Adaptive Assessment of Visualization Literacy

Yuan Cui, Lily W. Ge, Yiren Ding, Fumeng Yang, Lane Harrison, and Matthew Kay

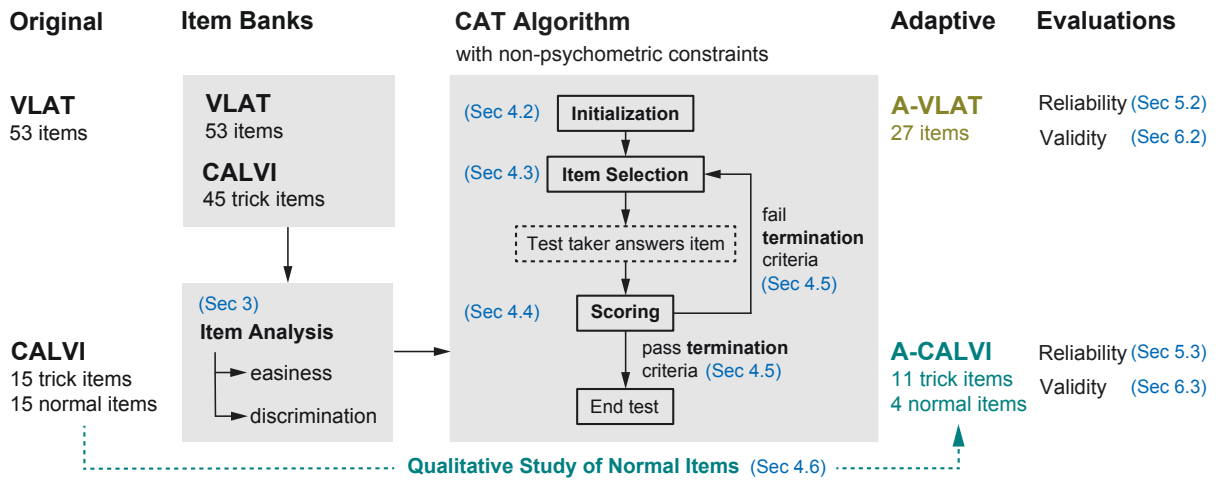

Fig. 1: The process of developing adaptive visualization literacy assessments: A-VLAT and A-CALVI. We start with the item banks of VLAT and CALVI and use the parameters from item analysis to construct the CAT algorithms for the adaptive assessments. We then evaluate their validity and reliability via four online studies. The annotations in blue are the components' corresponding sections.

Abstract—Visualization literacy is an essential skill for accurately interpreting data to inform critical decisions. Consequently, it is vital to understand the evolution of this ability and devise targeted interventions to enhance it, requiring concise and repeatable assessments of visualization literacy for individuals. However, current assessments, such as the Visualization Literacy Assessment Test (VLAT), are time-consuming due to their fixed, lengthy format. To address this limitation, we develop two streamlined computerized adaptive tests (CATs) for visualization literacy, A-VLAT and A-CALVI, which measure the same set of skills as their original versions in half the number of questions. Specifically, we (1) employ item response theory (IRT) and non-psychometric constraints to construct adaptive versions of the assessments, (2) finalize the configurations of adaptation through simulation, (3) refine the composition of test items of A-CALVI via a qualitative study, and (4) demonstrate 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. We discuss practical recommendations for using our CATs and opportunities for further customization to leverage the full potential of adaptive assessments. All supplemental materials are available at `https://osf.io/a6258/`.

Index Terms—Visualization literacy, computerized adaptive testing, item response theory

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 administered 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 [4,22] and the critical thinking ability to detect visualization misinformation [14], 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 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 [16–18,32] and used in practice by educational agencies, such as the Educational Testing Service (ETS) and College Board, to create large-scale standardized tests [10, 12, 28]. However, visualization literacy assessments have yet to take advantage of the benefits of adaptive testing.

In this paper, we adopt the CAT development framework [30] to create two short, adaptive visualization literacy tests: A-VLAT and A-CALVI, which are built upon the existing static assessments VLAT [22] and CALVI [14]. 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

- *Yuan Cui, Lily W. Ge, Fumeng Yang, and Matthew Kay are with Northwestern University. E-mail: charlescui@u.northwestern.edu, {wanqian.ge, fy, mjskay}@northwestern.edu.*
- *Yiren Ding and Lane Harrison are with Worcester Polytechnic Institute. E-mail: {yding5, ltharrison}@wpi.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

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

2 Background

2.1 Visualization Literacy

Within the visualization community, many researchers have studied visualization literacy through the development of assessments [4,14,22] and frameworks [5]. 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 [4]. 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 [22]. 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 [14]. 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 [22]. For example, Fig. 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 [22].

CALVI The items in CALVI were generated from 11 misleaders (i.e., ways a chart can lead to conclusions not supported by the data) and 9 chart types [14]. CALVI consists of 15 *trick* items, which are items whose visualizations contain a misleader, and 15 *normal* items (i.e., items based on well-formed visualizations)—CALVI includes normal items because people may encounter misleading visualizations mixed in with well-formed visualizations in their daily lives. It is worth noting that only test takers' performance on the trick items are used to estimate their abilities to detect visualization misinformation. The 15 trick items are selected to cover all 11 misleaders from CALVI's bank of 45 total trick items [14]. For example, Fig. 2.B is a trick item with misleader *Manipulation of Scales - Inappropriate Use of Scale Functions*, where the number labels on the y-axis do not match the spacings of the tick marks; and Fig. 2.C is a normal item. The trick items in CALVI each have an item easiness and an item discrimination parameter estimated using a 2-parameter IRT model [14].

Besides the assessments above, related fields such as statistical literacy also investigated graph interpretation abilities, and measurements were developed to assess graph comprehension. Specifically, studying how students interpreted graphical representations of distributions, DelMas et al. developed assessment items for the ability to interpret distributions [11]. Other researchers have also developed instruments that target younger audiences [25], and some focused more on the ability to critically interpret graphs, which is related to statistical literacy [1].

Although existing assessments provide measurements for visualization literacy and related abilities, we reiterate that they can be time consuming and inflexible due to their fixed, lengthy format. Next, we give an overview of computerized adaptive testing, which reduces the length of tests while maintaining similar measurement precision.

2.2 Computerized Adaptive Testing (CAT)

Computerized adaptive testing (CAT) is a form of computer-based test where the next item or the next set of items a test-taker sees depends on their performance on the previous items. The idea is to adaptively select items to tailor to a test-taker's ability in order to gain more information about them: for example, if someone performs poorly on difficult items, they will then be presented with an easier item. CAT has many advantages over traditional static assessments where everyone receives the same items: it can achieve similar measurement precision with fewer items, and it may better motivate test takers because the items are more appropriate for their ability level [27].

CAT has been broadly adopted in healthcare research. Clinical researchers have developed various CATs to measure patient characteristics, such as quality of life [16], anxiety [18, 32], and mental health disorders [17]. CAT has also been used to measure various forms of literacy, such as English literacy [13], health literacy [21], and math knowledge of university students [15].

CAT has also been widely applied to many large-scale standardized tests in practice, such as the Graduate Record Examinations (GRE) [12] and the Graduate Management Admission Test (GMAT) [28], and continues to grow in popularity: the SAT, a long-established college admissions test in the U.S., is planning to become digital and adopt adaptive testing starting in 2023 and 2024 [10].

Despite the prevalent prior applications and potential benefits of CAT, it has yet to be applied towards visualization literacy assessments. We address this gap and explore the benefits of CAT in visualization literacy, developing shorter, adaptive assessments, along with a systematic CAT development procedure across two existing visualization literacy tests.

2.3 Procedure of CAT Development

Although existing measurement methods are not adaptive, as explained in Sec. 2.1, they provide a valuable foundation for the development of CAT, which requires item parameters from IRT item analysis. We extend and adopt the framework proposed by Thompson and Weiss [30] to develop adaptive tests. Below, we outline steps from this framework that were instrumental to developing our test.

First, test developers need to conduct **item analysis** of the item bank using test tryout data (i.e., data collected on participants answering items before the test is deployed in practice). In this stage, IRT item analysis is applied on test tryout data to compute the item easiness (i.e., how easy an item is) and item discrimination (i.e., how well an item differentiates test takers of different levels of ability) parameters of the items in the bank. After the test developers obtain these parameters for the items, they can proceed to the **construction of CAT algorithms**. A CAT algorithm consists of four main components:

1. Initialization: initialize an estimate of the test-taker's ability;
2. Item selection: select the next item for the test-taker based on their current estimated ability (score);
3. Scoring: compute the score for a test-taker after each item;
4. Termination: end the test when the termination criterion is met.

To build a CAT algorithm, test developers use the test tryout data, the item parameters, and simulation studies to determine the configurations of these main components and demonstrate the measurement precision of the CAT. Although the framework thus far does not outline steps specifically for establishing reliability and validity, in the development of tests in general, they should be assessed [9]. Therefore, for both CATs we also evaluate the **reliability** (i.e., examining the extent to which the

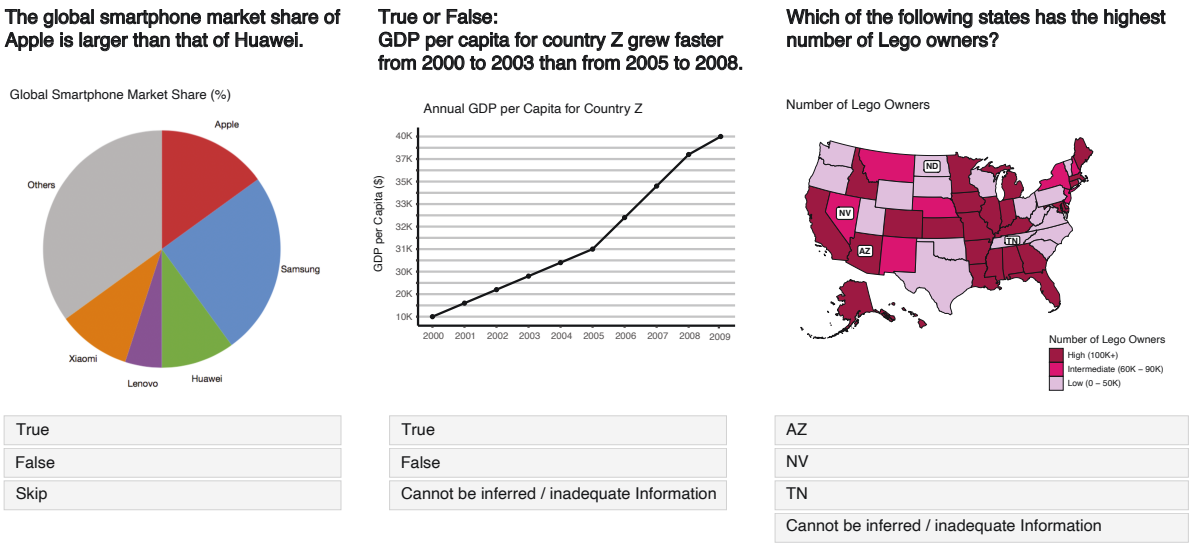

Fig. 2: Example items from VLAT and CALVI: (A) is an item from VLAT; (B) is a trick item from CALVI; (C) is a normal item from CALVI.

test produces consistent and stable results) and **validity** (i.e., ensuring that the test measures the ability that it is supposed to measure) [9].

3 Item Analysis

As outlined in Sec. 2.3, the first stage of CAT development is item analysis. In this section, we describe the IRT model used for item analysis to obtain parameters for both item easiness and discrimination. These item parameters are necessary for constructing the CAT algorithms, to be detailed in the following sections. Additional results from the item parameter analysis are available in supplemental materials.

3.1 2-Parameter IRT Model

IRT mathematically describes the relationship of a test taker's ability, item parameters, and the probability of them answering a particular item correctly. We use the 2-parameter IRT model to infer the item easiness and discrimination parameters of all VLAT and CALVI items. The item response function of the 2-parameter IRT model (Equation (1)) describes the relationship between the probability of a test taker with ability θ answering item i correctly (left-hand side) and the easiness b_i and discrimination a_i of item i (right-hand side). This function will be used in conjunction with test tryout data to compute the item parameters (i.e., a_i and b_i) for VLAT and CALVI items, and it will also be used in Sec. 4 to define and compute other key quantities (e.g., item information, standard error) in adaptive testing.

$$p_i(\theta) = \frac{1}{1+\exp(-a_i(\theta+b_i))} \tag{1}$$

3.2 Parameters of VLAT and CALVI Items

The developers of CALVI collected test tryout data and used the 2-parameter Bayesian IRT model to compute the item parameters of each item [14]. Therefore, we can directly use these item parameters to adaptively select the next item for A-CALVI.

Because item analysis for VLAT was conducted with CTT instead of IRT, we apply IRT analysis on VLAT items using their test tryout data.[1] We use a similar 2-parameter Bayesian IRT model as the one used for CALVI with adjustments to address the following difference between the two tests: CALVI has two types of items: normal items and trick items, and only the trick items are associated with test takers' abilities θ. Whereas in VLAT, a person's performance on every item counts. All other specifications are the same as the model from CALVI, provided in supplemental materials, along with the item parameters of the VLAT and CALVI items.

[1]This data was obtained from the authors of VLAT by personal communication.

4 Construction of A-VLAT and A-CALVI Algorithms

To construct A-VLAT and A-CALVI, we use test tryout data, item parameters, and simulation studies to determine the configurations of A-VLAT and A-CALVI: **(1) initialization** (Sec. 4.2), **(2) item selection** (Sec. 4.3), **(3) scoring** (Sec. 4.4), and **(4) termination** (Sec. 4.5). In addition, we conduct a qualitative study for A-CALVI to determine how to balance items across the misleading / normal categories (Sec. 4.6). Figure 3 presents these components and the flow of our CAT algorithm.

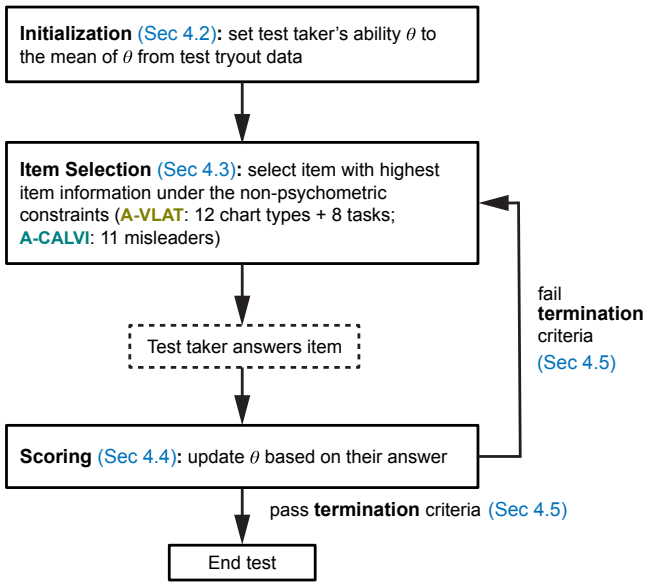

Fig. 3: The components and flow of our CAT algorithm: the algorithm begins by initializing ability θ for a test taker, which is then used to select the next item. The test taker then answers the item, and their θ gets updated based on the answer. If the termination criterion is satisfied, then the test ends; if not, the updated θ is used to select the next item and this process repeats.

4.1 Key Quantities in IRT-based CAT

The idea of CAT is to iteratively give a test taker an item that is best suited to their ability and iteratively update their estimated ability. In IRT-based CAT, an item that is best suited to a test taker with ability θ is the item that provides the most information about that test taker; this quantity is called *item information*. Equation (2) provides the mathematical definition of item information in the 2-parameter model [2]: on the left-hand side, $I_i(\theta)$ is the item information from test taker with ability θ answering item i; on the right-hand side, a_i is the item discrimination parameter of item i, and $p_i(\theta)$ is the probability of person with ability θ answering item i correctly (defined in Eq. (1)).

$$I_i(\theta) = a_i^2 p_i(\theta)(1 - p_i(\theta)) \tag{2}$$

Notice that $I_i(\theta)$ is maximized when $p_i(\theta) = 0.5$. This means that an item provides the most information about a test taker with ability θ when we are most uncertain about whether the test taker will get this item right.[2] Also, the greater the item discrimination a_i (an item's ability to differentiate test takers of different levels of ability), the greater the item information.

Under the context of item information, the item that is best suited to a test taker with ability θ is the item with the highest information for θ in the item bank. This will be used in item selection (Sec. 4.3).

Given the definition of item information, we can also define the *test information* [2] as the sum of the item information of its items:

$$I(\theta) = \sum I_i(\theta) \tag{3}$$

Another use of item information is to quantify the measurement precision of the test. The *standard error* [2] is defined as

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}. \tag{4}$$

The larger the test information, the smaller the standard error, and the more precise the measurement of the test on ability θ. The standard error will be used in the simulation studies to construct a metric for comparing the measurement precision of our adaptive tests with their original static counterparts, and this metric will provide evidence for selecting appropriate lengths for the adaptive tests (Sec. 4.5).

4.2 Initialization

As shown in Fig. 3, the first step of the CAT algorithm is to initialize θ for a test taker, which can then be used for item selection. After examining the test tryout data of both VLAT and CALVI, we find that the distributions of θ are approximately normal for both tests. Thus, most people's θ's are close to the mean of the distribution. Therefore, we initialize the θ of both adaptive tests using the mean of the θ's in respective test tryout data. We explore the implications and possibilities of different methods for initializing θ in the discussion section.

4.3 Item Selection

The next item in A-VLAT or A-CALVI will be the item that yields the highest item information for the test taker's current θ estimate. However, simply selecting without constraints could risk not covering the important aspects of the original test, which is a salient trade-off of shorter adaptive tests. This is referred to as *content balancing* [29, 31].

To ensure content balancing, we impose non-psychometric constraints to the item selection process such that A-VLAT covers all 12 chart types and 8 tasks of VLAT and A-CALVI covers all 11 misleaders of CALVI. We refer to them as *non-psychometric features*.[3]

This is achieved by the following logic: we initialize the list of non-psychometric features to contain all features needed to be covered. If the remaining number of items in the test is **greater than the length of this list**, then the item with the highest information is selected from the bank, and its associated non-psychometric features are removed from the list. If the remaining number of items in the test is **equal to the length of this list**, the highest-information item with at least one of the remaining features is selected, and its associated non-psychometric features are removed from the list.

4.4 Scoring

In adaptive testing, the θ (ability) estimate needs to be updated every time a test taker answers an item (shown in Fig. 3). We initialize θ to a prior distribution based on the test tryout data from participants taking VLAT and CALVI. Then, with each answer, we perform a Bayesian update to adjust the distribution of the θ estimate. We compute the final score of a test taker by taking the mean of the distribution of the θ estimate at the end of this iterative process.

4.5 Termination

We develop a fixed-length CAT for A-VLAT and A-CALVI. While CATs can be variable-length or fixed-length, variable-length tests can be difficult to implement in formal learning spaces with fixed time constraints, and require additional considerations and procedures to develop. We determine the lengths using simulations[4] by comparing the standard error of each person in fixed-length CAT of various lengths to that of the original static version of the test. In Sec. 4.5.1 and Sec. 4.5.2, we present details of the simulation studies resulting in **27 items for A-VLAT** and **11 trick items for A-CALVI**.

Because we select the original static tests as the baseline for both A-VLAT and A-CALVI, we define the following metric:

$$\text{relative difference in } SE(\theta) = \frac{SE_{\mathrm{A}}(\theta) - SE_{\mathrm{O}}(\theta)}{SE_{\mathrm{O}}(\theta)}, \tag{5}$$

where $SE_{\mathrm{A}}(\theta)$ is the standard error for θ in the adaptive test and $SE_{\mathrm{O}}(\theta)$ is the standard error for θ in the original test. This metric is a measure of relative precision of the adaptive test compared to the original version; the smaller relative difference in $SE(\theta)$ is, the more precise the adaptive test compared to the original version. For example, a relative difference in $SE(\theta)$ of 0.30 implies that for this θ, the standard error of the adaptive test is within 30% of the standard error of the baseline. Similarly, a relative difference in $SE(\theta)$ of 0 means that the adaptive and original versions are equally precise. Note that if this quantity is positive, the adaptive test has a larger standard error than the baseline for θ. If the quantity is negative, then the adaptive test has a smaller standard error than the baseline for θ.

4.5.1 Simulation for A-VLAT

In this simulation study, we compare the measurement precision of A-VLAT with VLAT and provide reasoning for why we select 27 items as the length for A-VLAT.

Data The item analysis of VLAT test tryout data on 191 participants yielded θ estimates for all 191 participants. We fit a normal distribution for θ using this data and generate a sample of 500 simulated persons (i.e., θ's). We treat the simulated θ's as the ground truth latent abilities for the simulated persons and refer to them as **true θ's** below.

Method Because our goal is to find a reasonable fixed length for A-VLAT, we build a simulation that realizes the 500 simulated persons' test results under A-VLAT of different lengths from 19 items[5] to 53 items and the original version of VLAT, which has a length of 53. We do so by (1) plugging in the true θ and the item parameters into Eq. (1) to compute the probability of correctness for each simulated person, (2) simulating the binary correctness outcome with that probability, and (3) computing the mean of the posterior distribution of the θ estimate as the final score for each person as described in Sec. 4.4.

From here, we compute the relative difference in $SE(\theta)$ for all 500 simulated persons.

[2] This should not be confused with the test taker being most uncertain about which answer is correct (i.e. randomly picking an option). For instance, if a test taker picks randomly in a 3-option item, $P(\text{correct})$ would be $1/3$ and $P(\text{incorrect})$ would be $2/3$—so we would be more certain that they would answer incorrectly than correctly. Uncertainty is maximized at $1/2$, not $1/3$.

[3] The list of non-psychometric features is provided in supplemental materials.

[4] The code and data for simulations can be found in supplemental materials.

[5] The non-psychometric constraint of A-VLAT dictates that it must cover all 12 chart types and 8 tasks. Since any item is associated with a chart type and a task, A-VLAT needs to have at least 19 items to satisfy the non-psychometric constraint in the worst case.

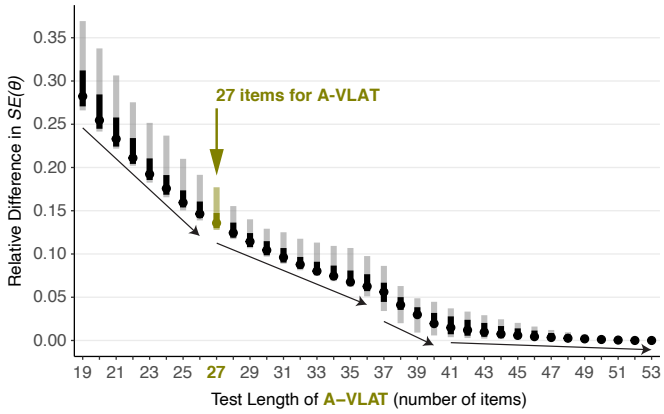

Fig. 4: The distribution of relative difference in $SE(\theta)$ from length 19 to 53 for A-VLAT: the dot is the median, the lighter bar is where 95% of the observations fall, and the darker bar is where 66% of the observations fall. The trend has approximately 4 segments as indicated by the arrows. We select 27 as the length of A-VLAT.

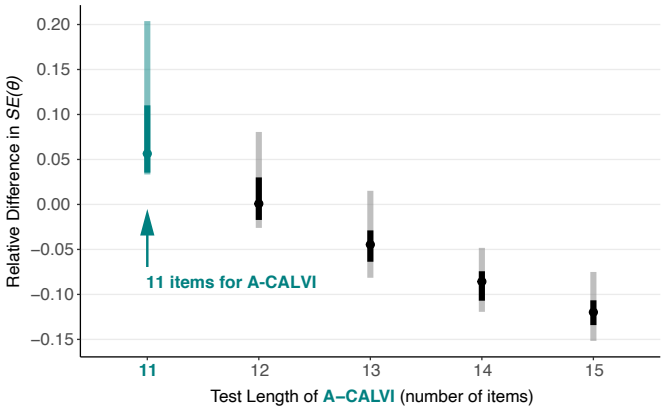

Fig. 5: The distribution of relative difference in $SE(\theta)$ from length 11 to 15 for A-CALVI, which shows a downward trend. We select 11 as the length of A-CALVI.

Results Figure 4 shows the 500 simulated persons' distribution of relative difference in $SE(\theta)$ for A-VLAT of length 19 to 53. Upon inspecting the trend, we find that this chart has approximately four segments: relative difference in $SE(\theta)$ experienced the sharpest drop from length 19 to length 26, a milder decrease from 27 to 36, another steep yet short downfall from 37 to 40, and finally plateaus after 41. Note that it never goes below 0 because the baseline uses all 53 items, and adaptively selecting 53 items would yield the same result.

After examining Fig. 4, we select 27 as the length of A-VLAT. This reduces the length of the original test by half, and the median and most observations of relative difference in $SE(\theta)$ are below 0.15, a reasonable loss of measurement precision in exchange for halving the test. It should be noted that there is no gold standard for selecting the perfect cutoff in the literature. Test administrators can use Fig. 4 to help customize the length based on their needs and constraints. We will discuss customization of our adaptive tests in Sec. 7.3.

4.5.2 Simulation for A-CALVI

In this simulation study, we compare the measurement precision of A-CALVI with CALVI and provide reasoning for why we select 11 as the number of trick items for A-CALVI.

Data Similar to the A-VLAT simulation, we used the θ estimates for all 497 participants from CALVI test tryout to generate a sample of 500 simulated persons. We treat all simulated θ's as the ground truth latent abilities for the simulated persons.

Method The same method in Sec. 4.5.1 was applied. we obtained the relative difference in $SE(\theta)$ for all simulated persons.[6]

Results Figure 5 shows a downward trend. From length of 13, the median and most observations of relative difference in $SE(\theta)$ fall below 0, indicating that A-CALVI outperforms the original CALVI in terms of measurement precision; this is because A-CALVI selects better items from the bank of 45 than the fixed 15 items in CALVI. We select 11 as the length of A-CALVI because it is shortest length that can cover all non-psychomtric features and the median and most of the observations of relative difference in $SE(\theta)$ at length 11 are below 0.10.

4.6 Composition of A-CALVI Items via Qualitative Analysis

The developers of CALVI included 15 items with well-constructed non-misleading visualizations (i.e. normal items) to accompany 15 items with misleading visualizations (i.e. trick items). This is because (1) if a test taker realizes that all items are about misleading visualizations, they could adopt the strategy of never selecting the obviously correct answers and (2) when people encounter visualizations in the real world, they see both normal, well-constructed and misleading visualizations, which makes distinguishing misleading ones difficult [14]. However, a test taker's performance on the 15 normal items does not affect the θ estimate in CALVI, and the normal items constitute half of the test, making it long and time-consuming. Therefore, we conduct a qualitative study to investigate the effect of normal items and whether they would be necessary in A-CALVI.

We recruited two groups of participants from **Prolific** for two conditions: (1) in *without-normal condition*, participants took A-CALVI without normal items (i.e., only 11 adaptively-selected trick items from the bank of 45 trick items); (2) in *with-normal condition*, they took A-CALVI with normal items (i.e., 11 adaptively-selected trick items + 11 fixed normal items). The 11 normal items were randomly sampled from the set of 15 normal items in the item bank of CALVI.

After participants in both conditions completed the items on the test, we asked them to answer a post-test survey of Yes/No and open-ended questions[7] to investigate the possible impact of the presence of normal items. These questions included whether they noticed items with misleading charts (i.e., trick items) and how they approached the rest of the test after noticing trick items. We also provided the participants with the items they saw and their responses (presented in the order in their test) to help them recall the test content. In the *with-normal* condition, we also explored how participants approached the rest of the test after noticing the presence of normal items.

Participants A total of 60 participants were recruited from **Prolific** for this qualitative study, with 30 for the *without-normal condition* and 30 for the *with-normal condition*. The data of one participant in the *with-normal condition* was discarded due to a data collection error, leaving a total of 59 participants. All participants fall under the following criteria: speak fluent English, have normal or corrected-to-normal vision, are between 18 and 65 years old, and are U.S. residents. Participants who failed more than half of the attention check questions were excluded, and all participants passed the attention check in this study. In addition, **Prolific** participants from similar studies that the authors conducted in the past were excluded.

Procedure We started by presenting the consent form and information page describing the structure of the study to the participants. We instructed participants that they need to select an answer for the current item/question before proceeding to the next one, and once they moved on, they could not return to previous ones. After the participants completed the test items, they were directed to the post-test survey.

Method We reviewed and coded participants' responses on the post-test survey to understand their experience with the test and how trick and normal items affected their approach to the test.

Results In the *without-normal condition*, 24 out of 30 participants reported they noticed the presence of trick items, after which 16 paid

[6]Similar to A-VLAT, A-CALVI needs at least 11 items to cover all 11 misleaders because each item in the bank is associated to only one misleader. The original CALVI has 15 trick items, so we selected 15 as an upper bound.

[7]The open-ended questions can be found in supplemental materials.

more attention or became more skeptical of the items, 6 did not change how they approached the rest of the items, and 2 gave unclear responses.

In the *with-normal condition*, 19 out of 29 participants reported that they noticed the presence of trick items, after which 14 of them paid more attention or double checked their answers, 4 did not change how they approached the rest of the items, and 1 participant gave an unclear response. When asked whether the presence of normal items changed their approach to answer trick items, 16 out of the 19 participants said no, and the remaining 3 participants mentioned that it only made them more suspicious of the normal items.

There seems to be little evidence that the presence of normal items affected how participants approached the trick items, but we acknowledge that this sample size is small and the possibility of normal items having an effect still exists. Thus, we refrain from completely removing normal items, but instead look to reduce the amount.

Our underlying goal is the same as CALVI—the test taker sees a mix of trick and normal items, and we reason about this from the the test takers' perspective. Based on the analysis of trick items in the original CALVI, there are certain misleaders that are substantially harder to identify, such as *Overplotting* or *Missing Data* [14], meaning these items are essentially perceived as "normal" in the eye of the test taker. Therefore, the items from such misleader categories can be effectively considered as "normal" items by the test takers (i.e., perceived "normal"). We also notice that it is highly likely that the option of "Cannot be inferred/inadequate information" is the correct answer when it appears in a trick item in A-CALVI. To prevent the strategy of always selecting this option when it appears, we decide to include in A-CALVI all of the normal items from CALVI that contain this option but it is not the correct answer: there are 4 such items. With the 4 *actual* normal items and the *perceived* "normal" items, we expect that, from the test taker's perspective, there would be a balanced mix of trick and "normal" items, following a similar format as the original CALVI. Thus, A-CALVI consists of 11 trick items selected adaptively by the CAT algorithm and 4 *actual* normal items. The total length of A-CALVI is 15, half of the length of the original CALVI.

4.7 Final Configurations of A-VLAT and A-CALVI

A-VLAT The final configuration initializes the first item by using the mean of θ's from the VLAT test tryout data to select the item with highest information in the bank. After the test taker answers each item, that item is removed from the bank, and θ is updated accordingly, which is then used to select the next item. Each time, A-VLAT selects the next item with the highest information under the constraint that the 12 chart types and 8 tasks must be covered before the test terminates. A-VLAT terminates after a test taker has answered 27 items.

A-CALVI The final configuration initializes the first item using the mean of θ's from the CALVI test tryout data to select the item with highest information in the bank. After the test taker answers each item, that item is removed from the bank, and θ is updated accordingly, which is then used to select the next item. Each time, A-CALVI selects the next item with the highest information under the constraint that the 11 misleaders must be covered before the test terminates. The 4 normal items are positioned at 4 randomly pre-determined locations among the trick items, but which normal item appears at which of the 4 positions is randomly determined at runtime for each test taker. A-CALVI terminates after a test taker has answered 11 trick items and 4 normal items.

With the configurations of A-VLAT and A-CALVI finalized, it is necessary to know whether the two tests can provide stable and consistent measurements and whether they measure the abilities they are designed to measure. Therefore, we conduct four online studies to evaluate the reliability (Sec. 5) and validity (Sec. 6) of the adaptive tests.

5 Reliability Evaluation

Reliability of a test is whether it can provide consistent and stable measurement. We provide evidence of **test-retest reliability**, a common way of assessing reliability and used in various domains to evaluate the reliability of CATs (e.g., [3, 20]).

Test-retest reliability evidence is gathered by administering a test to the same group of people twice with a time interval between the first and the second administrations and comparing participant's performances on each attempt. An interval of 1 or 2 weeks is typical [26]. For both A-VLAT and A-CALVI, we have the same study design: we recruit one group of participants from `Prolific`. Everyone takes the adaptive test and is invited back to take it again after a week.

We use the **intraclass correlation coefficient (ICC)** as the reliability metric recommended by literature [26]. Next, we describe the models to compute ICC for A-VLAT and A-CALVI and present the results.

5.1 Model for Reliability Metric

ICC can be computed via a random effects model, and we used a Bayesian measurement error version of the model because the measurement for each person is a distribution rather than a point estimate. It is important to take into account the measurement error when computing ICC, as otherwise it will be underestimated [33]. Our model is preregistered on OSF (see link in abstract) and can be found in supplemental materials.[8] The model is:

$$\begin{aligned}\theta^* &\sim \text{Normal}(\mu + \alpha_j, \sigma_\varepsilon)\\ \theta &\sim \text{Normal}(\theta^*, \text{se}(\theta))\\ \alpha_j &\sim \text{Normal}(0, \sigma_\alpha),\end{aligned}$$

where θ^* is the latent ability, μ is the unobserved mean, α_j is the random effect for person j, σ_ε is the standard deviation within-person, σ_α is the standard deviation between people, θ is the observed ability, and $\text{se}(\theta)$ is the measurement error of θ. For the priors of this model, we selected $\text{Normal}(0,1)$ and $\text{Normal}(-1,1)$ for μ for A-VLAT and A-CALVI, respectively, because they are approximately the distributions of the estimated abilities in the VLAT and CALVI test tryout data. We also selected $\text{Normal}^+(0,1)$ for σ_α and σ_ε.

For both the reliability of A-VLAT and A-CALVI, we ran our model with 4 chains, each with 20,000 iterations. We discarded 10,000 warmup iterations per chain and thinned the final sample by 5, yielding 8,000 total post-warmup draws.

This model outputs distributions of σ_α and σ_ε, which we then use to compute ICC with the following formula [23]:

$$\text{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}.$$

ICC can be understood as the ratio between the variance between people (σ_α^2) and the sum of variance between people and the (undesirable) within-person variance (σ_ε^2) [23].

5.2 Reliability of A-VLAT

Participants 90 participants were recruited in anticipation that around 60 would return to the second study so that we have a sample size close to 60. This sample size was determined by collecting pilot data with 20 participants, fitting the model with pilot data, generating simulated participants with the fitted model, and choosing the sample size that would yield a ±0.1 95% credible interval around the ICC.[9]

In the end, 47 participants completed both parts of the study. All participants fall under the same criteria listed in Sec. 4.6.

Model Diagnostics The minimal bulk effective and the minimum tail effective sample sizes are 6,757 and 7,067, and the $\hat{R}$ values are approximately 1.00 for all parameters.

Results Figure 6 shows the posterior density of ICC and a scatterplot of participants' scores on the first attempt versus those on the second attempt. **The median ICC is 0.98 with 95% CI of [0.87, 1.00]**, so A-VLAT has **excellent test-retest reliability** [8, 24].

5.3 Reliability of A-CALVI

Participants We determined the sample size by first following the same method in Sec. 5.2. However, in the model with measurement

[8] We initially pre-registered an implementation of this model that is difficult to fit and computes extra parameters (the latent θ^*s) that we do not need. We later found an alternative parameterization better suited for our purpose, so the results in this section are from the alternative parameterization.

[9] All pilot analyses for sample size determination are in supplemental materials.

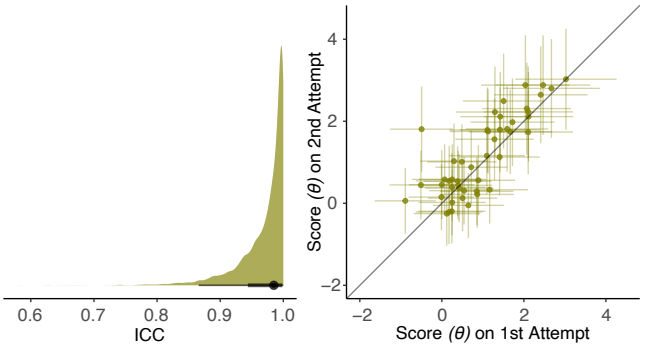

Fig. 6: Reliability of A-VLAT: the left plot is a posterior density plot of ICC with a point interval plot at the bottom; the point is the median, and the lighter and darker bars are the 95% and 66% intervals. The right plot is a scatterplot of the 47 participants' scores on the first and second attempts, where each point represents a participant, and the cross lines at each point represents the measurement error (95% intervals).

error on pilot data, the intervals of ICC did not shrink as the sample size of the simulated participants increased. Therefore, we use a sample size of 60 because it is consistent with the sample size of the reliability study for A-VLAT and the 95% intervals on σ_α and σ_ε have similar width as those in Sec. 5.2. In addition, we report the ICC results from models with and without measurement error because the interval of ICC may be large for the model with measurement error. Based on the attrition rate of the previous study, we recruited 115 participants, in anticipation that around 60 would return to the second study. In the end, 73 participants from the first part of the study returned to the second part. All participants fall under the same criteria listed in Sec. 4.6.

Model Diagnostics The minimal bulk effective and the minimum tail effective sample sizes are 7,272 and 7,560, and the $\hat{R}$ values are approximately 1.00 for all parameters.

Results Figure 7 shows the posterior density of ICC and a scatterplot of participants' scores on the first attempt versus those on the second attempt. **The median ICC is 0.98 with 95% CI of [0.82, 1.00]**, so A-CALVI has **excellent test-retest reliability** [8, 24]. Even though the model with measurement error is able to produce a small interval for ICC with data from this study, we report the result from the other model as stated in pre-registration: the median ICC is 0.77 with 95% CI of [0.66, 0.85]. We select the ICC from the model with measurement error as the final report for reliability because a model that does not account for measurement error consistently underestimates ICC. [33].

6 Validity Evaluation

We evaluate whether our adaptive tests are valid—that they measure what they are designed to measure—by computing the **convergent validity** [7]: the correlation between test takers' performance on the adaptive tests and their performance on the original static tests. This is an appropriate measure of the validity of the adaptive tests [19] because VLAT and CALVI are both valid measurements [14, 22] and the equivalence of the original and adaptive tests through strong correlation would demonstrate that the adaptive tests indeed measure the same abilities as the originals, hence are valid.

To compute the correlation coefficient for convergent validity, the same group of participants need to take both versions of the tests. Because retaking the tests may create learning and ordering effects, we use two groups of participants to counterbalance the effects. For both A-VLAT and A-CALVI, we have the same study design with two groups, both recruited from `Prolific`: each person takes both tests, with a week[10] in between sessions. Order is counterbalanced: half the participants (Group 1) takes the original test first, and the other half (Group 2) takes the adaptive test first. The procedure of our validity studies is the same as that in Sec. 4.6.

[10]Because of the similarity of content between the adaptive and original tests, we choose to have a one-week interval to mitigate learning effects.

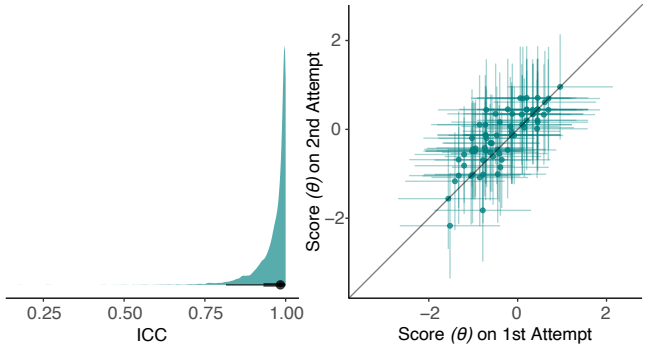

Fig. 7: Reliability of A-CALVI: the left plot is a posterior density plot of ICC; the right plot is a scatterplot of the 73 participants' scores on the first and second attempts with measurement error.

6.1 Model for Validity Metric

Correlation coefficients can be computed via a Bayesian bivariate Normal model. We again use a measurement error model. The model is preregistered on OSF (see link in abstract) and can be found in supplemental materials.[11] The bivariate Normal model is:

$$\begin{bmatrix}\theta_0^* \\ \theta_A^*\end{bmatrix} \sim \text{Normal}_2\left(\begin{bmatrix}\mu_0 \\ \mu_A\end{bmatrix}, \sigma_0, \sigma_A, \rho\right)$$
$$\theta_0 \sim \text{Normal}(\theta_0^*, \text{se}(\theta_0))$$
$$\theta_A \sim \text{Normal}(\theta_A^*, \text{se}(\theta_A)).$$

where θ_0^* and θ_A^* are the latent abilities measured from the original test and the adaptive test, with μ_0 and μ_A as their means, σ_0 and σ_A as their standard deviations, and ρ as the correlation between them. θ_0 and θ_A are the observed abilities, with measurement errors $\text{se}(\theta_0)$ and $\text{se}(\theta_A)$. In the implementation of this model, we re-parameterized μ_0 and μ_A in terms of the difference between them by subtracting the mean of θ_0 from all the observed θ's.

For the prior of the model, we selected $\text{Normal}(0,1)$ for the difference between θ_0 and θ_A, as 1 is a large difference in latent IRT ability. We picked $\text{Normal}(1, 0.5)$ for σ_0 and σ_A, because standard deviations should center around 1 as the IRT model is parametrized so that latent abilities have a standard deviation around 1. For ρ, we selected a uniform prior (the LKJ distribution with parameter $\eta = 1$).[12]

We ran our final model with 4 chains, each with 20,000 iterations. We discarded 10,000 warmup iterations per chain and thinned the final sample by 5, yielding 8,000 total post-warmup draws.

This model outputs the distribution of the correlation coefficient ρ for convergent validity.

6.2 Validity of A-VLAT

Participants 120 participants were recruited (60 for each group) in anticipation around 100 would return to the second part of the study so that we have a sample size close to 100. This sample size was determined by collecting pilot data with 10 participants in each group, fitting the model with pilot data, generating simulated participants with the fitted model, and choosing the sample size that would yield a ±0.1 95% credible interval around ρ.

In the second part of the study, 49 returned in Group 1 and 37 returned in Group 2, for a total of 86 participants. All participants fall under the same criteria listed in Sec. 4.6.

Model Diagnostics The minimal bulk effective and the minimum tail effective sample sizes are 7,607 and 7,394, and the $\hat{R}$ values are approximately 1.00 for all parameters.

[11]The model results in this section are based on an alternative implementation different than the initially pre-registered for the same reason in Sec. 5.1.

[12]We pre-registered LKJ distribution with parameter $\eta = 3$ as the prior of ρ in the initial implementation of the model for A-VLAT, because it needed a tight prior to converge. However, the alternative implementation—which does not estimate all θ^*s—converges much more easily, so we relaxed the prior on ρ.

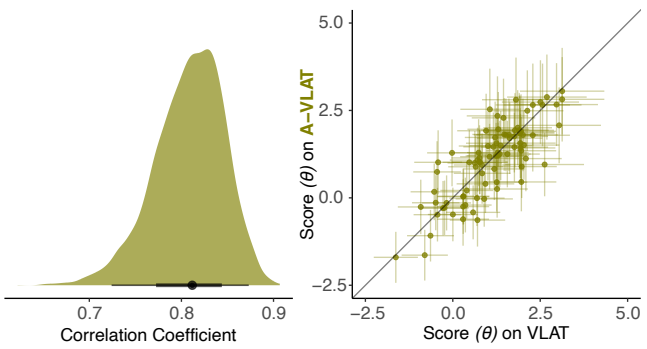

Fig. 8: Validity of A-VLAT: the left plot is a posterior density plot of correlation coefficient ρ; the right plot is a scatterplot of the 86 participants' scores on VLAT and A-VLAT with measurement error.

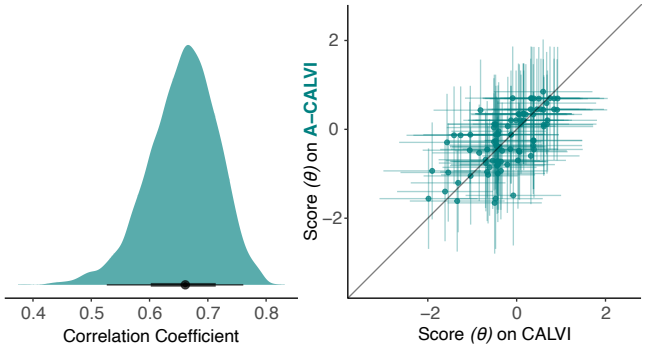

Fig. 9: Validity of A-CALVI: the left plot is a posterior density plot of the correlation coefficient ρ; the right plot is a scatterplot of the 90 participants' scores on CALVI and A-CALVI with measurement error.

Results Figure 8 shows the posterior density of ρ and a scatterplot of participants' scores on the adaptive test versus those on the original test. **The median ρ is 0.81 with 95% CI of [0.72, 0.87]**. While there is no gold standard for convergent validity, a correlation coefficient below 0.50 should be avoided and above 0.70 is recommended [6], so A-VLAT has **high convergent validity**.

6.3 Validity of A-CALVI

Participants A sample size of 100 was determined in the same way as in Sec. 6.2. Based on the attrition rate of the previous study, 140 participants were recruited (70 in each group), in anticipation that around 100 participants would return to the second part of the study.

In the first part of the study, one participant's data was discarded because they failed the attention check questions. In the second part, 47 and 43 returned in Groups 1 and 2, for a total of 90 participants. All participants fall under the same criteria listed in Sec. 4.6.

Model Diagnostics The minimal bulk effective and the minimum tail effective sample sizes are 6,455 and 5,090, and the $\hat{R}$ values are approximately 1.00 for all parameters.

Results Figure 9 shows the posterior density of ρ and a scatterplot of participants' scores on the adaptive test versus those on the original test. **The median ρ is 0.66 with 95% CI of [0.53, 0.76]**. Since a correlation coefficient below 0.50 should be avoided and above 0.70 is recommended [6], A-CALVI has **acceptable convergent validity**.

7 Discussion

7.1 How Reliable, Valid, and Precise Are CATs Really?

The results from Sec. 5 and Sec. 6 have shown that we have successfully created two reliable and valid adaptive tests, A-VLAT and A-CALVI, which measure the same abilities as the original static VLAT and CALVI. In addition, we collected data that can be used to demonstrate the measurement precision of both adaptive tests: in validity evaluation, every participant was asked to take both the adaptive and original versions. We can use this data to analyze the relative difference in $SE(\theta)$ of both adaptive tests as we did in Sec. 4.5 with simulation data.[13]

For A-VLAT, the distribution of relative difference in $SE(\theta)$ for 86 participants has a mean of 0.089 (i.e., for a person with average ability in this group, the standard error for them is within 8.9% of the standard error of the baseline.) with a standard deviation of 0.11.

For A-CALVI, the distribution of relative difference in $SE(\theta)$ for 90 participants has a mean of 0.023 with a standard deviation of 0.039.

We observe that results of relative difference in $SE(\theta)$ from actual participants reveal that both adaptive tests are similarly precise compared to their original static counterparts, in accordance with the simulation results in Sec. 4.5.1 and Sec. 4.5.2. This demonstrates the measurement precision of A-VLAT and A-CALVI and further supports the justification for the shortened lengths of the adaptive tests.

[13]The code for this analysis is provided in supplemental materials.

Another consideration of reliability is whether a test taker can recover quickly if they mistakenly select a wrong answer. First, we define d_i as the absolute value of the difference between the mean of the θ estimate distribution and the true θ of a test taker after answering the i-th item in the test. Then, "mistakenly selecting a wrong answer for item i" is defined as when $d_i > d_{i-1}$; i.e., when their θ estimate moves away from the true θ after answering an item. Naturally, "recovery" from choosing a wrong answer for item i happens at the first step i' where $d_{i'} \leq d_i$; i.e., their θ estimate moves back to being within at least d_i of their true θ, and recovery length is $i' - i$. We conducted simulations[14] to study the recovery lengths in our adaptive tests. We generated 500 simulated participants for each test as before, and simulated their response correctness to A-VLAT and A-CALVI. The median recovery lengths for A-VLAT and A-CALVI are 3 and 2 with standard deviations of 4.8 and 1.1, respectively. Therefore, test takers can recover quickly from a mistake in our tests, demonstrating good reliability.

7.2 Correlation within Chart Types, Tasks, and Misleaders

Items in the banks of VLAT and CALVI span a range of chart types, tasks, and misleaders. Studying the relationship between people's performance within these different categories of visualization features can shed light on how different aspects and skills of visualization literacy connect and transfer, which can then help us better understand this ability and potentially further optimize adaptive testing.

Actual participants have been asked to answer these visualization literacy items from prior work [14, 22], as well as the qualitative study, reliability evaluation, and validity evaluation in this paper. This gives us the opportunity to use this data to examine the relationship between participants' performance on different chart types, tasks, and misleaders. Therefore, we conducted an exploratory study to compute the correlations within these visualization features by using both test tryout data (VLAT and CALVI) and data collected in this paper and employing the same IRT models described in Sec. 3.[15] Next, we present some highlights from this exploration and discuss their implications.

VLAT Chart Types We found high correlations between performance on certain chart types: 0.83 with 95% CI of [0.72, 0.92] for scatterplot and bubble chart, 0.82 with 95% CI of [0.72, 0.90] for 100% stacked bar and stacked bar charts, and 0.81 with 95% CI of [0.69, 0.90] for scatterplot and line chart. There are also chart types without high correlations with others: choropleth map's and stacked area chart's correlation with other chart types do not exceed 0.38 and 0.52, respectively.

VLAT Tasks Many pairs of tasks have high correlations: 0.93 with 95% CI of [0.88, 0.97] for *retrieve value* and *make comparisons*, 0.89 with 95% CI of [0.81, 0.96] for *find extremum* and *make comparisons*, 0.79 with 95% CI of [0.70, 0.86] for *determine range* and *retrieve value*. Most tasks highly correlate with at least another task.

[14]The code for the simulations is provided in supplemental materials.
[15]The entire results can be found in supplemental materials.

CALVI Misleaders Most pairs of misleaders have weak or moderate positive correlations: *Manipulation of Scales - Inappropriate Scale Range* and *Manipulation of Scales - Inappropriate Use of Scale Functions* have a correlation of 0.75 with 95% CI of [0.58, 0.89], highest among all pairs of misleaders. Some misleaders have very weak correlations with others; *Overplotting* especially stands out as its correlations with the other misleaders range from -0.21 to 0.34. This is reasonable as *Overplotting* is one of the most difficult misleaders to judge correctly.

Although A-VLAT and A-CALVI are significantly shorter than their original versions, there are opportunities to further shorten them by leveraging these correlations. For instance, the VLAT tasks *retrieve value* and *make comparisons* are highly correlated, so test developers could make the item selection process less constrained by only covering one of the two instead of both. Further research on incorporating these correlations into CAT algorithms for visualization literacy is needed.

Moreover, these correlations are valuable quantitative data that informs taxonomy in visualization literacy: people's performance on tasks *retrieve value*, *find extremum*, and *make comparisons* are highly correlated, which means it may be reasonable to categorize them into one task. Additionally, when researchers and educators design curricula or tools to improve people's visualization literacy, they can draw insights from these correlations to create more efficient and effective interventions: since people's performances on answering items with scatterplots and bubble charts are strongly correlated, teachers may choose to focus on one, or group lessons together for increased efficiency.

7.3 Customization of Adaptive Tests

When using A-VLAT and A-CALVI, test administrators can customize many components of the tests, including initialization, item selection, and termination criterion. For initialization, we chose the means of the θ distributions from test tryouts. However, test administrators can explore other alternatives, such as randomly choosing the first item from the bank to decrease the item exposure (i.e., the frequency at which an item is administered across all administrations) of certain items. Test administrators can also customize the non-psychometric constraints in item selection. For example, if one wishes to focus on only covering all VLAT chart types or only covering all tasks, they could adjust the non-psychometric constraints to achieve that. Moreover, the results from the simulation studies in Sec. 4.5 can guide administrators to adjust the length of the adaptive test based on their time-limit constraints and requirement for measurement precision. For scoring, we recommend reporting the mean of the posterior distribution of θ as a point estimate (described in Sec. 4.4), as well as the raw correctness (i.e., number of items answered correctly over the total number of items answered) to help test takers intuitively interpret their final scores.

Test administrators may wonder about the reliability and validity of their customized tests. We think that validity would remain relatively stable as the lengths of adaptive tests change so long as the non-psychometric features are covered, because these features ensure that the adaptive and original versions are qualitatively similar. Also, we think that the more non-psychometric features not covered, the less valid the adaptive tests would likely be. However, if two features are very similar, as discussed in Sec. 7.2, not covering one may not have a noticeable effect on validity. In addition, test-retest reliability could vary if the length is too short or long. On the one hand, if too short, the θ estimates would be more susceptible to measurement error and have limited variability, which can affect reliability. On the other hand, if a test is too long, test takers' motivation, attention, and effort might change on the second attempt, causing the test-retest reliability to decrease. Although test length can affect reliability and validity in extreme cases, we believe that moderately changing the length of A-VLAT or A-CALVI would have a marginal effect on reliability and validity so long as the non-psychometric constraints are not violated.

For test administrators who only need to measure the visualization literacy of a group of people once, they can use A-VLAT and A-CALVI as alternatives to their original versions because they are much shorter, contain balanced content, and have similar measurement precision. For researchers and educators who want to conduct pre-post testing to evaluate the effectiveness of their interventions by administering a visualization literacy test once before and once after the intervention, we recommend using our adaptive tests for both administrations. For the second administration, they can choose to exclude certain overused items in the first attempt if they are concerned about item exposure. They could also use the estimated θ of a particular person in the first attempt to initialize this person's first item in the second attempt.

7.4 Limitation and Applicability in Practice

To create assessments under the IRT framework, it is necessary to construct items, run a test tryout study, and compute item parameters using IRT models. This process requires substantial resources as well as technical expertise on IRT modeling, which can be a barrier to implementing adaptive tests in practice. Moreover, adaptive assessments require more complex infrastructure to deliver than a standard series of questions and it is not trivial to add newly designed test items, due to the adaptive nature requiring computation at each step. However, given the benefits of CATs, there is a need to democratize community access to capabilities for test development and administration. Both A-VLAT and A-CALVI are therefore implemented using an experimental framework, designed to create flexible and reusable visualization assessments. One key feature is the decoupling of static and dynamic experiment/assessment content, which supports the implementation of CATs in a more generalized fashion. Consequently, A-VLAT and A-CALVI assessments are designed such that they can be reshaped by intermediate programmers, making it easier for researchers and practitioners to adapt these assessments for specific needs. As more assessments are developed in the visualization community, such infrastructure could play a role in developing adaptive versions of future tests, while also providing helpful validity, reliability, and general test analysis capabilities. Code is available in the paper's supplemental materials.

7.5 Automated Item Generation

Although A-VLAT and A-CALVI are half the lengths of their original versions, the size of the item banks are still limited, meaning that repeated administrations of the tests would increase the item exposure quickly. Larger banks can reduce item exposure and allow for more attempts of re-assessments without giving out the same items. However, generating large banks is costly, both in terms of time and resources. The current process is not scalable because it requires manually creating the visualizations and writing the question text for each item, as well as recruiting a large number of participants for test tryout in order to estimate the item parameters needed for adaptive testing.

There is an opportunity for automating the item generation process and inferring item parameters without recruiting a large number of actual participants: the items in the banks are made up of specific components that are common in all items, such as the visualizations, question text structure, and the answer options. One can potentially leverage visualization grammars and machine learning techniques to automatically generate items and infer their item parameters using substantial existing data. Future studies can investigate ways to scale and improve the item generation process.

8 Conclusion

To more efficiently measure visualization literacy, we develop A-VLAT and A-CALVI, which are adaptive versions of VLAT and CALVI that are half the length of the original tests. We conduct a qualitative study to improve the question composition of CALVI, and incorporate non-psychometric features of both assessments as constraints, ensuring that the adaptive tests cover all tasks and chart types (for VLAT) or misleaders (for CALVI) from the original assessments. We establish the reliability and validity of these adaptive tests via four online studies and associated analyses. We demonstrate how to apply CAT to create more efficient visualization literacy assessments that are reliable, valid, and precise. As visualization literacy can influence data-driven decisions, both A-VLAT and A-CALVI can be used to support the study of visualization literacy development and intervention evaluations through pre- and post-testing. Adaptive assessments can transform measurement practice in visualization literacy, and our work establishes a channel to support the adoption of more efficient assessments in practice.

Acknowledgments

We extend our gratitude to William Revelle and Elizabeth Tipton for their early-stage feedback. We thank Mandi Cai, Maryam Hedayati, and members of the MU Collective at Northwestern for their support and feedback. We are grateful for the Design Cluster program at the Center for HCI and Design at Northwestern, especially Haoqi Zhang, Kapil Garg, and Eleanor O'Rourke for their mentorship. We thank Bum Chul Kwon and other developers of VLAT for sharing their data. We are grateful for Matthew vonAllmen and others who volunteered their time in testing our system before the launch of our experiments. We also thank the online participants for their time and the reviewers for their insightful comments. This work was supported in part by grants from the National Science Foundation (#1815587, #2120750, #2127309 to the Computing Research Association for the CIFellows Project).

Supplemental Materials

All supplemental materials are available on OSF at `https://osf.io/a6258/`.

References

[1] K. Aoyama and M. Stephens. Graph interpretation aspects of statistical literacy: A Japanese perspective. *Mathematics Education Research Journal*, 15(3):207–225, 2003. doi: 10.1007/BF03217380 2

[2] F. B. Baker. *The Basics of Item Response Theory*. ERIC, 2001. 4

[3] D. Beiser, M. Vu, and R. Gibbons. Test-Retest Reliability of a Computerized Adaptive Depression Screener. *Psychiatric Services*, 67(9):1039–1041, 2016. PMID: 27079989. doi: 10.1176/appi.ps.201500304 6

[4] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete. A Principled Way of Assessing Visualization Literacy. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1963–1972, 2014. doi: 10.1109/TVCG.2014.2346984 1, 2

[5] K. Börner, A. Bueckle, and M. Ginda. Data Visualization Literacy: Definitions, Conceptual Frameworks, Exercises, and Assessments. *Proceedings of the National Academy of Sciences*, 116(6):1857–1864, 2019. doi: 10.1073/pnas.1807180116 2

[6] K. D. Carlson and A. O. Herdman. Understanding the Impact of Convergent Validity on Research Results. *Organizational Research Methods*, 15(1):17–32, 2012. doi: 10.1177/1094428110392383 8

[7] C.-L. Chin and G. Yao. Convergent validity. In *Encyclopedia of Quality of Life and Well-Being Research*, pp. 1275–1276. Springer, Dordrecht, 2014. doi: 10.1007/978-94-007-0753-5_573 7

[8] D. V. Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290, 1994. doi: 10.1037/1040-3590.6.4.284 6, 7

[9] R. J. Cohen, W. J. Schneider, and R. Tobin. *Psychological Testing and Assessment*. McGraw Hill LLC, New York, NY, 10th ed., 2022. 2, 3

[10] CollegeBoard. Transitioning to a digital SAT, 2022. `https://professionals.collegeboard.org/pdf/digital-sat-faculty-guide-v3-ada-v0.pdf`. 1, 2

[11] R. DelMas, J. Garfield, and A. Ooms. Using Assessment Items to Study Students' Difficulty Reading and Interpreting Graphical Representations of distributions. In *Fourth Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-4)*. Auckland, New Zealand, 2005. 2

[12] ETS. GRE general test structure, 2023. `https://www.ets.org/gre/test-takers/general-test/prepare/test-structure.html`. 1, 2

[13] B. Foorman, A. Espinosa, C. Wood, and Y.-C. Wu. Using Computer-Adaptive Assessments of Literacy to Monitor the Progress of English Learner Students. *Regional Educational Laboratory Southeast (REL 2016-149)*, 2016. 2

[14] L. W. Ge, Y. Cui, and M. Kay. CALVI: Critical Thinking Assessment for Literacy in Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI'23, pp. 1–18. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3581406 1, 2, 3, 5, 6, 7, 8

[15] F. B. Ghio, M. Bruzzone, L. Rojas-Torres, and M. Cupani. Preliminary Development of an Item Bank and an Adaptive Test in Mathematical Knowledge for University Students. *European Journal of Science and Mathematics Education*, 10(3):352–365, 2022. doi: 10.30935/scimath/11968 2

[16] C. Gibbons, P. Bower, K. Lovell, J. Valderas, and S. Skevington. Electronic Quality of Life Assessment Using Computer-Adaptive Testing. *J Med Internet Res*, 18(9):e240, 2016. doi: 10.2196/jmir.6053 1, 2

[17] R. D. Gibbons and F. V. deGruy. Without Wasting a Word: Extreme Improvements in Efficiency and Accuracy Using Computerized Adaptive Testing for Mental Health Disorders (CAT-MH). *Current Psychiatry Reports*, 21(67), 2019. doi: 10.1007/s11920-019-1053-9 1, 2

[18] R. D. Gibbons, D. J. Weiss, P. A. Pilkonis, E. Frank, T. Moore, J. B. Kim, and D. J. Kupfer. Development of the CAT-ANX: A Computerized Adaptive Test for Anxiety. *American Journal of Psychiatry*, 171(2):187–194, 2014. doi: 10.1176/appi.ajp.2013.13020178 1, 2

[19] B. F. Green, R. D. Bock, L. G. Humphreys, R. L. Linn, and M. D. Reckase. Technical Guidelines for Assessing Computerized Adaptive Tests. *Journal of Educational measurement*, 21(4):347–360, 1984. doi: 10.1111/j.1745-3984.1984.tb01039.x 7

[20] D. Guinart, R. de Filippis, S. Rosson, B. Patil, L. Prizgint, N. Talasazan, H. Meltzer, J. M. Kane, and R. D. Gibbons. Development and Validation of a Computerized Adaptive Assessment Tool for Discrimination and Measurement of Psychotic Symptoms. *Schizophrenia Bulletin*, 47(3):644–652, 2021. doi: 10.1093/schbul/sbaa168 6

[21] S. Kandula, J. S. Ancker, D. R. Kaufman, L. M. Currie, and Q. Zeng-Treitler. A New Adaptive Testing Algorithm for Shortening Health Literacy Assessments. *BMC Medical Informatics and Decision Making*, 11(52), 2011. doi: 10.1186/1472-6947-11-52 2

[22] S. Lee, S.-H. Kim, and B. C. Kwon. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):551–560, 2017. doi: 10.1109/TVCG.2016.2598920 1, 2, 7, 8

[23] D. Liljequist, B. Elfving, and K. Skavberg Roaldsen. Intraclass correlation – A discussion and demonstration of basic features. *PLoS ONE*, 14(7):e0219854, 2019. doi: 10.1371/journal.pone.0219854 6

[24] G. J. Matheson. We Need to Talk about Reliability: Making Better Use of Test-retest Studies for Study Design and Interpretation. *PeerJ*, 7:e6918, 2019. doi: 10.7717/peerj.6918 6, 7

[25] K. Peterman, K. A. Cranston, M. Pryor, and R. Kermish-Allen. Measuring Primary Students' Graph Interpretation Skills Via a Performance Assessment: A Case Study in Instrument Development. *International Journal of Science Education*, 37(17):2787–2808, 2015. doi: 10.1080/09500693.2015.1105399 2

[26] D. F. Polit. Getting Serious About Test–retest Reliability: A Critique of Retest Research and Some Recommendations. *Quality of Life Research*, 23:1713–1720, 2014. doi: 10.1007/s11136-014-0632-9 6

[27] M. Rezaie and M. Golshan. Computer Adaptive Test (CAT): Advantages and Limitations. *International Journal of Educational Investigations*, 2(5):128–137, 2015. 2

[28] L. M. Rudner. Implementing the Graduate Management Admission Test Computerized Adaptive Test. In *Elements of Adaptive Testing, Statistics for Social and Behavioral Sciences*, pp. 151–165. Springer, New York, NY, 2009. doi: 10.1007/978-0-387-85461-8_8 1, 2

[29] D. O. Segall. Computerized Adaptive Testing. *Encyclopedia of social measurement*, 1:429–438, 2005. 4

[30] N. A. Thompson and D. A. Weiss. A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research, and Evaluation*, 16(1), 2011. doi: 10.7275/wqzt-9427 1, 2

[31] H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, and R. J. Mislevy. *Computerized Adaptive Testing: A Primer*. Routledge, New York, 2nd ed., 2000. doi: 10.4324/9781410605931 4

[32] O. B. Walter, J. Becker, J. B. Bjorner, H. Fliege, B. F. Klapp, and M. Rose. Development and evaluation of a computer adaptive test for 'Anxiety'(Anxiety-CAT). *Quality of Life Research*, 16:143–155, 2007. doi: 10.1007/s11136-007-9191-7 1, 2

[33] R. Wilms, R. Lanwehr, and A. Kastenmüller. Do We Overestimate the Within-Variability? The Impact of Measurement Error on Intraclass Coefficient Estimation. *Frontiers in Psychology*, 11:825, 2020. doi: 10.3389/fpsyg.2020.00825 6, 7