NORTHWESTERN UNIVERSITY

Towards Practice-Integrated Human-AI Systems for Developing High-Quality Educational Assessments

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Yuan “Charles” Cui

EVANSTON, ILLINOIS

December 2025

Abstract

Assessments are foundational to education, shaping how teachers teach, how students learn, and how accountability is maintained across the educational ecosystem. Yet, creating and delivering high-quality assessments remains one of the most time-consuming aspects of teaching. Educators, especially those in under-resourced schools, often lack the time, training, and support to design effective assessments that are aligned with curricula and diverse student needs. While recent advances in large language models (LLMs) show the potential to reduce this burden, they also raise concerns about quality and teacher trust. In specific domains such as data visualization education, researchers and educators face similar challenges: existing assessments for measuring visualization literacy (the ability to read and interpret visualizations) are lengthy, rigid, and difficult to scale or adapt. Across both general and visualization-specific contexts, improving efficiency, maintaining quality, and building trust in assessment development and delivery requires systems that lower technical barriers, promote content reuse and adaptation, and integrate meaningfully into educators' existing workflows.

This dissertation addresses these challenges through three complementary projects that develop adaptive, scalable, practice-integrated solutions for assessment development and delivery. The first project introduces A-VLAT and A-CALVI, computerized adaptive tests for visualization literacy that maintain precision and reliability while halving test length. The second project presents VILA, a pipeline that leverages LLMs to generate visualization literacy questions at scale, combining automated generation with expert-driven evaluation to ensure quality and validity. The third project extends these ideas to general K–12 education, where we codesigned a web-based system, Ripplet, that supports multilevel reusable interactions with LLMs for creating, editing, and adapting assessments. Together, these projects contribute to the development of assessment systems that improve efficiency, scalability, and quality.

Acknowledgment

It took a village.

First, I would like to thank my advisor, Matthew Kay. I started my Ph.D. as a theory student but realized that I wanted to do more applied work midway through. It was a challenging time for me both as a student and as an early twenty-something who felt lost emotionally. I seriously considered leaving the program. Just on the verge of quitting the Ph.D., I took a seminar with Matt. He seemed kind, so I asked if I could work with him, and he took a chance on me. I had much to learn, and Matt has been patient, constructive, and always, always encouraging when teaching me the craft of research. Over the years, he has always made clear that he wants what makes me happy, whether it is a research project unrelated to his interest or a career outside of academia. I feel extremely fortunate to have an advisor who is invested in helping me find my path.

Fumeng Yang has been my de facto co-advisor. She was a postdoc in my lab when we first met. She had no obligation in mentoring me, but she was always generous with her time and helped me above and beyond. We became close collaborators over time, and she is my go-to person when I needed counsel. After she became a professor at the University of Maryland, she supported me to pursue a passion project as a visiting student in her lab. She taught me to be programmatic, resilient, and detail-oriented. Her impeccable aesthetic taste has also had a profound impact on me: research is not only about the science, but also about the art of presenting the science beautifully.

Lily Ge is my closest student collaborator. After I shifted my research focus, Lily let me work on a project with her. That was the first time in grad school when I felt joy doing research. That project turned out to be a success, and it inspired the subsequent projects in my dissertation. Lily is meticulous and humble, and she always held me accountable.

We challenged each other's ideas all the time and became better researchers. Ph.D. can be a lonely journey, and I am glad to have her as a companion on this road.

I am grateful for my collaborators in my dissertation projects. Lane Harrison and Yiren Ding at Worcester Polytechnic Institute provided crucial support for two projects, and they have offered me research and career advice outside of those projects. For my last thesis project, I worked with a team of brilliant undergraduate students at Northwestern, including Annabel Goldman, Jovy Zhou, Xiaolin Liu, Clarissa Shieh, Joshua Yao, Mia Cheng, April Shi, Laura Félix, Christopher Heo, Rachel Johnson, and Eric Lee. These students have taught me so much more than I could teach them.

I have been fortunate to have other support systems at Northwestern. The MU Colletive Lab is my base camp, and its members (Mandi Cai, Sheng Long, Abhraneel Sarma, Ziyang Guo, Taewook Kim, Maryam Hedayati, Paula Kayongo, and Hyeok Kim) created a loving and supportive environment. The Northwestern Institute on Complex Systems and HCI + D Center have generously provided funding for my research, where I also met my thesis committee members Eleanor O'Rourke and Michael Horn, who have given me invaluable advice. I have also received guidance from many other professors, including Steve Franconeri, Jason Hartline, Haoqi Zhang, Konstantin Makarychev, Aravindan Vijayaraghavan, Elizabeth Tipton, and Duri Long. The student leaders at CSPAC and CSSI have built a strong sense of community within the computer science department, and my life has been better because of it. Elysse Longiotti at the Northwestern Career Development Center has always encouraged me. My work would not have been possible without the incredible staff members of the CS department, including Julia Blend, Katie Winters, Bella Barrios, Melissa Duong, Jensen Smith, Ethan Walles, and Dru Redmond.

I am lucky to have had opportunities to explore my other research interests during

my Ph.D. I want to thank Lily Xu, Ana-Andreea Stoica, Sandro Radovanović, Matthew Olckers, Francisco J. Marmolejo Cossio, Sera Linardi, Wanyi Dai Li, and Rediet Abebe at EAAMO (previously Mechanism Design for Social Good), Daniel Ho and Derek Ouyang at Stanford RegLab, Ugofilippo Basellini, Irena Chen, and Monica Alexander at the Max Planck Institute for Demographic Research, and Wei Zhang, Yiru Chen, Jane Hoffswell, and Abhisek Trivedi at Adobe. I am especially grateful for the Data Science for Social Good Fellowship program (Rayid Ghani, Kit Rodolfa, my teammates, and all the mentors and fellows) at Carnegie Mellon University—it was everything I could have hoped for in a truly collaborative, supportive, and intellectually stimulating environment. These experiences shaped me as a researcher, and more importantly, I made life-long friends who continue to help me grow as a person.

Many other people helped during my Ph.D. My undergraduate mentor Samuel Taggart at Oberlin College, a former Northwestern CS Ph.D. graduate, continued to support via numerous email replies and Zoom calls. I am lucky to have met people in the data visualization and HCI community who were willing to spend their time helping me, including Leilani Battle, Steven Moore, Katy Williams, Alex Kale, Niklas Elmqvist, Mi Feng, Cindy Xiong, Xiaoying Pu, and Bum Chul Kwon.

This journey would not have been possible without friends and family. The past five years were made so much better thanks to the kindness and companionship of Jennah Gosciak, Zeyuan Zhang, Xinnan Cheng, Daniel Firebanks-Quevedo, Tao Hong, Martin Mancini, Yuhao Wu, Kennedy Casey, Jane Hsieh, Brian Wesley Hill, Yasmine Ramachandra, Catherine Wahlenmayer, and the MatSci basketball teams. I am especially grateful for Hannah Sommerlad and the Sommerlad family for their support. 最后，感谢爸爸妈妈无条件的爱，也感谢他们从我十五岁起就让我自由地追求自己的人生。

Table of Contents

Chapter 1

Introduction

1.1 Background and Motivation

Assessments—defined in this dissertation as paper- or computer-based instruments composed of answerable questions such as homework, quizzes, tests, and exams, rather than broader tasks like labs or class projects—are critical in the educational ecosystem. They affect how teachers teach and how students learn and ensure accountability in all stakeholders [20, 124, 126]. Formative assessments such as quizzes and homework help teachers monitor students' progress and guide instruction, while summative assessments such as unit tests and final exams provide a snapshot of student achievement for schools, parents, and policymakers [117, 130, 11]. Despite their central role, **creating and delivering high-quality assessments remains one of the most challenging and time-consuming responsibilities in education**. Teachers spend up to half their time writing and revising assessments, aligning them with standards, and adapting them for students with different needs [126, 140]. This intensive labor leaves little time for reflection or improvement, and it is compounded by systemic disparities: teachers in under-resourced schools often have fewer planning periods and limited access to high-quality materials. As a result, assessment quality and equity vary widely across classrooms [78, 124, 131, 45].

Efforts to improve the efficiency of creating and delivering high-quality assessments have a long history in educational research. Early computational methods sought to reduce teacher workload through template-based automatic question generation (AQG), where question developers encoded reusable question templates that could be programmatically instantiated with different parameters [37, 99, 78, 105, 134]. These approaches found some success in domains such as mathematics and medical licensing, but they often produced questions that were repetitive or lacked pedagogical depth. In recent years, advances in large language models (LLMs) have renewed interest in automating question creation [77, 65, 87, 31, 98]. LLM-based systems can generate multiple-choice, short-answer, or open-ended questions across a wide range of subjects. Yet these systems introduce new risks for quality and trust. LLMs may generate factually incorrect or ambiguous questions or overlook alignment with local curricula and classroom realities. More importantly, most existing tools treat teachers as passive consumers of AI output rather than as active collaborators. Assessment authoring is a personal and iterative process, grounded in teachers' pedagogical values and their nuanced understanding of student needs [37, 149, 84, 2]. Without mechanisms that preserve teacher agency and enable oversight, automation may produce large volumes of content that are efficient but unreliable—or worse, erode teachers' confidence in the assessment process itself.

Similar issues arise in specialized domains such as data visualization education, where assessing people' ability to interpret and reason about data visualizations—known as *visualization literacy*—is increasingly important [15, 81, 55]. Visualization literacy affects decision-making in high-stakes contexts ranging from public health to environmental policy, yet measuring this ability reliably and efficiently remains difficult. Existing assessments such as the Visualization Literacy Assessment Test (VLAT) [81] and the Critical Thinking Assessment for Literacy in Visualizations (CALVI) [55] provide strong founda-

tions but share many of the same limitations seen in general education: they are time-consuming to administer, rigid in structure, and labor-intensive to create, adapt, and maintain.

Across both general and visualization-specific contexts, a common theme emerges: assessments are vital but difficult to create, adapt, and deliver efficiently while maintaining quality and trust. Through the course of this dissertation, I explore these challenges and, in doing so, identify a set of design principles for addressing them. Across three projects, we find that effective assessment systems must (1) lower technical barriers for educators and researchers, (2) promote reuse and adaptation of existing high-quality materials, (3) integrate seamlessly into teachers' authentic workflows, and (4) preserve the transparency and control that underpin teachers' trust in AI-assisted processes. In K–12 classrooms, this means developing tools that empower teachers to collaborate with AI while retaining authority over generated content. In data visualization education, it means creating adaptive assessments efficient in measurement and automated methods to generate visualization questions that maintain quality. Together, these findings motivate my **thesis statement**:

> Assessments in K–12 classrooms and data visualization education are important but difficult for educators to create, adapt, and deliver. Designing systems that lower technical barriers, reuse and adapt existing content, and integrate into teachers' existing practices can help improve the efficiency of assessment development and ensure trust and quality.

1.2 Thesis Contributions

This dissertation responds to these challenges in K–12 educational assessments and visualization literacy assessments. Across three projects, I develop solutions that make assessments adaptive in measurement, scalable in content generation, and integrated into teachers' authoring practices.

1.2.1 Adaptive Assessment of Visualization Literacy (Chapter 3)

This work is done in collaboration with Lily Ge, Yiren Ding, Fumeng Yang, Lane Harrison, and Matthew Kay.

To improve the efficiency of assessment delivery in data visualization education, we developed computerized adaptive tests (CATs) for visualization literacy. Building on existing assessments, including VLAT and CALVI, we applied item response theory and adaptive testing algorithms to create A-VLAT and A-CALVI, which measure the same skills in about half the number of questions. Specifically, we (1) employed item response theory (IRT) and non-psychometric constraints to construct adaptive versions of the assessments, (2) finalized the configurations of adaptation through simulation, (3) refined the composition of test items (questions) of A-CALVI via a qualitative study, and (4) demonstrated the test-retest reliability (ICC: 0.98 and 0.98) and convergent validity (correlation: 0.81 and 0.66) of both CATs via four online studies. Although these results demonstrate that adaptive assessments can more efficiently and reliably measure visualization literacy, their adoption in practice has been limited. Despite frequent citations, few researchers have deployed these adaptive assessments, largely because hosting the adaptive algorithms requires setting up a backend service, which is an obstacle even for visualization

researchers. Moreover, because existing question banks are small, repeated administrations to the same participants often reintroduce identical questions, limiting longitudinal use. These challenges reveal the need to lower technical barriers and to develop scalable methods for generating high-quality visualization assessment questions.

1.2.2 VILA: Scalable Generation of Visualization Questions (Chapter 4)

This work is done in collaboration with Lily Ge, Yiren Ding, Lane Harrison, Fumeng Yang, and Matthew Kay.

Building on the need of having large question banks from the previous project, we built VILA (**V**isualization **I**tems generated by **LA**rge language models), a three-staged pipeline that uses large language models to generate visualization questions at scale. Working with visualization experts, we developed an evaluation rulebook that formalizes quality criteria and guided the review of over a thousand generated items (questions). The resulting VILA bank provides a validated collection of visualization questions. Through our evaluation, we found that 21% of generated questions contained errors, which we classified to reveal common pitfalls in LLM-generated assessment content for visualization. This analysis highlighted that maintaining question quality requires human oversight not only at the end of the generation process but throughout it. Human feedback can guide generation, correction, and validation iteratively. Despite the methodological contributions and frequent citations of VILA, adoption among visualization educators and researchers has remained limited, in part due to the technical barrier of setting up local Python environments and also the one-off nature of the generation workflow that is not naturally integrated into teachers' practices. These lessons reveal the need to design AI-based assessment systems that embed human oversight directly into the authoring process and fit into users' existing workflows.

1.2.3 Ripplet: Authoring Educational Assessments through Multilevel Reusable Interactions with LLMs (Chapter 5)

This work is done in collaboration with Annabel Goldman, Jovy Zhou, Xiaolin Liu, Clarissa Shieh, Joshua Yao, Mia Cheng, Matthew Kay, and Fumeng Yang.

Building on these lessons about the need for human oversight and better integration into educators' workflows, we develop Ripplet, a web-based and LLM-powered assessment authoring system. To ensure that the system can integrate into teachers' existing workflows, we partnered with 13 teachers to codesign Ripplet. Ripplet supports creating questions from diverse inputs, adapting content at multiple levels, and tracking and reapplying teachers' edits. Over seven months, Ripplet enabled teachers to create formative assessments they would not have otherwise made, shifted their practices from generation to curation, and helped them reflect more on assessment quality. In a user study with 15 additional teachers, compared to their current practices teachers felt the results were more worth their effort ($\mu = +1.93/10$, 95% CI: $[0.33, 3.53]$) and that assessment quality improved ($\mu = +1.11$ $[0.10, 2.11]$). Unlike VILA, Ripplet achieved sustained and organic adoption. The system is used by teachers across more than ten schools in four states and continues to attract new users through word of mouth, even after formal studies concluded. This reflects the importance of codesign in ensuring that systems fit into teachers' workflows. Rather than using LLMs to generate everything from scratch, Ripplet's design emphasizes reusing and adapting existing high-quality materials, allowing teachers to maintain ownership while benefiting from AI assistance. These outcomes demonstrate the thesis of this dissertation: systems which lower technical barriers, promote reuse and adaptation, and integrate seamlessly into teachers' authentic workflows can make assessment development more efficient and ensure assessment quality.

Chapter 2

Related Work

2.1 Assessment Development

Developing assessments is a time-intensive process [94, 140, 78, 124], consuming up to half of teachers' professional time [126, 125]. They need to align assessments with learning objectives, devise intellectually demanding tasks, and ensure instructional relevance [133, 56, 42]. However, most teachers receive little training or support for assessment development [124, 125, 131, 56, 86, 42, 95, 41, 131], leading to low-quality assessments and consequently uneven chances among students to demonstrate their knowledge [78, 124, 131, 45].

Researchers in education, psychology, and cognitive science have proposed measurement theories and models to guide assessment development, such as constructing detailed test blueprints [35, 136] and estimating item parameters (e.g., difficulty) [5, 50, 64]. Some argue that models of cognition and learning can inform educational assessment development, making assessments more effective in measuring student understanding [107]. However, translating such research into practice is not easy [94, 107]. In reality, teachers rarely adopt such principles when creating assessments, as these theories often

lack relevance to their day-to-day instruction [94, 128]. For instance, most theories of educational measurement are designed for large-scale or standardized testing, emphasizing psychometric properties rather than the formative and summative purposes of classroom assessment. As a result, these theories often feel disconnected from teachers' day-to-day realities, leading them to rely on experience and institutional norms instead of abstract theoretical principles. Consequently, existing work in test development theory provides insufficient practical guidance for how teachers currently develop assessments.

2.2 Automatic Question Generation

Questions are the building blocks of an assessment. Traditional question creation processes are highly costly and laborious due to their manual nature. They usually involve subject-matter experts writing and reviewing questions. Thus, researchers have sought ways to reduce the amount of human labor required in this process. This field of study is *automatic question generation* (AQG) [79]. AQG methods aim to ease the burden of creating questions [37, 99, 78], gaining momentum with the recent advances in generative AI [99, 78]. These methods range from template-based generation to neural and transformer models [105, 134, 148, 135, 113, 23, 141]. Recent LLM-based methods dramatically lowered the technical barrier to entry [77, 65, 87, 31, 98, 49, 46], and can take inputs like Bloom's Taxonomy and textbooks [90, 39, 89, 88] to generate questions of varying difficulty levels [26, 30] in multiple choice, free response, and fill-in-the-blank formats [99]. While these methods have shown promise in generating questions across diverse contexts, they often struggle to produce high-quality questions that can measure higher cognitive abilities [90, 39, 53, 30, 49].

Additionally, there exists a gap between these automated methods and teachers'

practices. Most automated solutions are not designed for teachers: some are end-to-end pipelines that require technical expertise (e.g., coding) to execute, while others use generic interfaces that lack targeted support for assessment authoring. In addition, effective assessments must closely align with individual teachers' curricula, institutional guidelines, and the unique needs of student cohorts [106, 41, 46]—requirements that cannot be met by sifting through thousands of generic questions. Assessment authoring is also a deeply personal practice that embodies teachers' pedagogical philosophy, beliefs, and years of accumulated knowledge about their students [37, 149, 84, 2]. Automated methods without careful design to involve human oversight do not fit real-world classrooms.

2.3 Human-AI Collaboration for Assessment Development

Prior work has explored ways to incorporate human oversight into LLM-based question generation [85, 73, 31, 48, 120]. TutorCraftEase allows teachers to generate questions with LLMs and accept, reject, and manually edit LLM outputs. This enables teachers to produce quality questions more efficiently [73]. Similarly, ReadingQuizMaker helps college instructors make questions for reading quizzes, providing control for when to request AI assistance; the authors found that instructors preferred this collaborative approach over an AI-only process [85]. While these systems show promise in mitigating the issues of fully automated approaches, they operate largely at the level of individual questions and provide limited support for reusing and adapting questions. Assessment authoring is inherently **multilevel**: it requires not only generating individual questions but also structuring them into a holistic assessment, tweaking specific parts of a question to fit students' needs, balancing difficulty and topic coverage, and aligning the collection to curricular goals. In addition, as we find in Chapter 5, assessment authoring is also **it-**

erative: teachers move between creating and adapting questions, searching for relevant materials, reviewing and restructuring the content, and refining their requirements for an assessment. Designers of assessment development technology should consider the multilevel and iterative properties of this process.

2.4 Computerized Adaptive Assessment

After developing the content of an assessment, educators often face the challenge of assessment delivery. Traditional assessment delivery methods often rely on fixed tests that treat all students the same, regardless of their individual ability levels. These assessments can be lengthy and inefficient. Moreover, they can increase the risk of disengagement or even cheating, as students face repetitive or poorly matched questions that fail to adapt to their skill level [121].

Computerized adaptive testing (CAT) is a form of computer-based assessment where the next question or the next set of questions a test-taker sees depends on their performance on the previous questions. The idea is to adaptively select questions to tailor to a test-taker's ability in order to gain more information about them: for example, if someone performs poorly on difficult questions, they will then be presented with an easier question. CAT has many advantages over traditional static assessments where everyone receives the same questions: it can achieve similar measurement precision with fewer questions, and it may better motivate test takers because the questions are more appropriate for their ability level [121].

CAT has been broadly adopted in healthcare research. Clinical researchers have developed various CATs to measure patient characteristics, such as quality of life [58], anxi-

ety [60, 138], and mental health disorders [59]. CAT has also been used to measure various forms of literacy, such as English literacy [51], health literacy [72], and math knowledge of university students [57]. CAT has also been widely applied to many large-scale standardized tests in practice, such as the Graduate Record Examinations (GRE) [47] and the Graduate Management Admission Test (GMAT) [122], and continues to grow in popularity: the SAT, a long-established college admissions test in the U.S., became digital and adopted adaptive testing since 2023 and 2024 [36]. Despite the prevalent prior applications and potential benefits of CAT, it has yet to be applied in data visualization education.

Chapter 3

Adaptive Assessment of Visualization Literacy

This work is done in collaboration with Lily Ge, Yiren Ding, Fumeng Yang, Lane Harrison, and Matthew Kay. This chapter is published as a full paper at the IEEE VIS *Conference in 2023 [38].*

3.1 Introduction

Visualization literacy—an individual's ability to understand and interpret visualizations—can significantly impact data-driven decisions. People may rely on data visualizations showing the spread of infectious diseases to make personal health decisions, to decide between treatment options in medical settings, to make financial decisions, or to engage with social or political topics. Inaccurate interpretations of such visualizations may lead to faulty reasoning and decisions, as well as harmful outcomes.

In the study of visualization literacy, there is a persistent need for accurate, reliable, and timely ways to assess people's ability. For example, visualization researchers may want to track the progress of this ability over time, or to empirically evaluate interventions designed to enhance a target group's ability to interpret visualizations. Such efforts require concise, quantitative assessments of visualization literacy that can be adminis-

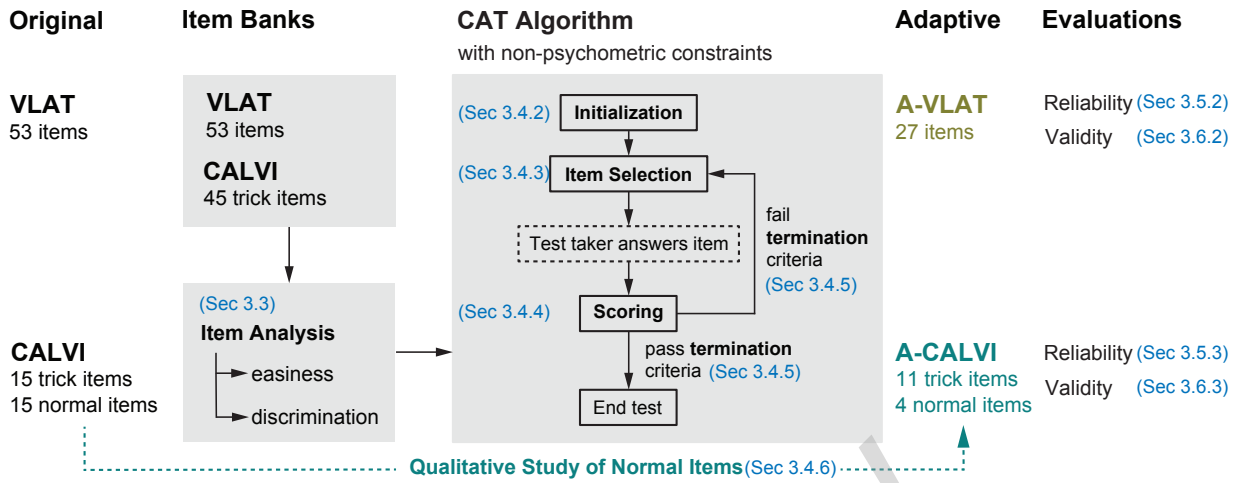

Figure 3.1: The process of developing adaptive visualization literacy assessments: A-VLAT and A-CALVI. We start with the item banks (i.e., question banks) of VLAT and CALVI and use the parameters from item analysis to construct the CAT algorithms for the adaptive assessments. We then evaluate their validity and reliability via four online studies. The annotations in blue are the components' corresponding sections.

tered to individuals multiple times, allowing for the measurement of skill development and the evaluation of intervention efficacy through pre- and post-testing.

While researchers have devised assessments to measure both basic visualization literacy skills [81, 15] and the critical thinking ability to detect visualization misinformation [55], these tests can be time-consuming. To address this, we apply *computerized adaptive testing* (CAT) to visualization literacy. CAT's core principle is to adaptively select test items (i.e., questions)[1] for a test taker based on their performance on previously-answered items, and it can provide a precise estimate of the test-taker's abilities with fewer items. CAT has been widely applied in healthcare science [58, 60, 138, 59] and used in practice by educational agencies, such as the Educational Testing Service (ETS) and College Board, to create large-scale standardized tests [47, 122, 36]. However, visualization literacy assessments have yet to take advantage of the benefits of adaptive testing.

[1]In this dissertation, the terms *item* and *question* are used interchangeably.

In this project, we adopt the CAT development framework [132] to create two short, adaptive visualization literacy tests: A-VLAT and A-CALVI, which are built upon the existing static assessments VLAT [81] and CALVI [55]. First, we compute *item parameters* (how easy items are and how well items in the bank separate test takers of different abilities) with Item Response Theory (IRT). Using the item parameters, we construct adaptive algorithms to select items for test-takers. We include non-psychometric constraints in our algorithms to ensure the tests contain balanced content; i.e., to ensure A-VLAT covers all 12 chart types and 8 tasks from VLAT and A-CALVI covers all 11 misleaders from CALVI. We also conduct a qualitative study to refine the composition of items in A-CALVI, reducing its length by a further 11 items. Ultimately, we contribute:

1. A demonstration of the refinement of visualization literacy assessments through adaptive testing, including the incorporation of non-psychometric features of existing assessments.

2. Two valid and reliable adaptive visualization literacy tests, A-VLAT (27 items) and A-CALVI (15 items), which are half the length of their non-adaptive counterparts.

3. Evidence from four online studies demonstrating the test-retest reliability (ICC: 0.98 and 0.98) and convergent validity (correlation: 0.81 and 0.66) of these tests.

In addition, we discuss how cumulative results from visualization literacy studies can better inform our understanding of the relationship between the constructs measured by literacy assessments (e.g., what correlations between chart types might tell us about how people understand visualizations). We also provide recommendations for using and customizing visualization literacy assessments based on test administrators' needs. Our work just scratches the surface of the potential for adaptive testing in visual-

ization; we believe it offers a path to shorter, repeatable, and more reliable assessment of visualization literacy.

3.2 Background

3.2.1 Visualization Literacy

Within the visualization community, many researchers have studied visualization literacy through the development of assessments [15, 81, 55] and frameworks [14]. Boy et al. used Item Response Theory (IRT) to generate a set of tests that aims to assess people's ability to interpret line charts, bar charts, and scatterplots [15]. Similarly, to test people's ability to interpret visually represented data, Lee et al. developed a Visualization Literacy Assessment Test (VLAT) that contains 53 multiple-choice items [81]. Later, considering the complexity of visualization literacy as a construct, Ge et al. expanded on prior definitions to incorporate the ability to identify and reason about visualization misinformation and developed a Critical Thinking Assessment for Literacy in Visualizations (CALVI), which has a bank of 45 items [55]. We rely heavily on both VLAT and CALVI to develop our adaptive tests, A-VLAT and A-CALVI, so we explain VLAT and CALVI in more detail below.

VLAT. The items in VLAT were generated from 12 chart types with 8 tasks such as *retrieve value* and *make comparisons*, and each test taker is expected to take all 53 items [81]. For example, Figure 3.2.A is a VLAT item in a pie chart that asks viewers to make comparisons. Each item in VLAT has an item difficulty and item discrimination index obtained from Classical Test Theory (CTT) analysis [81].

CALVI. The items in CALVI were generated from 11 misleaders (i.e., ways a chart can lead