org/10.2307/2289447

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267. http://doi.org/10.1016/s1364-6613(97)01080-2

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, *6*, 532. http://doi.org/10.3389/fpsyg.2015.00532

Skau, D., & Kosara, R. (2016). Arcs, angles, or areas: Individual data encodings in pie and donut charts. *Computer Graphics Forum*, *35*(3), 121–130. http://doi.org/10.1111/cgf.12888

Skau, D., Harrison, L., & Kosara, R. (2015). An evaluation of the impact of visual embellishments in bar charts. *Computer Graphics Forum*, *34*(3), 221–230. http://doi.org/10.1111/cgf.12634

Smith, C. L. (2013). Factors affecting conditions of trust in participant recruiting and retention. In *Proceedings of the 2013 Workshop on Living Labs for Information Retrieval Evaluation*, 13–14. https://doi.org/10.1145/2513150.2513161

Smyth, M. M., & Scholey, K. A. (1996b). Serial order in spatial immediate memory. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *49*(1), 159–177. https://doi.org/10.1080/713755615

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, *5*(1), 61–77. http://doi.org/10.1002/acp.2350050106

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29. http://doi.org/10.1037/h0093759

St-Cyr, O., & Hollands, J. G. (2003). Judgments of proportion with graphs: Object-based advantages. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(13), 1659–1662. http://doi.org/10.1177/154193120304701314

STEM Partnerships Forum. (2018). Optimising STEM industry-school partnerships: Inspiring Australia's next generation. *Carlton South Education Counci*l,1–103.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, *19*(1), 97–140. http://doi.org/10.1016/0364-0213(95)90005-5

Stoeger, H., & Ziegler, A. (2008). Evaluation of a classroom based training to improve self-regulation in time management tasks during homework activities with fourth graders. *Metacognition and Learning*, *3*(3), 207–230. http://doi.org/10.1007/s11409-008-9027-z

Stone, E. R., Gabard, A. R., Groves, A. E., & Lipkus, I. M. (2015). Effects of numerical versus foreground-only icon displays on understanding of risk magnitudes. *Journal of Health Communication*, *20*(10), 1230–1241. https://doi.org/10.1080/10810730.2015.1018594

Sun, Y., Li, S., Bonini, N., & Su, Y. (2011). Graph-framing effects in decision making. *Journal of Behavioral Decision Making*, *25*(5), 491–501. http://doi.org/10.1002/bdm.749

Talbot, J., Setlur, V., & Anand, A. (2014). Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2152–2160. http://doi.org/10.1109/TVCG.2014.2346320

Tan, J. P.-L., Yang, S., Koh, E., & Jonathan, C. (2016). Fostering 21st century literacies through a collaborative critical reading and learning analytics environment. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 430–434. http://doi.org/10.1145/2883851.2883965

ten Have, P. (1999). *Doing conversation analysis: A practical guide*. Sage Publications.

Tervakari, A.-M., Silius, K., Koro, J., Paukkeri, J., & Pirttila, O. (2014). Usefulness of information visualizations based on educational data. *2014 IEEE Global Engineering Education Conference (EDUCON)*, 142–151. http://doi.org/10.1109/EDUCON.2014.6826081

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66–73. http://doi.org/10.1037/0022-0663.95.1.66

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11(1), 99-113.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. http://doi.org/10.1038/381520a0

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. http://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. http://doi.org/10.1080/13546783.2013.844729

Torii, K. (2018). Connecting the worlds of learning and work: Prioritising school-industry partnerships in Australia's education system. *Mitchell Institute*, 2, 1–31.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127–145. http://doi.org/10.1016/S0953-5438(00)00031-X

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition*, *14*(4-8), 411–443. http://doi.org/10.1080/13506280500195250

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. http://doi.org/10.1016/0010-0285(80)90005-5

Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology. General*, *114*(3), 285–310. https://doi.org/10.1037/0096-3445.114.3.285

Tseng, P., & Bridgeman, B. (2010). Information at hand is detected better in change detection. *Journal of Vision*, *10*(7), 209–209. http://doi.org/10.1167/10.7.209

Tufte, E. (2001). Reading and writing with images: a review of four texts. Visual Explanations: Images and Quantities, Evidence and Narrative. *Computers and Composition*, *18*(1), 89–91. Elsevier Inc. http://doi.org/10.1016/S8755-4615(00)00050-5

Tufte, E. R. (1990). *Envisioning Information*. Graphics Press.

Van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). Analytics in higher education: Establishing a common language. *EDUCAUSE learning initiative*, 1(1), l-ll.

Van den Haak, M. J., De Jong Technical, M., & Schellens, P. J. (2007). Evaluation of an informational web site: Three variants of the think-aloud method compared. *Technical Communication*, *54*, 58–71.

van Montfort, X. (2007). What makes a difference: A method to determine whether a change in an image affects the perceived gist. *Behavior Research Methods*, *39*(3), 399–406. http://doi.org/10.3758/BF03193009

van Wijk, J. J. (2013). Evaluation: A Challenge for Visual Analytics. *Computer*, *46*(7), 56–60. http://doi.org/10.1109/MC.2013.151

VanRullen, R., & Thorpe, S. J. (2016). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, *30*(6), 655–668. http://doi.org/10.1068/p3029

Vargo, S., & Lusch, R. (2012). The nature and understanding of value: A service-dominant logic perspective. In *Special Issue – Toward a Better Understanding of the Role of Value in Markets and Marketing*. Emerald Group Publishing Limited, 9, 1–12. https://doi.org/10.1108/S1548-6435(2012)0000009005

Vaskevich, A., & Luria, R. (2018). Adding statistical regularity results in a global slowdown in visual search. *Cognition*, 174, 19–27. https://doi.org/10.1016/j.cognition.2018.01.010

Venkatraman, Overmars, & Wahr. (2019). Visualization and experiential learning of mathematics for data analytics. *Computation*, *7*(3), 37–13. http://doi.org/10.3390/computation7030037

Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., & Duval, E. (2011). Dataset-driven research for improving recommender systems for learning. In *Proceedings of the 1st International Conference Learning Analytics & Knowledge*, 44–53. https://doi.org/10.1145/2090116.2090122

Verbert, K., Ochoa, X., De Croon, R., Dourado, R. A., & De Laet, T. (2020). Learning analytics dashboards. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge*, 35-40. http://doi.org/10.1145/3375462.3375504

Viberg, O., Khalil, M., & Baars, M. (2020). Self-regulated learning and learning analytics in online learning environments. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge*, 524-533. http://doi.org/10.1145/3375462.3375483

Vozniuk, A., Holzer, A., & Gillet, D. (2014). Peer assessment based on ratings in a social media course. In *Proceedings of the 4th International Conference on Learning Analytics and Knowledge*, 133-137. http://doi.org/10.1145/2567574.2567608

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*, 53(4), 5-23. https://doi.org/10.1007/BF02504682

Ware, C. (2012). *Information Visualization: Perception for Design*, Morgan Kaufmann.

Wei, L., & Hindman, D. B. (2011). Does the Digital Divide Matter More? Comparing the Effects of New Media and Old Media Use on the Education-Based Knowledge Gap. *Mass Communication and Society*, *14*(2), 216–235. http://doi.org/10.1080/15205431003642707

Welsh, M., Burns, N., & Delfabbro, P. H. (2013). The Cognitive Reflection Test - how much more than Numerical Ability? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 1587-1592.

Wengraf, T. (2020). Qualitative Research Interviewing. SAGE Publications, 1-49. http://doi.org/10.4135/9781849209717

Whitelock, D., Twiner, A., Richardson, J. T. E., Field, D., & Pulman, S. G. (2015). OpenEssayist - a supply and demand learning analytics tool for drafting academic essays. In *Proceedings of the 5th International Conference on Learning Analytics & Knowledge*, ACM, 208–212. http://doi.org/10.1145/2723576.2723599

Winne, P. H., & Hadwin, A. F. (1998). Studying as Self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in Educational Theory and Practice*, Routledge. http://doi.org/10.4324/9781410602350

Wise, A. F. (2014). Designing Pedagogical Interventions to Support Student Use of Learning Analytics. In *Proceedings of the 4th International Conference on Learning Analytics & Knowledge*, ACM, 203-211. http://doi.org/10.1145/2567574.2567588

Wise, A. F., Zhao, Y., & Hausknecht, S. N. (2014). Learning analytics for online discussions: Embedded and extracted approaches. *Journal of Learning Analytics*, *1*(2), 48–71.

Wolfe, J. M. (2010). Visual search. *Current Biology*, *20*(8), R346–R349. http://doi.org/10.1016/j.cub.2010.02.016

Wong, B. T.-M., & Li, K. C. (2019). A review of learning analytics intervention in higher education (2011–2018). *Journal of Computers in Education*, *7*(1), 7–28. http://doi.org/10.1007/s40692-019-00143-7

Wu, C.-C., Wang, H.-C., & Pomplun, M. (2014). The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes. *Vision Research*, *105*(C), 10–20. http://doi.org/10.1016/j.visres.2014.08.019

Xiao, Z., Wauck, H., Peng, Z., Ren, H., Zhang, L., Zuo, S., et al. (2018). Cubicle: An adaptive educational gaming platform for training spatial visualization skills. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, 91-101. http://doi.org/10.1145/3172944.3172954

Yannakakis, G. N., & Martínez, H. P. (2015). Ratings are Overrated! *Frontiers in ICT*, *2*. http://doi.org/10.3389/fict.2015.00013

Yau, N. (2013). Data points: Visualization that means something (1. Aufl.). Wiley.

Yin, R. K. (2009). Case study research and applications: Design and methods (4 ed.). Sage Publications.

Yoon, S. Y. (2011). *Psychometric Properties of the Revised Purdue Spatial Visualization Tests: Visualization of Rotations (The Revised PSVT-R).* ProQuest.

Yue, C. L., Castel, A. D., & Bjork, R. A. (2012). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition*, *41*(2), 229–241. http://doi.org/10.3758/s13421-012-0255-8

Zhao, J. & Yu, R.Q. 2016. Statistical regularities reduce perceived numerosity. *Cognition*. 146, (2016), 217–222.

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the Subjective Numeracy Scale: Effects of Low Numeracy on Comprehension of Risk Communications and Utility Elicitations. *Medical Decision Making*, *27*(5), 663–671. http://doi.org/10.1177/0272989X07303824

Zimmer, H. D., & Liesefeld, H. R. (2011). Spatial information in (visual) working memory. In A. Vandierendonck & A. Szmalec (Eds.), *Current issues in memory. Spatial working memory*, 46–66. Psychology Press.

Zimmerman, B. J. (2012). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, *81*, 329–339.

Zimmerman, B. J., & Bandura, A. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, *29*, 663–676.

Zuk, T., Schlesier, L., Neumann, P., Hancock, M. S., & Carpendale, S. (2006). Heuristics for information visualization evaluation. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, 1-6. http://doi.org/10.1145/1168149.1168162

Appendix.

Experiment 5 interview script

	Welcome
	Thank you for joining me today, I really appreciate it. It is feedback from students, who are doing the work of learning, that will help us to improve learning analytics tools. (Repeat the part about the consent and reiterate about withdrawing any time they want.) I am recording this interview to help me with transcription. Once it is transcribed, this recording will be deleted. Do you have any questions?
1	Just a housekeeping thing… If I ask you a question again, or it seems like I'm asking the same question, it is because I'm doing my best to understand your point of view. Anything you can tell me about how you were feeling at the time, and how you understood what the visualizations displayed will go a long way toward helping to improve the usability of learning analytics visualizations.
2	So, are you ready to get started?
3	Would you please state your name and the class in which you used the LAD?
	Numeracy
4	In preparation for this interview you were asked to complete a quick survey about your perceived numerical ability and your preference for numerical information. What do you think that it indicated?
5	This is pretty similar to what you entered in the survey. (Share survey results.) - OR - What you entered in the survey was slightly different. (Share survey results.)
6	How would you rate your numerical abilities?
7	Do you have a preference for information to be presented in words or in text?
	Online Discussions
8	What you want to get out of IAT (222 or 334)? (In response to if they hesitate or seem like they don't want to answer.) I ask because I want to get an idea of how important this class is to you. For example, is it required or is it an elective?

9	Can you give me an idea of what your previous experience in online learning discussions has been? (Both classes have an online discussion the week before the one with the LAD.)
10	Was this discussion similar or different?
11	Had you participated in an online course discussion before this class?
12	Would you say that you are comfortable learning online, or not really?
13	(Prompt in case response to the previous question is not descriptive.) What do you usually do in discussions like this? How would you approach them? For example, do you tend to post early or late, define these discussions confusing and hard to follow, do you do all the readings before you do the discussion, or do you check the discussion first?
14	When you started this discussion activity, what were your goals?
15	What if you had to choose only one of the following four goal descriptions to describe your desire for only this learning activity, which one do you think is most accurate? Would you say that your goal was to get high marks, to do the minimum amount of work, to learn as much as possible, are to avoid a low mark? (These categories correspond to the four achievement goal orientation categories.)
16	What did you understand about the assignment requirements? (If reply is short.) What were the expectations for this assignment? For example, how was it graded?
17	What was your approach to the current discussion? What did you do first?
18	Did you relate to the subject matter of the discussion?
	First Time Accessing LAD
19	Okay now it is time to review the visualizations! This is a re-creation of the LAD from the first time that you accessed it. I want you to take your time and tell me everything that you remember. For example, why did you click on the LAD the very first time you access the visualization?
20	The first time that you viewed the visualization, why did you look? What information were you looking for?

21	How did you think you were doing? What part of the visualization, you know the features of the visualization, did you use to come to this conclusion?
22	Did you form your opinion based on only your performance, your performance in comparison to your peers, or something else?
23	(Alternate question) I noticed that you looked at the LAD before you posted. Would you tell me why?
24	(Alternate question.) I noticed you posted, then looked at the LAD almost immediately. Could you tell me why?
25	The first time you looked at the LAD, why did you look? (overview, comparison, see how others are doing, to see something in particular, to see progression in time, change as a result of my last post)
26	Did you expect what you saw? Did anything in the LAD surprise you?
27	What did you do next?
	Second Time Accessing LAD
28	(Describe things that happen next to set the scene) for example (a day later, after your third posts, etc.) When you viewed the LAD next, this is what it looked like. Can you describe how you thought you were doing, based on what you saw?
29	What features of the LAD did you base this opinion on?
30	Did anything you saw surprise you?
31	Why did you return to look at the LAD again?
32	What did you do next?
	Subsequent Times Accessing LAD - Why did they return to the LAD?
33	Can you please describe what you understand from this (indicate LAD)? I am really looking for a summary of what you looked out within the LAD, what you did or didn't understand from it, things like that.
34	The next time you looked at the visualization was ___ (ex. immediately, after several people had posted). Why did you access it then?

35	When you looked at the visualization here, how did you think you were doing based on what you saw? How certain were you (feeling of knowing)? How did you feel you were doing? How confident were you in this?
36	(Alternate) What aspects of the LAD led you to believe this? Did you form this opinion based on your performance, your performance in comparison to your peers, or something else?
37	(Alternate question) you looked at the visualization then posted (however many) minutes later. What did you do between those times? (Example additional research? Reviewed others' posts? Reread the instructions? Emailed someone?)
	Overall Use of LAD - Why did they use the LAD?
38	You tended to (ex. post first then look, look then post, waited until the end, etc.). Why?
39	You mentioned at the beginning of this interview that you felt like you were ___ overall in this class. Did using the LAD change this opinion of your performance that you had going into the assignment?
40	Did your goals change at any point in discussion, and if so, why?
41	Were you surprised by anything that you saw in the LAD?
42	Did you have difficulty understanding the LAD at any point, or were there things that were unclear?
43	Were there any times when you felt the LAD wasn't accurate?
44	Did you find the visualizations helpful? If not, why? Was there a particular point in the discussion that you found the visualizations useful?
45	If available, would you use LAD in future courses? What kind of courses would you find them helpful in?