IJET

Dembitzer, L., Zelikpvitz, S., & Kettler, R. (2017). Designing computer-based assessments: multidisciplinary findings and student perspectives. *International Journal of Educational Technology, 4*(3), 20-31.

Original Paper

Journal homepage: http://educationaltechnology.net/ijet/

OPEN ACCESS

Designing computer-based assessments: multidisciplinary findings and student perspectives

Leah Dembitzer[a,*], Sarah Zelikovitz[b] & Ryan J. Kettler[c]

[a]*Center for Health Education, Medicine, and Dentistry, Lakewood, USA*
[b]*College of Staten Island of CUNY, Staten Island, NY, USA*
[c]*Rutgers, the State University of New Jersey, Piscataway, NJ, USA*

KEYWORDS

Accessibility
Accommodations
Computer-based assessment
Usability

ABSTRACT

A partnership was created between psychologists and computer programmers to develop a computer-based assessment program. Psychometric concerns of accessibility, reliability, and validity were juxtaposed with core development concepts of usability and user-centric design. Phases of development were iterative, with evaluation phases alternating with development phases. The system was evaluated by the team using an accessibility measure, standard usability heuristics, and student questionnaires (n = 131). In its final form, the assessment was satisfactory to students, although many students did not use the provided testing accommodations, and reliability analyses were in the acceptable range. Recommendations include updating and creating common language, standards, heuristics, and measures to develop and evaluate computer-based assessments.

Introduction

As in many public policy fields, the current climate in education increasingly emphasizes data-based decision making, with assessments used to generate the data (Spillane, 2012). Many initiatives in education have therefore focused on developing assessments for students that are useful for teachers, administrators, and policy makers (e.g. Curriculum Based Measures and tiered screening in Positive Behavior Support and Response to Intervention paradigms). In conjunction with this, educators and researchers are eager to harness the power of technology to assist in achieving the assessment goals. In one such example, the Race to the Top (RTT; U.S. Department of Education, 2009) federal grant funded two consortia to create computer-based standardized assessments linked to the Common Core standards.

As these initiatives mark a step forward for education, there is abundant literature addressing these trends. The RTT consortia have produced manuals explaining their use of current research and theory such as universal design, multi-tiered systems of support, and matching standards (PARCC, 2013; Smarter Balanced, 2013). Additionally, there exists a large research base addressing the impact of using computer-based assessment (CBA) instead of paper and pencil tests (Mead & Drasgow, 1993; Wang, Jiao, Young, & Brooks, 2007), discussing student use of supports (Higgins, Fedorchak, & Katz,

* *Corresponding author.* E-mail address: leah.dembitzer@gmail.com

2012), and examining the universal design aspect of CBA (Thompson, Johnstone, & Thurlow, 2002, Dolan et al., 2005; Ketterlin-Geller, 2005; Russell, Hoffman, & Higgins, 2009). However, there is a dearth in the education literature detailing the process of developing these assessments with a focus on the necessary partnerships with computer science experts. Indeed, many psychometricians and test developers seem unaware of graphic user design principles (Parshall, 2002). For example, concepts such as usability and user centered design are necessary parts of any software engineering project; these concepts are often ignored in industry design (Seffah & Metzker, 2004). In general, if technological support is not applied properly, it can lead to inefficiency and frustration (Te'eni, Carey, & Zhang, 2007). Specifically, when thinking about CBAs, a substandard design can be distracting to students, causing results of the assessment to be skewed (Weinwerth, Koenig, Brunner, & Martin, 2014). This paper models the necessary steps in creating a partnership, using the illustrative example of the Universal Design and Accommodations (UD&A) Project which featured that model.

Computer-Based Assessment

Computer-Based Assessment (CBA) refers to tests administered to students by computer. The purpose of using CBAs is to increase the efficiency of test administration and scoring and ensure the standardization of testing procedures. CBAs have also been able to offer innovations in test administration that cannot be available for paper and pencil tests, such as embedded accommodations and modifications, as well as the model of Computer Adaptive Testing (CAT; Parshall, Davey, and Pashley, 2000). There are also numerous challenges inherent in using CBAs. Students must be adequately trained in the user interface in order to reduce any advantage computer-savvy students may have. In addition, technological capabilities in schools often must be upgraded in order to offer the CBA of choice. However, the most important challenge to address when using CBAs is attaining or maintaining sufficient score quality for the intended inferences, as reflected by various types of reliability and validity evidence.

Psychological Perspectives in Test Development

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), provide guidance for the development and interpretation of educational and psychological tests. The recommendations are based on current research and theory in assessment and measurement. The *Standards* highlight the components of fairness or accessibility, reliability, and validity evidence necessary to support the use of a test.

Fairness/accessibility

Beddow (2012) describes test accessibility as the degree to which a test allows examinees to demonstrate their knowledge on a construct. An accessible test item allows a student entry to the target content being measured without barriers. For example, a multiple-choice test item must be visible for its content to be accessible. If the features of test items require certain test-taker characteristics (such as sensory perception or reception) that are not intended to be measured, the test is less accessible for students who have impairments in those areas, and will cause inferences from scores to be less valid. In terms of CBA, accessibility can be an important component because test-taker characteristics such as lack of computer literacy or poor screen-reading abilities can negatively impact student performance on the targeted constructs of reading, mathematics, or social studies.

Reliability of scores

Reliability refers to the consistency of a set of scores. Reliability is a necessary precursor to validity (Kaplan, 1987). Lack of adequate reliability indicates error in test scores, and can reflect issues with accessibility that must be addressed. Although required by the *Standards* (APA, AERA, NCME, 2014), many research studies in this domain do not focus on reliability (Kettler, 2011). Reliability is a critical consideration that must be included in CBA development.

Validity of inferences

According the *Standards* (APA, AERA, & NCME, 2014), the validity of a test refers to the evidence and theory that support using the scores to make inferences. Validity evidence comes in many forms. Validity evidence based on *test content* ensures the included content matches the proposed interpretation; evidence based on *response processes* describes the event of test-taking and how the skills, abilities, and processes used support the interpretation; *internal structure* evidence involves the structural makeup of the test construct and its match to proposed interpretation; and evidence based on *relations to other variables* explores the test's relationship to outside criteria or to other scores for similar or dissimilar constructs. When working with CBAs, all kinds of validity evidence are important; it is particularly important to ensure that the test does not include any user-interface related construct-irrelevant content (Fulcher, 2003), and that the response processes favor the proposed way of solving test problems.

Computer Design Perspectives in Test Development

User-centric design

From a computer design perspective, often the first issue software engineers and designers address is the purpose of a system, considering questions such as: What is this system intended to do? What will be learned from the system? How will the system be used? Who will use it? Good design is user-centric, whereby the prototype, design, and evaluation are user-driven, such that the users will answer these questions (Abras, Maloney-Krichmar, & Preece, 2004; Norman & Draper, 1986; Norman, 2013). In the case of CBAs, there are actually two tiers of users: (a) the students taking the test and the (b) teachers, proctors, administrators, and psychologists who interact with the students, and subsequently administer and interpret the tests. The designers of a system of this nature have to assure it is accessible to all students, and that it can be administered and interpreted easily by teachers and other stakeholders.

Usability

Usability is the extent to which a system allows those who interact with it to achieve their goals to a desired degree within a specified context (International Standardization Organization, 2013). The field of Human Computer Interaction (HCI) addresses usability issues and provides principles to maximize usability. HCI researchers have underscored that it is particularly important students feel natural and comfortable with the user-interface on CBAs, since a lack of understanding of the interface can cause frustration, and subsequently change the results of the assessment. In addition, CBAs that are not rated high on the usability scale will require more training. This, in itself, does not allow for fairness in the CBA because there are some students who will learn the system faster than others (Harms & Adams, 2008). Research has been conducted on user-interfaces for children (Fails, Guha, & Druin, 2013), indicating that intuition of adults is not always the correct way to approach user-design targeted to children. A small subset of this work has been conducted with teenagers (ages 13-18), the target group for the current study (Zeising & Katterfeldt, 2013). Teenagers have been around technology their entire lives, and have specific ideas regarding the appearance and functionality of an interface (Fitton & Bell, 2014). For usability to play an important role in the design of a CBA, it is important the designers become familiar with the user; evaluate, design, and produce prototypes iteratively by incorporating feedback; and embrace the standards of the HCI community.

Usability heuristics

The HCI community uses a set of standard heuristics to evaluate the usability of a system. These heuristics are principles that can be applied to design of user interfaces and aid in the recognition of problems during the design process. Keeping these heuristics foremost in the design process, and iterating with users, allows for a more usable design. These usability heuristics emphasize the clarity of the system, assuring there is a match to the real world and adhering to industry consistency and standards. There are also heuristics that specify ease of use, such as minimizing the amount users must remember, having an aesthetically pleasing interface, and allowing users to undo and redo items. An important set of these heuristics includes features that help the user such as (a) ability to recover from errors that are made inadvertently, (b) system status visibility for smoother navigation, and (c) help and documentation in all areas.

These generic heuristics are used for all different types of applications, and should be customized to the particular application and set of intended users (Weinerth, 2014). In the CBA discussed in this paper, that application is a reading comprehension test and the users are teenagers. CBA has particular challenges with respect to user interfaces, because the result of the assessment can be dependent on test-taker level of comfort with the interface (Martin, Koenig, & Weinerth, 2013). In addition, teenagers are a particular group that has not been studied extensively by the HCI community. Teenagers have specific qualities; they have extremely varied experience with computer interfaces, and that experience is often based on parental choice in addition to their own. In addition, it may be hard for teenagers to articulate exactly what is good or bad about a particular interface (Fitton, Read, & Horton, 2013).

The Universal Design & Accommodations Study

No studies have been conducted on integrating system design perspectives with best practices in assessment design. The UD&A Study attempted to address this gap in the literature by creating a partnership between computer scientists and psychologists, detailing the process and lessons learned, and providing data from the collaboration.

Procedures

Weinerth et al. (2014) provide a model for developing CBAs that can reduce students' cognitive burden of learning a new system. Their model emphasizes an iterative workflow with design and evaluation occurring in alternating stages. The researchers propose a development process in four phases, with the evaluation phase functioning as a feedback loop that brings the process back into the development and design phases. The first development phase is to specify the context of use, including users, tasks, and environment. This may involve some form of task analysis, as well as a realistic evaluation of the hardware and software resources available for the deployment of the CBA. The second phase is an outgrowth of the context of use and includes developing a list of requirements necessary in the computer program. Usability requirements, instructions for users, and scoring design are also developed at this phase. The third phase includes developing prototypes and models based on the requirements, and the fourth phase is the aforementioned evaluation phase. Weinerth et al. (2014) recommend using actual respondents and both objective and subjective methods to evaluate the usability, as well as incorporating usability heuristics to evaluate the system.

Phase 1: Task Analysis

According to Weinerth et al. (2014), task analysis should identify the users, the tasks, and the testing environment. One computer scientist and one school psychologist partnered to conduct the task analysis, write the questions, and develop the prototype, with the team eventually expanding to two psychologists and two programmers. The addition of two team members ensured both practical and research expertise were represented. The lesson learned from this phase was that all needs and tasks must be thoroughly discussed and delineated together, with opportunities for questions and answers throughout the process. Psychologists may not understand the limits of programming, while computer scientists may not understand specific study details such as those about communicating with students and collecting data. Optimal design emanates from discussion among team members to ensure such gaps in understanding do not influence the final product.

Users

The development team decided the users in this case would be both students who take the test and practicing psychologists who analyze the data (a separate group from the psychologists included on the design team). These two distinct groups of users were an integral part of the design process, ensuring the design was user-centric for both constituencies. Teachers were not considered a separate group of users in this project, since in practice they would likely receive the data through the psychologists. The decision to exclude teachers as a group of users simplified the programming demands, but necessarily meant the test would not be as usable for teachers on their own.

Tasks

The psychologists on the design team developed the list of tasks the computer system would facilitate. Initially, this was a simple list, and it was further expanded in Phase 2 of development. Tasks were considered in terms of the students using the test, the psychologists administering the tests, and the psychologists examining the data from the tests. Students needed to enter their user IDs, take a training session on use of the test, read two passages and answer multiple choice questions for each, use an audio or extra time accommodation (in some sessions), and answer demographic and evaluation questionnaires. The administration tasks included giving the correct form of the test to each student and opening the task on the computer in a secure manner. Data tasks included handling uniquely identified data from each student on answers to questions (regardless of correctness), time taken, number of times accommodations were accessed, and final grade and aggregate data.

Testing environment

Early on, decisions needed to be made about the technology available in schools to support the computer system. In many cases, this information was unknown, so systems needed to be developed in a versatile manner that could adapt to the broadest range of capabilities. The decision was made to put the system on a hard drive as instead of web-based, due to the robustness of this approach. Putting the system on a hard drive required manual consolidation of data before final results could be tallied. During field testing, difficulties arose that necessitated revisiting some of these original decisions.

Phase 2: Develop Lists of Requirements

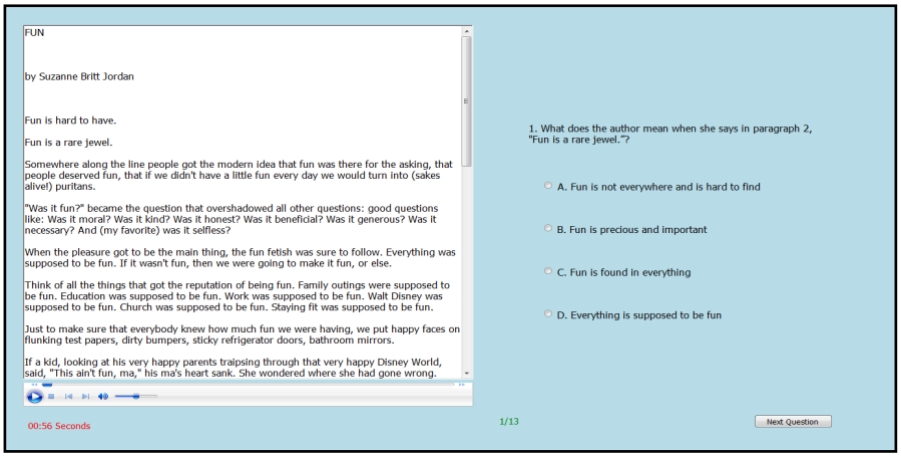
In this development phase, the team of psychologists and programmers developed lists of requirements of the system, including usability requirements. This phase was also iterative, because often a requirement requested by the psychologists did not fit with the programming plan, and would need to be reworked into a viable solution. For example, the psychologists did not realize the training examples had to include 3 questions, so the *previous* and *next* buttons could be seen on one screen whenever the middle question was active. Also during Phase 2, the programmers demonstrated for the psychologists the ease with which certain data can be collected using a CBA. For example, the system was able to compute the time a student spent on each individual question, as well as indicate which questions caused students to hit the *Play Question* button for a read aloud accommodation.

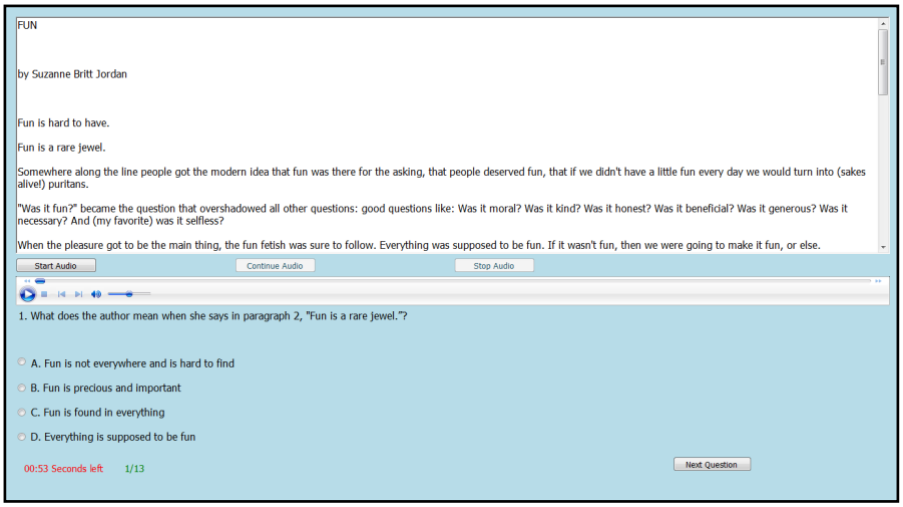
Also, while all project requirements were shared with the programmers, not all programmer requirements and concerns were shared with the psychologists. For example, it was not relevant to the psychologist developers which programming language would best achieve the goals, as long as the goals would be achieved. This phase could be conceptualized into two parts: (a) developing a list of requirements for the tasks and (b) developing a list of requirements for programming. Operating from two different perspectives, the psychologists and the computer scientists were able to generate a fuller list. The psychologists requested a requirement that the testing screen adjust to appear similarly across various screen shapes and sizes, while the computer scientists indicated that being able to go back and forth between questions is probably a desirable feature.

Phase 3: Develop Prototype

In many educational arenas, participants are stressed and timing is rushed. In this project, more complete information provided to the programmers enabled a better final product, while late changes induced confusion and frustration. The psychologists would make last-minute changes to the language of the training and instructions, which would add tasks to the programmers who were already working with system requirements. In the interest of efficiency, it was recommended that all text included in tests be as close to a final stage as possible before beginning to work on the system.

This development process was also iterative, because although detailed lists of requirements were created, the product often looked different than the psychologist users expected, necessitating additional tweaking. Also, it proved useful to the programmers to show the psychologists the product in stages before proceeding to add the next features. Figure 1 depicts some changes from the original design to second and third iterations.





Phase 4: Evaluation

During the evaluation phase, five eleventh- and twelfth-grade students piloted the program under testing circumstances and returned their feedback about usability. In addition, a team of school psychology graduate students evaluated the program using the Test Accessibility and Modification Inventory - Accessibility Rating Matrix (TAMI - ARM; Beddow et al., 2009). The programmers reviewed the prototype using the Nielson's usability heuristics (Nielsen, 1994) and focused on the feedback these aspects, modifying the interface accordingly.

Pilot evaluation

A convenience sample of five general education eleventh and twelfth grader students piloted the CBA concurrent with the TAMI review. Students discussed their comments afterward, and the objective data of testing time, scores, and accommodations used were reviewed. The students reported they found the system intuitive and easy to use. They indicated they were comfortable with the layout of the pages and items. Students requested buttons be more clearly labeled, and they criticized the voice for the audio accommodation. The pilot group used an average of one third of the time available to complete each section, and none of the students used the extra time accommodation.

TAMI-ARM evaluation

The Test Accessibility and Modification Inventory (TAMI; Beddow et al., 2009) was created to evaluate test questions, materials, formats, and stimuli to ensure maximal accessibility. Created based on theories of UD, accessibility, fairness, and cognitive load, the development of TAMI was also aided by practical research on testing accommodations, item writing, and item modification (Elliott et al., 2010). TAMI includes categories of Passage/Stimulus, Item Stem, Visuals, Answer Choices, and Page and Item Layout, yielding ratings for each category as well as a Total Item Accessibility Rating.

The 52 items of the CBA in its final form were coded for accessibility by two raters using the TAMI-ARM (Beddow et al., 2009). Of 52 items, 50 were coded in agreement, as indicated by the two separate ratings being within 1 point of each other (Beddow, Kettler, & Elliot, 2010). Another expert rater reviewed all items that were not equal, and raters met to reach consensus in such cases.

Figure 1: Two iterations followed the original prototype design. Colour and layout changes were necessitated, and audio was piloted after the general prototype was approved.

Most of the TAMI categories refer to the accessibility of the item as written (e.g. Passage/Stimulus, Item Stem, Answer Choices). One category - Page/Item Layout - refers to the presentation of the item on the paper or CBA. The TAMI-ARM also provides a comments section allowing raters to make suggestions for improvements that can increase the accessibility of the test. Table 1 depicts the Layout Ratings and Total Ratings for the items. A combined 61% of items received Total ratings of 3 or 4 (Maximally Accessible for All or Most Test-Takers), and 38% received a Total rating of 2 (Maximally Accessible for Some Test-Takers), with none receiving a rating of a 1. Of the 20 items that received a Total Rating of a 2, areas of concern were Passage (19 out of 20), Item Stem (17 out of 20), Answer Choices (14 out of 20), and Page/Item Layout (10 out of 20).

Table 1
TAMI ARM Results

	Page/Item Layout		Total Rating	
ARM Rating and Explanation	*n*	%	*n*	%
4 – Maximally Accessible for All Test-Takers	*16*	31	*7*	13
3 – Maximally Accessible for Most Test-Takers	*31*	60	*25*	48
2 – Maximally Accessible for Some Test-Takers	*5*	10	*20*	38
1 – Inaccessible for Many Test-Takers	*0*	0	*0*	0
Total	*52*	100	*52*	100

Note. TAMI ARM = Test Accessibility and Modification Inventory Accessibility Rating Matrix.

The Layout ratings were higher than the Total ratings, with 90% of items receiving Layout ratings of 3 or 4, and 10% of items receiving a Layout rating of 2. The comments in the Layout section referred to specific items and indicated items should be embedded in the passage, and information should be contained on one screen instead of multiple screens.

Usability heuristics evaluation

The computer scientists on the team used Neilsen's ten usability heuristics that are general principles for user interface design (Nielsen, 1994) and adapted them to CBA for teenagers. These usability heuristics are the standard in the HCI community, and evaluation of a software system on these heuristics reflects the intuitiveness of the system for users, as well as ease of interaction with the system. These heuristics are “broad rules of thumb, and not specific usability guidelines” (Nielsen, 1994). This allows the flexibility to interpret these heuristics as fitting the particular software application that is the focus of this paper.

Table 2. Documentation of Usability Heuristics

Visibility of system status	• The interface specifies which question out of the total number a student is reading. • Time remaining is clearly written at the bottom corner of the screen. The color changes when little time is left.
Match between system and the real world	• The multiple-choice questions were set up in a way that is familiar to high school students from all the exams that they have taken throughout their school years. • The interface for the audio player and the previous and next buttons are the expected icons that teenagers know.
User control and freedom	• Students can move back and forth between questions and change answers as desired.
Consistency and standards	• A sample passage and two test passages were given in the CBA, all in the same form with the same expected feedback.
Error Prevention	• Users are presented with a confirmation option before they commit to the final action of submitting each portion of the exam. • All possible clickable buttons were tested to make sure that none took students to any screen that was different than expected. • Only valid user IDs were accepted to take the test. • Data for each student was saved in separate files to prevent contamination.
Recognition rather than recall	• Passages remained at the side of the screen while students answered questions, so that the passage could be referenced. All buttons were clearly labeled and on the screen throughout the exam.
Flexibility and efficiency of use	• Students have a choice of listening to the passage, stopping during the audio, resuming, or beginning again. They can also listen to a question if necessary, but no student is forced to listen to audio.
Aesthetic and minimalist design	• Only the needed components for the CBA are visible on the screen at any time. • Research was done as to the clearest font and color, and those were chosen. • The design is minimalistic so students are not distracted by pictures and extra words. Only one question is shown to the user at a time to minimize complications.
Help and documentation	• The CBA begins with a training session that is done in two parts. One is a set of screen shots from the CBA, with clearly marked parts and directions. The second is a sample, simple test that students can take to try the system. At any point during the training, students can go back and restart the training, so that students can feel very comfortable with the system.

The programmers documented how the elements of the interface and programming components reflected the use of the heuristics. Table 2 lists Nielsen's heuristics with the documented ways the system followed them.

Phase 5: Back to Development

The feedback incorporated in this phase included re-labelling the buttons from "NEXT" and "PREVIOUS" to "NEXT QUESTION" and "PREVIOUS QUESTION", as well as re-labelling the audio buttons to be "Read Passage to Me" and "Read Question to Me." System requirements were more clearly delineated as well.

Although TAMI raters indicated that certain items should be embedded in the passage, student users reported they found the design confusing and preferred a vertical side-by-side alignment of passage and questions. The problem with the audio voice could not be addressed within the current project. The time limit norms were kept although the pilot group used only a third of the time available.

In this phase, decisions needed to be made about which feedback to incorporate. When the available research and the users offered differing perspectives, such as for embedding the questions in the passage, the user perspective was accommodated. Consideration of the heuristics was weighted the most in making these decisions, and the field testing offered more opportunity for clarifying these points.

Phase 6: Field Testing

Participants

Students (n = 131) were recruited from three high schools in New Jersey. The sample was primarily female (75%), and ethnically diverse, with 60% identifying as European American, 43% as African American, 10% as Latino/a American, 16% as Other, and 2% as Asian American/Pacific Islander. Four percent of the students reported they received Special Education services. Assessed using grade level Curriculum Based Measures (CBMs), 34% of students presented with functional impairments in reading fluency.

Measures

Participants were administered CBMs to establish reading fluency levels, and completed an evaluation form and a demographic questionnaire. The CBA itself contained two forms of a reading comprehension test, with and without accommodations.

Curriculum based measures

The CBMs were created by taking three grade level passages from *Bader Reading and Language Inventory*, *Fifth Edition* (Bader, 2005) and formatting the passages through Intervention Central's CBM generator to make Oral Reading Fluency (ORF) measures. According to Rasinksi (2005), reading fluency and performance can be adequately measured and predicted using ORF CBMs.

Reading comprehension test

The Reading Comprehension Test was a computer program in forms A and B. Each form included two passages and 26 multiple choice reading comprehension questions. The passages and questions were taken from the National Assessment of Educational Progress public access item bank, with any constructed response questions rewritten as multiple-choice questions. Participants completed both forms in two separate timed sessions. During one administration, the test included accommodation options to have the reading passage and questions read aloud and to allow the time limit to increase by 50%.

Demographic questionnaire

Participants completed a demographic questionnaire on the computer which included information such as gender, age, educational placement, and ethnic/racial background.

Evaluation survey

The evaluation survey was used to gain qualitative and quantitative information regarding the subjective experience of the participants in completing the reading comprehension test. Questions addressed organization, clarity, and ease of use of the testing platform, as well as whether the accommodations offered were helpful. This survey was completed by computer as well. The evaluation questions used are presented in Appendix A.

Data analysis

The students' qualitative data was analyzed through an open coding and inductive process (Patton, 2002) to develop themes and subsequently re-examined to confirm hypotheses. Reliability was measured by computing Cronbach's alpha. Exploratory correlational analyses reflected the relationships between student performance on CBMs, performance on the reading comprehension test, and use of accommodations, indicating areas in need of exploration for the establishing of content validity.

Procedures

School administrators were initially contacted for permission, and students were subsequently recruited from these schools. The schools were offered a professional development session on the evidence based selection of testing accommodations for their participation. Students were offered incentives in the form of gift cards and raffles for participation. The students completed two forms of the computer-based reading comprehension test, one in each of two conditions (with options of testing accommodations and without options of testing accommodations), and then completed a demographic questionnaire and an evaluation survey. Approval to conduct this study was obtained from the IRB of the university of the lead author.

Field testing results

Although testing had been exhaustive during the previous steps, new problems that had not been anticipated arose during field testing. With the system requirements updated, the test was able to function on computers at all three schools. However, the audio was controlled differently and separately in one of the schools. The students needed individual technological support to allow the audio accommodation to work. In addition, it was discovered that with certain computer systems, sound volume was not allowed to be controlled directly from the test and again needed technological support for intended use. Otherwise, the test was easy to administer as reported by the psychologist administrators, and the data was easily compiled.

Cronbach's alpha was calculated for each form by condition. Table 3 indicates the reliability of the forms for the entire sample. Form A had a reliability of .83 for the non-accommodated condition and .84 for the accommodated condition. Form B had a reliability of .81 for the non-accommodated condition and .75 for the accommodated condition. This reliability range is acceptable for a research test, but should be higher for a test used for high-stakes testing. Error lowering reliability could have emanated from the written items or from the interface; future research is necessary to address this distinction.

Table 3
Cronbach's Alpha for Reading Comprehension Test

	Items	*N*	Cronbach's Alpha (95% C.I.)
Non-accommodated Form A	26	61	.83 (.76-.88)
Accommodated Form A	26	62	.84 (.78-.89)
Non-accommodated Form B	26	62	.81 (.74-.87)
Accommodated Form B	26	60	.75 (.65-.83)

Note. C.I. = confidence interval.

Exploratory correlational analyses indicated CBM scores had a medium positive correlation with non-accommodated (.43) and accommodated comprehension scores (.36). In their review of the literature, Shinn et al. (1992) have found correlations between ORF CBMs and norm-referenced reading tests ranging from .60 to 90, with most around .80, indicating that ORF CBMs and reading tests represent highly related and possibly overlapping constructs. The medium correlations found in this study indicate the constructs of reading fluency and reading comprehension on this test are related but distinct.

Overall, qualitative analysis from student surveys was positive, with students mentioning the word "easy" 18 times in the data. The students who used the word "easy" did not necessarily achieve high grades on the test. Two students commented this test was no better or worse than regular tests. For example, one student wrote, "It was generic and was no better or worse than any other computer-based reading comprehension test i [sic] have ever taken." The students' use of "easy" seemed to reflect that use was within the regular realm of student experience.

Twenty-nine students commented on the positive presentation of the computer based test. Specifically, students liked that the passage and questions were readily available at the same time. One student said, "The story is always on the left side and can be scrolled. We don't have to flip any pages." Twelve students commented on the directions and explanations accompanying the test, stating that they were helpful. One said, "The practice part was also good for people who did not understand it. That is also something that should stay in the test. It showed you how to answer questions and what everything meaned [sic], so it was helpful." Four comments disagreed with this view, with one stating, "…the tutorial part is more common sense and should not be required to take."

Although most students who used the audio presentation accommodation found it helpful, there were 31 students who wrote negative comments about the technical quality and style of the audio. For instance, one wrote, "The audio was read to me in a monotone computerized voice, and it sounded staticky and harsh on the ears when I tried to listen to it using the headphones." Another wrote, "It did not read in an efficient manner for my needs. I expected it to read faster and flow better." Eleven students commented that they do not need the audio presentation and prefer to read on their own, such as the one who wrote, "It might be helpful for some people, but i [sic] like to read the story myself to fully understand it."

Twenty-eight students commented that they did not need the extra time accommodation, such as one who stated, "I did not need the extra time, so that was not a problem for me." Fifteen students wrote that the extra time was psychologically helpful to them and made them feel better about testing (none of these actually used the accommodation). One said, "It made me feel like I didn't have to rush." Another wrote, "It gave me time to think twice about my answers and check if they could possibly be corrected or not." These students did not do well on the test; they only received three and four correct out of 13, respectively.

As with the audio presentation accommodation, students were concerned for other students who may need it. Therefore, many students said it was helpful even if they personally did not use it. One wrote, "For the people who take longer to read its [sic] good to know they have more time to read and answer the question the best they can."

High percentages of students responded favorably to questions about the overall organization of the test (94-98%). Eighty-two percent of students responded favorably when asked specifically about the time remaining feature, 91% responded favorably when queried about the vertical presentation of the passage and questions, and 67% responded favorably about the training module.

When asked about the accommodations, only 53% of students considered the audio accommodation helpful, and 61% endorsed that the extra time was helpful. Overall, students did not use many accommodations in the field testing. The mean number of times accommodations were accessed for the sample was 3.01 and extra time was only used twice altogether.

Discussion

The *Standards* mandates reliability and validity evidence in the development of tests, and the development process is the starting ground on which these concerns can be addressed. Some key points for future development were noted from this study. As in all design, an iterative process is necessary. The rapid cycling between both teams ensured that specifications not previously considered were able to be added, and the scheduled evaluation phases ensured that no concerns were overlooked. Therefore, both rapid informal formative and systematic summative evaluation is recommended when developing a CBA.

There is a lack of awareness and different terminology between disciplines, causing difficulty in development. Psychometric concerns of reliability and validity do match up to design concerns of usability, but the disciplines differ in the gathering of evidence. Standardized language detailing expectations both from a computer science standpoint and from a psychometric standpoint would be a better starting ground.

Even when the system was in its final evaluation phase, new and unexpected problems arose. Therefore, evaluation from all users (students, teachers) should be incorporated in fully developed CBAs to ensure valid functioning which leads to interpretable results.

Limitations

One limitation of this study was the decision to exclude teachers as users. This only allowed measurement and student concerns to be at the forefront, while teacher perspectives could have been a contribution.

Another limitation was in the platform decision for the exam. Different teenagers are comfortable with various interfaces and operating systems; this CBA was Windows based, and favored those comfortable with Windows/Microsoft software.

Future Research

It is necessary to develop standard usability heuristics specifically for CBAs that can be used to provide evidence of accessibility. The TAMI (Beddow et al., 2009) addresses concerns both from an item writing perspective and from a design perspective. Perhaps two measures or a two part measure can address the separate concerns. Further research is also needed on the accessibility of CBAs. For example, in this study, researcher evaluation and student evaluation differed with regard to embedding the item within the passage, with evaluators considering it negative for accessibility and students having a positive view. Further research on specific design features can only benefit this burgeoning field. Further, the CBA developed in this study can be improved based on feedback from the evaluation phase. Evidence for such tests can help determine whether validity is sufficient to allow for inferences relevant to high-stakes decision making.

Conclusions

Technology offers tremendous opportunities to the field of educational assessment in terms of convenience, efficiency, and features. However, technological application on its own does not guarantee a user experience that will facilitate valid inferences from scores. The convergence of concepts such as user-centric design and accessibility show the necessity of a collaborative process between computer science design experts and psychometricians designing a CBA, as well as further research aimed at setting standards for this process.

REFERENCES

Abras, C., Maloney-Krichmar, D., Preece, J. (2004) User-Centered Design. In Bainbridge, W. *Encyclopedia of Human-Computer Interaction*. Thousand Oaks: Sage Publications.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Bader, L. A. (2005) *Bader reading and language inventory, 5th edition.* Pearson.

Beddow, P. A. (2012). Accessibility theory for enhancing the validity of test results for students with special needs. *International Journal of Disability, Development and Education*, *59*(1), 97-111. doi: 10.1080/1034912X.2012.654966

Bocij, P., & Greasley, A. (1999). Can Computer-Based Testing Achieve Quality and Efficiency in Assessment?. *International Journal of Educational Technology*, *1*(1), n1.

Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of UD to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment*, *3*(7). Available from http://www.jtla.org

Fails, J. A., Guha, M. L., & Druin, A. (2013). Methods and techniques for involving children in the design of new technology for children. *Foundations and Trends® in Human–Computer Interaction*, *6*(2), 85-166.

Fitton, D., & Bell, B. (2014, September). Working with teenagers within HCI research: understanding teen-computer interaction. In *Proceedings of the 28th International BCS Human Computer Interaction Conference on HCI 2014-Sand, Sea and Sky-Holiday HCI* (pp. 201-206). BCS.

Fitton, D., Read, J. C. C., & Horton, M. (2013, April). The challenge of working with teens as participants in interaction design. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (pp. 205-210). ACM.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language testing*, *20*(4), 384-408.

Harms, M., & Adams, J. (2008, March). Usability and design considerations for computer-based learning and assessment. In *Meeting of the American Educational Research Association (AERA)*.

Higgins, J., Fedorchak, G., & Katz, M. (2012). *Assignment of Accessibility Tools for Digitally Delivered Assessments: Key Findings.* Dover, NH: Measured Progress.

International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, *6*(2), 143-171.

International Standardization Organization. *Ergonomics of human-system interaction: Part 210 human centred design for interactive systems.* Retrieved from: https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing UD for assessment. *Journal of Technology, Learning, and Assessment*, *4*(2). Available from http://www.jtla.org

Kettler, R. J. (2011). Holding modified assessments accountable: Applying a unified reliability and validity framework to the development and evaluation of AA-MASs. In M. Russell & M. Kavanaugh (Eds.), *Assessing Students in the Margins: Challenges, Strategies, and Techniques* (pp. 311–334). Charlotte, NC: Information Age Publishing.

Kettler, R.J., Braden, J.P., & Beddow, P.A. (2011). Test-taking skills and their impact on accessibility for all students. In S.N. Elliott et al. (Eds.) *Handbook of Accessible Achievement Tests for All Students*, (pp. 147-159). Springer.

Kettler, R.J., Feeney-Kettler, K.A., Palladino, M.A., Zahra, L.P., & Rodriguez, J.C. (2013). A comprehensive framework for evaluating systems for screening preschool behaviors. In C.H. Qi & T. Stanton-Chapman (Eds.). *Preschool children: Education, language, social functioning and behavioral issues*. Hauppauge, New York: Nova Science Publishers, Inc.

Martin, R., Koenig, V., & Weinerth, K. (2013). *The importance of human-computer interactions in computer-based assessment.* Educational Measurement and Applied Cognitive Science. Retrieved from: https://www.taotesting.com/wp-content/uploads/2014/09/Romain-Martin-Pres_Bern.pdf

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449. Doi: 10.1037/0033-2909.114.3.449

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. *Proc. ACM CHI'94 Conf.*(Boston, MA, April 24-28), 152-158.

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic Books (AZ).

Norman, D. A., & Draper, S. W. (1986). User centered system design. *Hillsdale, NJ*, 1-2.

Parshall, C. G. (2002). *Practical considerations in computer-based testing*. Springer Science & Business Media.

Parshall C.G., Davey T., Pashley P.J. (2000) Innovative Item Types for Computerized Testing. In: van der Linden W.J., Glas G.A. (eds) *Computerized Adaptive Testing: Theory and Practice*. Springer, Dordrecht. Doi: 10.1007/0-306-47531-6_7

Partnership for Assessment of Readiness for College and Careers (2013). *Accessibility features and accommodations manual, 1st edition.* Retrieved from: http://www.parcconline.org/parcc-assessment-policies

Rasinski, T. V., Padak, N. D., McKeon, C. A., Wilfong, L. G., Friedauer, J. A., & Heim, P. (2005). Is reading fluency a key for successful high school reading? *Journal of Adolescent & Adult Literacy*, *49*(1), 22-27. doi: 10.1598/JAAL.49.1.3

Russell, M., Hoffman, T., & Higgins, J. (2009). NimbleTools: A universally designed test delivery system. *Teaching Exceptional Children*, *42*(2), pp 6-12.

Seffah, A., & Metzker, E. (2004). The obstacles and myths of usability and software engineering. *Communications of the ACM*, *47*(12), 71-76.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, *21*(3), 469-479.

Smarter Balanced Assessment Consortium (2013). *Usability, accessibility, and accommodations guidelines.* Retrieved from: http://www.smarterbalanced.org/wordpress/wp-content/uploads/2013/09/SmarterBalanced_Guidelines_091113.pdf

Spillane, J. P. (2012). Data in practice: Conceptualizing the data-based decision-making phenomena. *American Journal of Education*, *118*(2), 113-141. Doi: 10.1086/663283

Te'eni, D., Carey, J. M., & Zhang, P. (2005). *Human-computer interaction: Developing effective organizational information systems*. John Wiley & Sons.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *UD applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved September 30, 2013, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, *67*(2), 219-238. Doi: 10.1177/0013164406288166

Weinerth, K., Koenig, V., Brunner, M., & Martin, R. (2014). Concept maps: A useful and usable tool for computer-based knowledge assessment? A literature review with a focus on usability. *Computers & Education, 78*, 201-209. Doi: 10.1016/j.compedu.2014.06.002

U.S. Department of Education (2009). *Race to the Top Program Executive Summary*. Washington, DC: Author.

Zeising, A., Katterfeldt, E. (April, May 2013). Where is the 'like' button? Going beyond usability when designing for and with teens. Paper presented at *CHI'13,* Paris, France.

Appendix A

Evaluation Survey

The following questions address your experience in taking the computer reading comprehension test:

1. Have you ever taken a reading comprehension test on a computer before participating in this study?
a. Yes
b. No

2. Do you think this test was presented in an organized way?
a. Yes
b. No
Please explain

3. Did you find the test format easy to use?
a. Yes
b. No
Please explain

4. Did you find the time remaining feature (pictured) helpful?
a. Yes
b. No

5. Did you find the passage on the side helpful?
a. Yes
b. No

6. Did you find the training module at the beginning of the testing session helpful?
a. Yes
b. No

7. Did you find the audio presentation of the passages and questions helpful?
a. Yes
b. No
Please explain

8. Did you find the extra time option helpful?
a. Yes
b. No
Please explain

9. Was it easy to understand how to access the audio presentation and the extra time?
a. Yes
b. No
c. If no, what could have helped?

10. Please share any further information about your experience taking this test that could be helpful. Thank you!