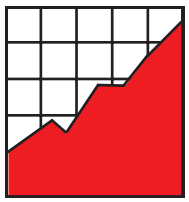
Technical Report 42

Considerations for the Development and Review of Universally Designed Assessments



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

Technical Report 42

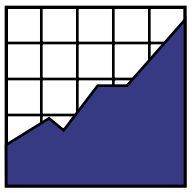
Considerations for the Development and Review of Universally Designed Assessments

Sandra J. Thompson • Christopher J. Johnstone • Michael E. Anderson • Nicole A. Miller

November 2005

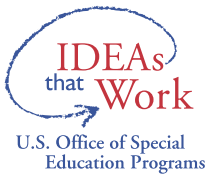
All rights reserved. Any or all portions of this document may be reproduced and distributed without prior permission, provided the source is cited as:

Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The research reported in this paper was supported by a grant (#H324D020050) from the U.S. Department of Education, Office of Special Education Programs, Directed Research Division. Points of view expressed herein are those of the authors, and not necessarily those of the U.S. Department of Education or Offices within it.



NCEO Core Staff

Deb A. Albus
Christopher J. Johnstone
Jane L. Krentz
Kristi K. Liu
Ross E. Moen
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Sandra J. Thompson
Martha L. Thurlow, Director

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://nceo.info

The University of Minnesota is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation.

This document is available in alternative formats upon request.

Acknowledgements

NCEO extends its sincere appreciation to the expertise of the individuals who provided us with thoughts, feedback, and suggestions in order to further develop and refine the considerations for universally designed assessments:

Karen Barton, CTB McGraw Hill

Sheryl Burgstahler, DO-IT Center, University of Washington

Margo Gottlieb, Illinois Research Center

Tom Haladyna, Arizona State University

Tracey Hall, CAST, Inc.

Barbara Henderson, American Printing House for the Blind

Scott Marion, National Center for the Improvement of Educational Assessment

Ken Olsen, Mid South Regional Resource Center

Marge Petit, National Center for the Improvement of Educational Assessment

Charles Stansfield, Second Language Testing, Inc.

Gerald Tindal, University of Oregon

Carol Traxler, Gallaudet University

Tim Vansickle, Minnesota Department of Education

Executive Summary

Universal design is an approach to educational assessment based on principles of accessibility for a wide variety of end users. Thompson, Johnstone, and Thurlow described seven elements of universally designed assessments in their 2002 report entitled *Universal Design Applied to Large Scale Assessments*. Elements of universal design include inclusive test population; precisely defined constructs; accessible, non-biased items; tests that are amenable to accommodations; simple, clear and intuitive procedures; maximum readability and comprehensibility; and maximum legibility. Since the 2002 report, Universal Design Project staff have examined research from a variety of fields in an effort to specify how elements of universally designed assessments can be put into practice.

This report describes the development of a "considerations of universally designed assessments" form based on Thompson et al.'s original elements. Considerations are specific questions for test designers to take into account while designing assessments. This report provides the original list of considerations from Thompson et al., then describes a validation process, whereby assessment and content area experts participated in a Delphi study. The Delphi study illuminated expert consensus on some considerations and disagreement on others. All expert commentary is captured in the text of this paper and in Appendix C (in tabular form), and a revised list of considerations is found in Appendix D.

Based on the comprehensive work represented in this report, several recommendations are presented for the use of the considerations of universal design at all stages of test development:

1. Incorporate elements of universal design in the early stages of test development.
2. Include disability, technology, and language acquisition experts in item reviews.
3. Provide professional development for item developers and reviewers on use of the considerations for universal design.
4. Present the items being reviewed in the format in which they will appear on the test.
5. Include standards being tested with the items being reviewed.
6. Try out items with students.
7. Field test items in accommodated formats.
8. Review computer-based items on computers.

Table of Contents

Introduction

The term universal design has been applied to a variety of educational approaches over the past several years. For instance, universal design for learning was first described by the Council for Exceptional Children (CEC) in a *Research Connections* article (CEC, 1999). Likewise, Thompson, Johnstone, and Thurlow (2002) of the National Center on Educational Outcomes (NCEO) described universal design approaches to large-scale assessment. In their initial paper on universal design of assessments, Thompson et al. outlined seven elements of universally designed assessments (inclusive assessment population; precisely defined constructs; accessible, non-biased items; amenable to accommodations; simple, clear and intuitive procedures; maximum readability and comprehensibility; and maximum legibility). Although elements of universal design provide guidance to states and assessment companies about design issues, there is still a need for specific information concerning what considerations should be made in test development in order to make tests accessible to a wide range of students.

This report summarizes the process of developing and refining a list of considerations for the universal design of statewide assessments for all students, including students with disabilities and English language learners. The staff of the Universal Design Project at NCEO, working closely with experts in the fields of assessment, disability, content areas (reading and math), and language acquisition, completed this version of considerations in the summer of 2004. This revision was one of three, which followed the compilation of an initial set of considerations identified from a literature review of multiple content areas (see Thompson, et al., 2002). The first version included stakeholder input from the Council of Chief State School Officers (CCSSO) conference on large-scale assessment in 2003. Following CCSSO feedback, a second version (a Delphi review, see description later in the text) was developed by NCEO in partnership with the Minnesota Department of Education, with a primary focus on students with limited English proficiency. This report describes the process of refining the considerations during a third validation study conducted by the Universal Design Project at NCEO. This is the third version of the considerations for use by test developers and item reviewers. This report also discusses the process used to validate the considerations, the issues that arise when using these considerations, and recommendations for use.

Purpose of the Study

The purpose of this report is to describe the process of developing and refining a set of considerations for item developers and item review teams to take into account in the universal design of inclusive, standardized, statewide assessments. Although the goal of this process was to find design strategies that maximize the accessibility of tests and test items, a larger goal was

to create an instrument to guide careful consideration of the elements of test design in order to discover issues in items that may be problematic.

What is Universal Design?

More than 20 years ago, Ron Mace, an architect who was a wheelchair user, began to actively promote a concept he termed "universal design." Mace was adamant that his field did not need more special purpose designs that serve primarily to meet compliance codes and may also stigmatize people. Instead, he promoted design that works for most people, from the child who cannot turn a doorknob to the elderly woman who cannot climb stairs to get to a door (Mace, 1998).

The term universal design is found in the newly reauthorized Individuals with Disabilities Education Act of 2004 (Public Law No: 108-446). Specifically, IDEA of 2004 states that:

> The State educational agency (or, in the case of a districtwide assessment, the local educational agency) shall, to the extent feasible, use universal design principles in developing and administering any assessments under this paragraph (§ 612(a)(16)(E).

Universal design is specifically defined in the U.S. Assistive Technology Act of 2004 (Public Law No. 108-364-ATA 2004) as follows:

> [A] concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly accessible (without requiring assistive technologies) and products and services that are interoperable with assistive technologies.

Assessments that are universally designed are designed from the beginning, and continually refined, to allow participation of the widest possible range of students, resulting in more valid inferences about performance. These assessments are based on the premise that each child in school is a part of the population to be tested, and that test results should not be influenced by disability, gender, race, or English language ability. Universally designed assessments are not intended to eliminate individualization, but they may reduce the need for accommodations and various alternative assessments by eliminating access barriers associated with the tests themselves.

The elements of universal design, according to Thompson et al., are:

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear and intuitive procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

From these elements, universal design staff constructed considerations for universally designed assessments. The considerations are a list of specific questions that help test designers locate potential design issues in items. The considerations are listed in Table 1.

Table 1: Considerations for Universally Designed Assessment Items

Does the item…
Measure what it intends to measure • Reflect the intended content standards (reviewers have information about the content being measured) • Minimize skills required beyond those being measured
Respect the diversity of the assessment population • Accessible to test takers (consider gender, age, ethnicity, socio-economic level) • Avoid content that might unfairly advantage or disadvantage any student subgroup
Have clear format for text • Standard typeface • Twelve (12) point minimum for all print, including captions, footnotes, and graphs (type size appropriate for age group) • Wide spacing between letters, words, and lines • High contrast between color of text and background • Sufficient blank space (leading) between lines of text • Staggered right margins (no right justification)
Have clear pictures and graphics (when essential to item) • Pictures are needed to respond to item • Pictures with clearly defined features • Dark lines (minimum use of gray scale and shading) • Sufficient contrast between colors • Color is not relied on to convey important information or distinctions • Pictures and graphs are labeled
Have concise and readable text • Commonly used words • Vocabulary appropriate for grade level • Minimum use of unnecessary words • Idioms avoided unless idiomatic speech is being measured • Technical terms and abbreviations avoided (or defined) if not related to the content being measured • Sentence complexity is appropriate for grade level • Question to be answered is clearly identifiable

Table 1: Considerations for Universally Designed Assessment Items (continued)

Allow changes to its format without changing its meaning or difficulty (including visual or memory load) • Allows for the use of braille or other tactile format • Allows for signing to a student • Allows for the use of oral presentation to a student • Allows for the use of assistive technology • Allows for translation into another language
Does the test…
Have an overall appearance that is clean and organized • All images, pictures, and text provide information necessary to respond to the item • Information is organized in a manner consistent with an academic English framework with a left-right, top-bottom flow
In addition to the other considerations, a computer-based test should have these considerations:
Layout and design • Sufficient contrast between background and text and graphics for easy readability • Color is not relied on to convey important information or distinctions • Font size and color scheme can be easily modified (through browser settings, style sheets, or on-screen options) • Stimulus and response options are viewable on one screen when possible • Page layout is consistent throughout the test • Computer interfaces follow Section 508 guidelines **Navigation** • Navigation is clear and intuitive; it makes sense and is easy to figure out • Navigation and response selection is possible by mouse click or keyboard • Option to return to items and return to place in test after breaks **Screen reader considerations** • Item is intelligible when read by a text/screen reader • Links make sense when read out of visual context (“go to the next question” rather than “click here”) • Non-text elements have a text equivalent or description • Tables are only used to contain data, and make sense when read by screen reader **Test specific options** • Access to other functions is restricted (e.g., e-mail, Internet, instant messaging) • Pop up translations and definitions of key words/phrases are available if appropriate to the test • Students are able to record their responses and read them back (and have them read back using text-to-speech) as an alternative to a human scribe, but only if student has experiences with this mode of expression and chooses it for the test **Computer capabilities** • Adjustable volume • Speech recognition available (to convert user’s speech to text) • Test is compatible with current screen reader software • Computer-based option to mask items or text (e.g., split screen) • Computer software for test delivery is designed to be amenable to assistive technology

Delphi Review

We conducted a Delphi review to determine the usefulness of existing considerations for universally designed assessments. The intent of the Delphi review was to invite experts in the fields

of assessment, special education, academic content, and language acquisition to give input on the considerations and modify them accordingly (Adler & Ziglio, 1996). The Delphi method is a structured process of using a series of questionnaires to gather the combined input from a group of persons with expertise related to a specific area or population. The method has been used in the social science and public health fields since the mid-1970s (Adler & Ziglio, 1996). Delphi studies allow participants to give their own informed opinion on an issue. The input is then compiled and returned to the participants who can respond to further questions, respond to the input from the other participants, and revise their own comments if desired. All iterations of Delphi are anonymous.

This Delphi study took place entirely by e-mail. Participants were unaware of who was invited to participate in the study, who elected to participate, and the individuals who provided feedback (anonymity was maintained throughout the study). All suggestions and comments were given equal weight.

Participants

Universal Design Project research staff identified a group of experts to review the considerations for universally designed assessments. To ensure that important areas of expertise were represented, a chart was created and participants were recommended based on their expertise in one or more of the identified areas (see Table 2). These individuals were then invited to participate in the Delphi review before the first Delphi questionnaire was sent out. The resulting group of Delphi participants represented experts in the field of assessment, assistive technology, computer-based testing, reading, math, second language acquisition and testing, disability consultation, and special education.

Table 2: Expertise and Participants

Vision	Barbara Henderson
Computer-based testing, learning disabilities	Gerald Tindal
Item analysis	Karen Barton
Second language acquisition and testing	Margo Gottlieb
Second language acquisition, testing, and translation	Charles Stansfield
Physical disabilities	Sheryl Burgstahler
Hearing_	Carol Traxler
Science	Scott Marion
Psychometrics	Tom Haladyna
Assistive technology	Tracy Hall
Math	Marge Petit
Special education assessment	Ken Olsen
State Assessment Director	Tim Vansickle

Delphi Process

The first Delphi survey (Delphi Form 1—see Appendix A) was developed to obtain specific feedback on the considerations draft presented by NCEO. Expert participants were provided ample opportunity to comment on the considerations or add to the list. The participants were asked first to rate the importance of each individual consideration on a five point Likert scale. They then were asked to comment on any of the considerations about which they felt strongly positive or negative. They could also pose questions on the form. Finally, they were asked to add any additional considerations and rate the importance of their additions. The participants were instructed to try to think about the considerations in terms of their usefulness for test developers and item reviewers.

In July 2004, the first Delphi survey (Delphi Form 1) was e-mailed to the participants. Each participant was given seven days to review the considerations and e-mail comments back to NCEO. The comments and ratings were returned by 13 of 14 participants. These were compiled at NCEO and a second survey was developed (Delphi Form 2–see Appendix B).

The second survey (Delphi Form 2) included a list of anonymous individual ratings and the mean from all ratings assigned to each consideration. All comments made by the participants on the first form were included in the second form. Participants were asked to comment on results from the initial survey, were probed on specific issues by NCEO researchers, and were asked to comment on the 15 considerations suggested by participants (the majority relating to computer-based testing). The second survey was e-mailed out at the beginning of August 2004 and participants were again given seven days to return their comments via e-mail. The comments were complied by the staff at NCEO in mid-August, 2004 (see Appendix C).

Response Rates

The original list of considerations (Delphi Form 1) was sent out via e-mail to 14 experts for review. Thirteen of 14 (93%) experts returned Delphi Form 1. The second survey (Delphi Form 2) was again sent out to the original 14 participants. The same thirteen participants returned the second survey (one participant did not participate in either survey). The feedback on both surveys was extensive.

Results

Using the feedback from both Delphi surveys, Universal Design Project staff revised the considerations for universally designed assessments (see Table 3). The considerations that had originally

been sent to reviewers were rated as somewhat important to extremely important (from 2.67 to 5), with an average of very important (i.e., 4.3) to consider in designing and reviewing assessments. One consideration was deleted based on expert feedback, while others were added or revised. The primary additions to the considerations were the expansion of the considerations for computer-based testing. In addition, there were several additions to the discussion points for the consideration note sections. All changes to the considerations are shown in Table 3, with additions marked by underlines and deletions shown by strikethroughs.

Table 3: Summary of Consideration Ratings and Changes

Does the item…	Range	Mean
Measure what it intends to measure		
• Reflect the intended content standards (reviewers have information about the content being measured)	5–5	5.00
• Minimize knowledge and skills required beyond ~~those being~~ what is intended for ~~measured~~ measurement.	3–5	4.33
Respect the diversity of the assessment population		
• ~~Accessible~~ Sensitive to test taker~~s~~ characteristics and experiences (consider age, gender, ethnicity, ~~and~~ socio-economic level, region, disability, and language)	4–5	4.75
• Avoid content that might unfairly advantage or disadvantage any student subgroup	4–5	4.64
Have clear format for text		
• Standard typeface	3–5	4.00
• Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (type size appropriate for age group)	3–5	4.09
• ~~Wide spacing between letters, words, and lines~~	2–5	3.09
• High contrast between color of text and background	3–5	4.09
• Sufficient blank space (leading) between lines of text	2–5	2.82
• Staggered right margins (no right justification)	2–5	3.36
Have clear visuals (when essential to item)		
• ~~Pictures~~ Visuals are needed to ~~respond to item~~ answer the question	3–5	4.56
• ~~Pictures~~ Visuals with clearly defined features (minimum use of gray scale and shading)	4–5	4.45
• ~~Dark lines (minimum use of gray scale and shading)~~	3–5	3.82
• Sufficient contrast between colors	1–5	3.64
• Color alone is not relied on to convey important information or distinctions	2–5	3.91
• ~~Pictures and graphs~~ Visuals are labeled	3–5	3.91
Have concise and readable text		
• Commonly used words (except vocabulary being tested)	1–5	4.18
• Vocabulary appropriate for grade level	4–5	4.83
• Minimum use of unnecessary words	1–5	4.17
• Idioms avoided unless idiomatic speech is being measured	3–5	4.67
• Technical terms and abbreviations avoided (or defined) if not related to the content being measured	4–5	4.73
• Sentence complexity is appropriate for grade level	1–5	4.45
• Question to be answered is clearly identifiable	5–5	5.00

Table 3: Summary of Consideration Ratings and Changes (continued)

Allow changes to its format without changing its meaning or difficulty (including visual or memory load)		
• Allows for the use of braille or other tactile format	3–5	4.67
• Allows for signing to a student	3–5	4.55
• Allows for the use of oral presentation to a student	3–5	4.36
• Allows for the use of assistive technology	3–5	4.45
• Allows for translation into another language	1–5	3.64
Does the test…		
Have an overall appearance that is clean and organized		
• All visuals (e.g., images, pictures) and text provide information necessary to respond to the item	3–5	4.50
• Information is organized in a manner consistent with an academic English framework with a left-right, top-bottom flow	4–5	4.33
• Booklets/materials can be easily handled with limited motor coordination (consideration was added)	0–5	4.00
• Response formats are easily ~~correlated~~ matched to question	0–5	3.43
• Place for student to take notes (on the screen for CBT) or extra white space with paper-pencil	0–5	3.82
In addition to the other considerations, a computer-based test should have these considerations:		
Layout and design		
• Sufficient contrast between background and text and graphics for easy readability	4-5	4.67
• Color alone is not relied on to convey important information or distinctions	2–5	3.92
• Font size and color scheme can be easily modified (through browser settings, style sheets or on-screen options)	2–5	4.08
• Stimulus and response options are viewable on one screen when possible	3–5	4.67
• Page layout is consistent throughout the test	4–5	4.75
• Computer interfaces follow Section 508 guidelines (www.section508.gov)	0–5	3.56
Navigation		
• Students have received adequate training on use of test delivery system	0–5	4.46
• Navigation is clear and intuitive; it makes sense and is easy to figure out	4–5	4.92
• Navigation and response selection is possible by mouse click or keyboard	3–5	4.67
• Option to return to items and return to place in test after breaks	3–5	4.60
Screen reader considerations		
• Item is intelligible when read by a text/screen reader	3–5	4.58
• Links make sense when read out of visual context. (“go to the next question” rather than “click here”)	4–5	4.67
• Non-text elements have a text equivalent or description	3–5	4.30
• Tables are only used to contain data, and make sense when read by screen reader	3–5	4.36

Table 3: Summary of Consideration Ratings and Changes (continued)

Test specific options		
• Access to other functions is restricted (e.g., e-mail, Internet, instant messaging)	3–5	4.55
• Pop up translations and definitions of key words/phrases are available if appropriate to the test	3–5	4.08
• Students writing online can get feedback on length of writing on-demand in cases where there is a restriction on number of words.	0–5	2.67
• Students are able to record their responses and read them back ~~(or have them read-back using text-to-speech) as alternative to human scribble, but only if student has experiences with this mode of expression and chooses it for the test~~ as an alternative to human scribe.	0–5	3.69
• Students are allowed to create persistent marks to the extent that they are already allowed to paper-based booklets (e.g., marking items for review, eliminating multiple choice items, etc.)	0–5	4.17
Computer capabilities		
• Adjustable volume	3–5	4.50
• Speech recognition available (to convert user's speech to text)	1–5	3.67
• Test is compatible with current screen reader software	3–5	4.25
• Computer-based option to mask items or text (e.g., split screen)	0–4	3.00
• Computer software for test delivery is designed to be amenable to assistive technology	0–5	3.91

Notes that were added to the considerations address some of the anticipated issues that might arise when using the considerations. While we tried to keep the list of considerations brief and user-friendly, it was clear from participant comments that more explanation about the intent and issues surrounding the considerations needed to be presented close to the considerations in note form. The notes are not meant to be used as definitive judgment of the "good" or "bad" quality of an item or design feature. Instead, the notes are intended to add clarity to the considerations, help elucidate important issues, and help generate discussion.

Discussions About Selected Considerations

In addition to providing greater clarity to several of the considerations, many of the respondents in the Delphi review pointed out that using some of the considerations depended on the content being tested. Extensive discussion focused on issues of construct vs. content validity and the minimization of construct-irrelevant variance. There was also extensive discussion on the validity and practicality of the translation of assessments to languages other than English. In this section of the report, we present a detailed review of these discussions. Considerations about which few comments were made and no clarification was deemed necessary are not discussed. Responses to all considerations, however, can be found in Appendix C.

Consideration: *"Reflects the intended content standards (reviewers have information about the content being measured)."*

Following a discussion by Universal Design Project staff, Delphi participants were asked to comment on whether the first consideration should remain "Reflects the intended content standards (reviewers have information about the content being measured)" or whether it should be reworded "Reflects the intended *construct* (reviewers have information about the *construct* being measured)." Although opinions leaned toward changing the wording (Yes = 6, No = 3, Combination wording = 1, Did not state position but provided information to consider when making the decision = 2, Don't know = 1), only two of the participants in favor of using the term "construct" provided reasoning. One suggested that construct "would fit better with the professional terminology," while the other stated that "content is topical, constructs are conceptual. This difference in meaning is huge. Furthermore, construct is a term used in APA standards and is deeper than content."

The participants who wished the consideration to remain the same provided critical information about what to think about before a decision could be made. Specifically, one participant suggested that we consider our audience: "Construct is a formal term that theorists use. Content standards [are] what practitioners understand." Another participant suggested we consider what the terms imply: "…construct is a sort of overarching concept (i.e., reading) whereas content standards are…narrower (e.g., reproduces capital letters)…If the test is supposed to be a standards-based achievement test, then it must address standards. If not, then the item need only address the construct."

Ultimately, Universal Design Project staff decided to retain the term "content." This term appears to be consistent with the link of items to standards, and avoids the apparent confusion surrounding the term "construct." It should be noted, however, that the term "construct" may still be useful, especially if item developers (who are familiar with the concept of constructs) are using these considerations.

Consideration: *"Minimize knowledge and skills required beyond ~~those being~~ what is intended for ~~measured~~ measurement."*

The second consideration under review was altered slightly based on participant input. Initially, this consideration stated, "Minimize skills required beyond those being measured." This was changed to "Minimizes knowledge and skills required beyond what is intended for measurement" following several suggested alternate phrases. In addition to suggestions on phrasing, Delphi participants expressed concern that item writers or reviewers might interpret this consideration in such a way as to "…separate skills too much…[and thus run the risk that] we'll wind up with tests that measure isolated, basic skills." Still others expressed the belief that this consideration has direct relevance for the measurement of "higher level thinking." Yet, as another reviewer

questioned, "how…the other skills (are) defined and targeted" would be important in guiding item writers and reviewers. One participant summed up the issue by saying that it "…depends on how discrete the standards are; minimal skills can be embedded in more complex contextualized items. Ultimately, it depends on what you are measuring."

Consideration: *"~~Accessible~~ Sensitive to test takers~~*~~ characteristics and experiences (consider age, gender, ethnicity, ~~and~~ socio-economic level, region, disability, and language."*
The third consideration was changed from "Accessible to test takers (consider age, gender, ethnicity, and socio-economic level" to "Sensitive to test taker characteristics and experiences (consider gender, age, ethnicity, socio-economic level, region, disability, and language)." When asked about including the term "bias" in this consideration, participants were somewhat divided. While some indicated that bias should be included to "reference systematic variance that interferes with making a valid inference," others clarified that "bias and accessibility are separate issues from a review standpoint, though obviously related." Keeping participants' suggestions and reasoning in mind, it was decided that the term "bias" would be included in the note portion of the consideration and that the demographic variables would be expanded from four to seven, reflecting the need for greater sensitivity to the experiences of very diverse populations.

Consideration: *"Standard typeface."*
When considering the clarity of the format for text in assessments, most participants agreed that a standard typeface was important. There was, however, confusion about the meaning of "standard." Some Delphi participants had interpreted this consideration as implying that a single standard font existed, as illustrated in the following comment: "There is no standard typeface, thus the myriad fonts used in various publisher's files, even within the same text or textbook." In order to reduce confusion over the meaning of the term, however, it was determined that additional clarification was needed. Consequently, the following was added to the note section: "Use clear, common, familiar, and consistent fonts," followed by examples of font.

Consideration: *"Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (type size appropriate for age group)."*
When considering which font size to select, several Delphi participants noted the importance of considering the font style. Given the fact that a 12-point font can vary in size depending upon the font style, an additional issue was included in the note section. As suggested, one consideration (width of spacing between letters) was combined with font. One participant stated "Wide spacing is not necessarily best; proper font selection is more important." Consequently, this consideration was added to the note section of the consideration addressing font.

Consideration: *"High contrast between color of text and background."*
When considering the use of color in text or background, participants suggested going beyond the issue of contrast to consider print density. Specifically, one participant stated, "[E]ven with

sufficient color contrast, color blind users may not be able to distinguish text and background. [I] suggest you further recommend high *print density* contrast. This would also avoid isoluminance effects for non-visually-impaired students." ("Isoluminance" is the point at which two colors have an equivalent luminous intensity, or brightness.) Based on these comments, information on print density and isoluminance was added to the note section for the consideration addressing format for text.

Consideration: *"~~Pictures~~ Visuals are needed to ~~respond to item~~ answer the question."*
The use of visuals resulted in considerable discussion ranging from issues surrounding limiting visuals, the use of visuals to provide only redundant information, and the benefits/drawbacks of using visuals in relation to specific disabilities. In relation to the content of visuals, for example, it was suggested, "Pictures, line art, etc. should be related to the item [and] should enhance understanding, [but] not [be] required for understanding, with the exception of data tables like on math and science tests." Additionally, another Delphi participant stated, "often there are pictures used that are not redundant with the text but that are relevant to the item and to the construct." Consequently, it was suggested that the wording of this consideration take this idea into account. Rather than dramatically change the wording of this consideration, qualifying information was provided to the note portion below the consideration addressing the idea that clear and well-designed graphics or pictures should add value for students who need a visual cue.

Consideration: *"Commonly used words (except vocabulary being tested)."*
When considering the vocabulary used in assessments, both for directions and specific items, many Delphi participants commented on the need for greater clarity surrounding the specification that the text be comprised of "commonly used words." Several participants suggested that the term "age-appropriate" was preferable, while another suggested adding "concise and readable." Ultimately, the greatest concern with this particular consideration was that there be some acknowledgement that the words selected should be common, "with the exception of subject specific terminology…" In other words, the "item should consist of commonly understood words or vocabulary…" except when knowledge of specific vocabulary is being tested. One participant also suggested that the vocabulary be "…consistent with each specific grade level," with another suggesting "at or below grade level [when] reading is not the primary construct tested." As a result of this feedback, additional clarification was added to the wording of the consideration (i.e., the consideration was changed from "Commonly used words" to "Commonly used words (except vocabulary being tested)" as well as in the note section following the consideration.

Consideration: *"Allows for translation into another language."*
Perhaps the most controversial consideration of all was "Allows for translation into another language." One and one-half pages of initial comments, questions, and suggestions were followed by an additional one and one-half pages of responses, comments, questions, and suggestions.

The response of one participant summarized a number of the issues that participants grappled with when determining the appropriateness of this consideration:

> "This is a questionable and highly controversial issue, particularly when one realizes that such a standard is impossible to meet. About 72% of our LEP students are Spanish speakers, but the other 28% represent many diverse languages. How do we accommodate and what is the theoretical rationale and what is the technology for doing this? Is it possible? Is it beneficial?"

In reference to the impracticality of translating tests into the less commonly represented language groups, some participants questioned the fairness of accommodating some students (e.g., Spanish speakers) and denying others. Another stated "What harm is done by helping the 72% of LEP students who speak Spanish? We provide accommodations to others where possible, but some would propose that a translated test is harmful. Poppycock!"

Participants also suggested some disagreement in terms of the quality of the translations/skill of the translators. A primary problem with translation, however, was clear: "The limitation is money. Translation must be cost effective like everything else in education. You can't provide translated tests for very small numbers. The Lau decision (*Lau v. Nichols*, 1974) and other civil rights decisions make it clear that numbers dictate expectations of school systems." Given the cost, customized dictionaries were suggested as a possible alternative to fully translated tests.

Besides the practicality/impracticality of translating tests, one area of considerable debate surrounded the validity of the inferences that can be made from scores derived from translated tests. Some participants expressed the belief that translated tests reduced the validity of scores ("Data analysis has shown these to be less than valid measures of student performance."), or that certain translations would result in less valid scores ("Some critical and relevant word/concepts [do] not translate into every language."). Others, however, made the argument that there are few instances where concepts do not translate:

> "Minnesota translates to Hmong and Somali. Only in these languages are there relevant words/concepts that do not translate easily into English. The other languages of state assessment (Spanish, Russian, Chinese, Korean, Haitian Creole) almost never pose a problem for translating words or concepts. Professional translators will tell you they can translate almost any word or idea, and if they encounter one they can't, they will tell you that too."

Another participant added, "Translation is no more a threat to validity than a change in option order or a change in font. Such changes might generate a miniscule change in item difficulty, but they don't affect validity... [Translation] is the exact same test stated in a different language." Yet others brought up the issue of validity in reference to a specific construct being measured.

For example, two participants stated that translating English language arts (ELA) tests would invalidate the inferences that could be made from the scores. In light of NCLB legislation, a participant brought up a final important point of consideration: "A translated test is always much less of a threat to validity and score comparability than an alternate assessment," suggesting that a translated test is preferable to alternate assessment measures for English language learners.

Two reviewers suggested that this consideration be eliminated given the controversy, at least until more research was available. Ultimately, Universal Design Project research staff decided to retain this consideration, acknowledging the issues item writers and reviewers face as they incorporate this consideration into the test construction/revision process. This information was included in the note section following the consideration.

Summary of Revisions

At the completion of the study, the Universal Design Project staff revised the original considerations based on Delphi responses (Appendix D). The most extensive revisions were made to the content and wording of the considerations. Some of the most significant changes to the considerations that resulted from the Delphi process are described here:

1. Wording of several of the considerations was revised using feedback from the Delphi review participants. For example, "Minimize skills required beyond those being measured" was changed to "Minimize knowledge and skills required beyond what is intended for measurement" and "Accessible to test takers (consider age, gender, ethnicity, and socio-economic level)" was changed and expanded to "Sensitive to test taker characteristics and experiences (consider gender, age, ethnicity, socio-economic level, region, disability, and language)."

2. Computer-based testing considerations were expanded. Much of the useful feedback for this section came from reviewers who are familiar with the development of computer-based tests. With these revisions, the section of considerations for computer-based testing was clarified and redundancies with other considerations were eliminated.

3. Notes were added to the considerations. These notes discuss some of the anticipated issues that might arise when using the considerations. While we tried to keep the list of considerations brief and user-friendly, it was clear that more explanation about the intent and issues surrounding the considerations needed to be presented on the same page. The notes are intended to add clarity to the considerations and help elucidate important issues. Notes also provide evidence of the complexity of some

of the considerations and illustrate that considerations are not static rules, but general principles that aid in flagging potentially problematic items.

4. One font-dependent consideration ("Wide spacing between letters, words, and lines") was eliminated. Instead it was included in the note section for "Have a clear format for text."

5. Relevant research citations were added to the considerations so that people wanting to investigate a certain issue in more depth would have the resource citations at hand (see Appendix E).

6. We created a review checklist of the considerations for item reviewers and developers (see Appendix F). This form is intended to be used by item reviewers and developers who have received training on the considerations. It consists of a list of the considerations, without the supporting text. Using this form, item reviewers and developers can go through items and flag for further discussion areas of concern or alteration. For item reviewers, there is an additional form on which comments may be recorded explaining why some aspect of an item was flagged (Appendix G).

Issues Related to Universal Design

One of the most important outcomes of this review process was the identification of issues that surround the development of universally designed assessments. These issues highlight the complexities of a process without easy answers. The issues discussed in this section are not meant to be an exhaustive list of the challenges related to the universal design of assessments, but instead provide some guidance about the challenges that might be encountered when using the considerations.

1. **Universal design is not a cure all**. Just because a test is universally designed, or has used the elements of universal design to guide its development, does not mean that a test is accessible to all students. The considerations recommended in this report are just that, considerations. They are meant to be used to guide test developers and reviewers in creating tests that are accessible to the greatest number of students possible. However, some changes to a test that might make it more accessible to one group of students, might actually make it less accessible to another group. For example, eliminating or altering an illustration accompanying an authentic reading text may clarify an item by removing a distraction for some students. On the other hand, eliminating it may remove or change some useful context for the passage. Issues of accessibility need to be carefully considered and discussed openly so that informed

decisions can be made without hindering the construct being tested. Universal design can be a useful tool for developing better assessments, but it is not a tool that can magically make all tests accessible to all students.

2. **Universal design does not replace accommodations**. While universal design may remove some barriers for students with disabilities and English language learners, it in no way eliminates the need for testing accommodations. Some students may still need accommodations such as large print or assistive technology. A goal of universally designed assessments is to anticipate common accommodations and design tests that allow accommodations to be more easily integrated into the format of the test.

3. **Universal design does not replace good instruction**. The goal of universal design is to think about the full range of students taking an assessment so that they all can demonstrate what they have learned. A student who has not had an opportunity to learn the material tested will not be helped by a universally designed test.

4. **Universal design does not lower standards**. Some may perceive a universally designed assessment to be a "watered-down" or "easier" assessment. It is important to make clear the purpose of universal design is to make sure that the content being tested is more universally accessible to all of the students taking the test and thus a better measure of student learning.

5. **Technology use is challenging**. The quality of technology available across schools is an important issue when creating a computer-based assessment. It is difficult to anticipate what accessibility issues will arise when a test is administered on a variety of different systems with a variety of assistive technologies. Trying to anticipate these issues is important, however, when reviewing items.

Recommendations

These considerations can be used to make assessments more universally accessible to the entire population of test takers. Here are some specific recommendations for the use of the considerations of universal design at all stages of test development.

1. **Incorporate elements of universal design in the early stages of test development**. Universally designed assessments present an opportunity to bring more people to the table in the early stages of test development including experts in disability, language acquisition, and technology. These experts are able to give more structured input at different stages of the test development process if they understand universal design

and have these considerations for item development and review at hand. It is more cost effective to consider universal design in the early stages of item development, rather than at the end when items have already been developed and field-tested.

2. **Include disability, technology, and language acquisition experts in item reviews.** Every effort should be made to involve experts in item review who can judge whether items meet all of the considerations.

3. **Provide professional development for item developers and reviewers on use of the considerations for universal design**. Explanation and discussion of each consideration will ensure use by item developers and reviewers.

4. **Present the items being reviewed in the format in which they will appear on the test**. When item reviewers examine items to be included in an assessment, it is important to format items as closely as possible to how they will appear on the test. Since many of the considerations have to do with format, it is not useful to look at items that are not in the font, size, or format in which they will appear in the actual test booklet.

5. **Include standards being tested with the items being reviewed**. Above all other considerations, the first consideration—does the item measure what it intends to measure—is of primary importance in constructing universally designed assessments. Consequently, item review teams using the considerations of universal design to guide their work must have the standard (grade level expectations) that each item is intended to test at hand. It is only by knowing what an item is intended to test that reviewers can judge whether an element of the item might interfere with student access. Each item needs to be presented with the corresponding standard being tested in that item.

6. **Try out items with students**. Some of the elements of an item that distract or confuse students are not easily recognizable by adults or native English speakers. For this reason, trying items out with students by conducting think-aloud studies can provide valuable information about whether an item is testing the content intended (Thompson, Johnstone, & Miller, in press).

7. **Field test items in accommodated formats.** In order to ensure that the content an item is intended to measure is not being changed when an accommodated format of a test is being used, include students using accommodated test formats in field tests. While this can add additional expense to the field test, there are ways of doing such studies that can progressively build a database. For example, a field test could focus on the use of certain accommodated formats one year and others the next, building up a database for the various forms of the test. Again, qualitative data from student

interviews in this area can provide important information that can be used to improve items.

8. **Review computer-based items on computers.** To judge whether computer-based items are universally designed, item reviewers need to use the technology that will be used to deliver the test. Using a paper print-out of an assessment does not allow a review team to meaningfully consider the format of the test.

Conclusion

We hope that the process detailed in this report has produced not only a better set of considerations of universally designed assessments for all students, but has also clarified some of the opportunities and challenges that universally designed assessments present. While using universal design does not guarantee the accessibility of any test to all students, using the considerations to openly discuss issues of test design throughout the test development process can make any assessment more inclusive. Making the process of test development more transparent, informed, and focused on the needs of the entire population of students will help ensure that the assessment results are more meaningful for the widest range of students.

References

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: interactions with student language background*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Adler, M., & Ziglio, E. (Eds.). (1996). *Gazing into the Oracle: the Delphi method and its application to social policy and public health*. London: Jessica Kingsley Publishers.

Anderson, R.C., Hiebert, E.H., Scott, J.A., & Wilkinson, A.G. (1985). *Becoming a nation of readers*. Urbana, IL: University of Illinois, Center for the Study of Reading, National Institute of Education, National Academy of Education.

Arditi, A. (1999). *Making text legible*. New York: Lighthouse.

Assistive Technology Act of 2004 (Brief Title: ATA 2004). (P.L.108-364).

Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement, 32,* 323–340.

Brown, P.J. (1999). *Findings of the 1999 plain language field test*. Newark, DE: University of Delaware, Delaware Education Research and Development Center.

Calhoun, M.B., Fuchs, L., & Hamlett, C. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly, 23,* 271–282.

Carter, R., Dey, B., & Meggs, P. (1985). *Typographic design: Form and communication*. New York: Van Norstrand Reinhold.

Cole, C., Tindal, G., & Glasgow, A. (2000). *Final report: Inclusive comprehensive assessment system research, Delaware large scale assessment program*. Eugene, OR: Educational Research Associates.

Council for Exceptional Children (1999). *Universal design: Research connections*. Retrieved September 3, 2004, from the World Wide Web: http://ericec.org/osep/recon5/rc5sec1.html

Fuchs, L., Fuchs, D., Eaton, S., Hamlett, C., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67* (1), 67–92.

Gaster, L., & Clark, C. (1995). *A guide to providing alternate formats*. West Columbia, SC: Center for Rehabilitation Technology Services. (ERIC Document No. ED 405689)

Gregory, M., & Poulton, E.C. (1970). Even versus uneven right-hand margins and the rate of comprehension in reading. *Ergonomics, 13* (4), 427–434.

Grise, P., Beattie, S., & Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test modifications make a difference. *Journal of Educational Research, 76,* 35–40.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

Hanson, M.R. (1997). *Accessibility in large-scale testing: Identifying barriers to performance*. Delaware: Delaware Education Research and Development Center.

Hanson, M.R., Hayes, J.R., Schriver, K., LeMahieu, P.G., & Brown, P.J. (1998). *A plain language approach to the revision of test items*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 16, 1998.

Harker, J.K., & Feldt, L.S. (1993). A comparison of achievement test performance of nondisabled students under silent reading plus listening modes of administration. *Applied Measurement, 6,* 307–320.

Hartley, J. (1985). *Designing instructional text* (2nd Edition). London: Kogan Page.

Heines. (1984). *An examination of the literature on criterion-referenced and computer-assisted testing*. ERIC Document Number 116633.

Hoener, A., Salend, S., & Kay, S.I. (1997). Creating readable handouts, worksheets, overheads, tests, review materials, study guides, and homework assessments through effective typographic design. *Teaching Exceptional Children, 29,* (3), 32–35.

Individuals with Disabilities Educational Improvement Act (Brief Title: IDEA 2004). (P.L. 108-446).

Johnstone, C.J., Miller, N.A., & Thompson, S.J. (in press). *Using the think aloud method (cognitive labs) to evaluate test design*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington DC: Council of Chief State School Officers.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles, CA: Center for Research on Standards and Student Testing.

Lau vs. Nichols, 414 U.S. 563, 94 S.Ct. 786 (1974).

MacArthur, C.A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: handwriting, word processing, and dictation. *Journal of Special Education, 21* (3), 22-42.

Mace, R. (1998). *A perspective on universal design*. An edited excerpt of a presentation at Designing for the 21st Century: An International Conference on Universal Design. Retrieved January, 2002, from the World Wide Web: www.adaptenv.org/examples/ronmaceplenary98.asp?f=4.

Menlove, M., & Hammond, M. (1998). Meeting the demands of ADA, IDEA, and other disability legislation in the design, development, and delivery of instruction. *Journal of Technology and Teacher Education*. *6* (1), 75–85.

Muncer, S.J., Gorman, B.S., Gorman, S., & Bibel, D. (1986). Right is wrong: An examination of the effect of right justification on reading. *British Journal of Educational Technology, 1* (17), 5–10.

National Research Council. (1999). *High stakes: testing for tracking, promotion, and graduation*. In J. Heubert & R. Hauser (Eds.), Committee on Appropriate Test Use. Washington, DC: National Academy Press.

Osborne, H. (2001). In other words…communication across a life span…universal design in print and web-based communication. *On Call* (January). Retrieved January, 2002, from the World Wide Web: www.healthliteracy.com/oncalljan2001.html.

Popham, W.J. (2001). *The truth about testing: An educator's call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W.J., & Lindheim, E. (1980). The practical side of criterion-referenced test development. *NCME Measurement in Education, 10* (4), 1–8.

Rakow, S.J. & Gee, T.C. (1987). Test science, not reading. *Science Teacher, 54* (2), 28–31.

Schiffman, C.B. (1995). *Visually translating materials for ethnic populations*. Virginia: ERIC Document Number ED 391485.

Schriver, K.A. (1997). *Dynamics in document design*. New York: John Wiley & Sons, Inc.

Sharrocks-Taylor, D., & Hargreaves, M. (1999). Making it clear: A review of language issues in testing with special reference to the National Curriculum Mathematics Tests at Key Stage 2. *Educational Research, 41* (2), 123–136.

Silver, A.A. (1994). Biology of specific (developmental) learning disabilities. In N.J. Ellsworth, C.N. Hedley, & A.N. Barratta, (Eds.), *Literacy: A redefinition*. New Jersey: Erlbaum Associates.

Smith, J.M., & McCombs, M.E. (1971). Research in brief: The graphics of prose. *Visible Language, 5* (4), 365–369.

Szabo, M., & Kanuka, H. (1998). Effects of violating screen design principles of balance, unity, and focus on recall learning, study time, and completion rates. *Journal of Educational Multimedia and Hypermedia, 8* (1), 23–42.

Thompson, D.R. (1991). *Reading print media: The effects of justification and column rule on memory.* Paper presented at the Southwest Symposium, Southwest Education Council for Journalism and Mass Communication, Corpus Christi, TX. (ERIC Document Number 337 749)

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S.J., Johnstone, C.J., & Miller, N.A. (in press). *Universally designed assessments from the end user's perspective: Using a think aloud method* (Policy Directions). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study. *Exceptional Children, 64* (4), 439–450.

Tinker, M.A. (1963). *Legibility of print.* Ames, IA: Iowa State University Press.

Trotter, A. (2001). Testing computerized exams. *Education Week, 20* (37), 30–35.

West, T.G. (1997). *In the mind's eye: Visual thinkers, gifted people with dyslexia and other learning difficulties, computer images, and the ironies of creativity.* Amherst, NY: Prometheus Books.

Worden, E. (1991). *Ergonomics and literacy: More in common than you think.* Indiana. (ERIC Document Number 329 901)

Zachrisson, G. (1965). *Studies in the legibility of printed text.* Stockholm: Almqvist and Wiskell.

Appendix A

Delphi Review of Test Item Considerations (Form 1)

Rating scale for importance:

5=Extremely important to consider; 4=Very important to consider; 3=Important to consider; 2=Somewhat important to consider; 1=Not important to consider.

Scales adapted from Ziglio (1996).

Considerations when reviewing any test item:	Subject Responses	Mean	Please insert your comments here…
Does the item…			
Measure what it intends to measure • Reflects the intended content standards (reviewers have information about the content being measured) • Minimize skills required beyond those being measured			
Respect the diversity of the assessment population • Accessible to test takers (consider age, gender, ethnicity, and socio-economic level) • Avoids content that might unfairly advantage or disadvantage any student subgroup			
Have a clear format for text • Standard typeface • Type size appropriate for age group (12 point minimum for all print, including captions, footnotes, and graphs) • Wide spacing between letters, words, and lines • High contrast between color of text and background • Sufficient leading (blank space) between lines of text • Staggered right margins (no right justification)			

Have clear pictures and graphics (when essential to item) • Pictures are needed to respond to item • Pictures with clearly defined features • Dark lines (minimum use of gray scale and shading) • Sufficient contrast between colors • Color is not relied on to convey important information or distinctions • Pictures and graphs are labeled			
Have concise and readable text • Commonly used words • Vocabulary appropriate for grade level • Minimum use of unnecessary words • Idioms avoided unless idiomatic speech is being measured • Technical terms and abbreviations avoided (or defined) if not related to the content being measured • Sentence complexity is appropriate for grade level • Question to be answered is clearly identifiable			
Allow changes to its format without changing its meaning or difficulty (including visual or memory load) • Allows for the use of braille or other tactile format • Allows for signing to a student • Allows for the use of oral presentation to a student • Allows for the use of assistive technology • Allows for translation into another language			

Does the test...			
Have an overall appearance that is clean and organized • All images, pictures, and text provide information necessary to respond to the item • Information is organized in a manner that is consistent with an academic English framework with a left-right, top-bottom flow			
In addition to the other considerations, a computer-based test should have these considerations:			
Layout and design • Sufficient contrast between background and text and graphics for easy readability • Color is not relied on to convey important information or distinctions • Font size and color scheme can be easily modified (through browser settings, style sheets, or on-screen options) • Stimulus and response options are viewable on one screen when possible • Page layout is consistent throughout the test			
Navigation • Navigation is clear and intuitive; it makes sense and is easy to figure out • Navigation and response selection is possible by mouse click or keyboard • Option to return to items and return to place in test after breaks			
Screen reader considerations • Item is intelligible when read by a text/screen reader • Links make sense when read out of visual context. (“go to the next question” rather than “click here”) • Non-text elements have a text equivalent or description			

• Tables are only used to contain data and make sense when read by screen reader			
Test specific options • Access to other functions is restricted (e.g. e-mail, Internet, instant messaging) • Pop up translations and definitions of key words/phrases are available if appropriate to the test			
Computer capabilities • Adjustable volume • Speech recognition available (to convert user's speech to text) • Test is compatible with current screen reader software			

Items on this form are based on information presented in Thompson, Johnstone, & Thurlow (2002, *Universal Design Applied to Large Scale Assessments*, Synthesis Report 44); Thompson & Thurlow 2002, *Universally Designed Assessments: Better Tests for Everyone!*, Policy Directions 14), and Kopriva (2002, *Ensuring Accuracy in Testing for English Language Learners*, CCSSO SCASS-LEP Consortium), as well from NCEO staff brainstorming and input received from participants in the Universal Design Pre-conference Clinic at the CCSSO Large Scale Assessment and Accountability Conference in San Antonio, Texas, June 2003 and input from a joint project/Delphi review with the Minnesota, Nevada, and South Carolina Departments of Education.

Appendix B

Delphi Review of Test Item Considerations (Form 2)

Rating scale for importance:

5=Extremely important to consider; 4=Very important to consider; 3=Important to consider; 2=Somewhat important to consider; 1=Not important to consider.

Scales adapted from Ziglio (1996).

Considerations when reviewing any test item:	Subject Responses	Mean	Please insert your comments here…
Does the item…			
Measure what it intends to measure			1-Ultimately this (is the) most important. Some of this can be accomplished within the concept of universal design…there may be content, knowledge, skills, and/or abilities that may not lend themselves well to UD in some instances.
• Reflects the intended content standards (reviewers have information about the content being measured)	55555555555	**5**	1-Agree
Should this be changed to read: "Reflects the intended construct (reviewers have information about the construct being measured)?	N/A	N/A	

• Minimize skills required beyond those being measured	334444555555	**4.33**	1-While I think you could argue for a "5" for the 2nd statement here, I worry that if we try to separate skills too much, we'll wind up with tests that measure isolated, basic skills. 2- Both cognitive demand and specific content should be included in the review. Content to content match is insufficient (see Conserving Math Construct (CMC) template). 3-I would rather it said "Minimizes skills required beyond those explicitly in the standard." 4-Very good observation. Relevant to higher level thinking. 5-Multiple response options should be available, perhaps as a pretest assessment. 6-Important but cannot override ability to measure all content areas to assessed (e.g., draw a graph of the results). 7-Crucial for test validity for deaf test takers. 8-This may be difficult with disabled students—but try. 9-Perhaps should be reworded as "Minimize skills required beyond those intended for measurement." Assure measurement of the intended construct. Assumptions about level of achievement possible/abilities (language, sensory, motor, background knowledge, etc.) can interfere with the ability to accurately measure the intended construct.
Based on these comments, would you change the wording of this consideration. If so, what should it say?	N/A	N/A	
Respect the diversity of the assessment population • Accessible to test takers (consider age, gender, ethnicity, and socio-economic level)	444555555555	**4.75**	
Should we add issues such as: Regional differences? Students with disabilities? Language minority students? Others? **Note: This section relates to what bias review committees typically address. Does the word "bias" need to be included?**	N/A	N/A	

• Avoids content that might unfairly advantage or disadvantage any student subgroup	44445555555	**4.64**	
Have a clear format for text • Standard typeface	33334445555	**4**	1-What is meant by "standard?" If you mean the same throughout a form, then 1. If you mean an acceptable typeface for clarity, then 4. 2-"Standard typeface" does not communicate in a world where everyone has access to hundreds of fonts—specify fonts and criteria re: serifs. 3-To do otherwise would introduce construct irrelevant variance (CIV). 4-Tests may be different between but should be the same within. 5-This should be "selection of typeface," wherein the best proven typefaces are used. 6-General comment for all of these points: no one set of values will work for all students. All we can hope for is a reasonable compromise unless we embed flexibility, as per universal design principles. Perhaps explicitly state here that this consideration addresses print materials exclusively. 7-Serif recommended for print; sans-serif recommended for computer displays. For large-print booklets, consider specific fonts (see APH guidelines: http://sun1.aph.org/edresearch/lpguide.htm)
Instead of the word "standard" should we use the word "common," "familiar," or "clear?" Please comment. **Would you change the wording entirely for the above consideration? If so, how?**	N/A	N/A	
• Twelve (12) point minimum for all print, including captions, footnotes, and graphs (type size appropriate for age group)	33444444555	**4.09**	
Additional comments on print size?	N/A	N/A	

• Wide spacing between letters, words, and lines	22233333445	**3.09**	1-Dependent on age. 2-To do otherwise would introduce construct-irrelevant variance (CIV). 3-Wide spacing is not necessarily best; proper font selection is more important. 4-Need to define precisely. Also consider that wide spacing may have deleterious effect on non-visually impaired students.
Should we retain the above consideration?	N/A	N/A	
• High contrast between color of text and background	33344445555	**4.09**	1-Also is one needed on color intensity? 2-And selection of paper/ink color. 3-Even with sufficient color contrast, color blind users may not be able to distinguish text and background. Suggest you further recommend high *print density* contrast. This would also avoid isoluminance effects for non-visually-impaired students.
Additional comments on color and background?			
• Sufficient leading (blank space) between lines of text	23334444455	**2.82**	1-After having just had a conversation on this, I'll up my rating to 3. 2-"Standard typographic leading (blank spaces) between lines of text."
• Staggered right margins (no right justification)	22333334455	**3.36**	
Have clear pictures and graphics (when essential to item)			
• Pictures are needed to respond to item	335555555	**4.56**	1-Does this mean how important is it for there to be picture-based items? In some speaking tests, pictoral prompts are often the best way of getting students to produce language (without giving them directions and thus affecting their responses). 2-Deaf examinees tend to focus on pictures when provided, more than on text. A picture should be related to the correct response, not be distraction to "trick" certain students (more than others). 3-Pictures, line art, etc. should be related to the item, should enhance understanding not required for understanding with the exception of data tables like in math and science tests. 4-Perhaps reword as "only construct-relevant visuals."

Should we change consideration to read: "When pictures are used, they should be redundant with the text as much as possible."?	N/A	N/A	
• Pictures with clearly defined features	44444455555	**4.45**	1-As we noted previously, this seems like catering to the lowest common denominator. Might there be instances where grayscale and shading are most appropriate for providing construct-relevant information to the student? If so, then alternate representations would need to be considered for visually impaired students.
• Dark lines (minimum use of gray scale and shading)	33334444455	**3.82**	1-Pictures with clearly defined features" covers this point. This now seems redundant and could be removed.
• Sufficient contrast between colors	13333444555	**3.64**	1-Again, color may be a different and additional issue for computer based assessments. 2-"Color alone is not relied on to convey important information or distinctions."
• Color is not relied on to convey important information or distinctions	23334445555	**3.91**	1-This is perhaps too vague to be useful as a guideline.
• Pictures and graphs are labeled	33334444555	**3.91**	
Have concise and readable text			
• Commonly used words	12445555555	**4.18**	1-Unless it is a vocabulary test. 2-Commonly used words are not necessarily easy…they can carry multiple meanings. 3-What does this mean in a "content" assessment? 4-Again phrasing is inconsistent. This is not appropriate for vocabulary/ language assessment. 5-If the construct is about unique words, then common words might make sense. 6-Important, but highly subjective. 7-That is, use common meaning, not low-frequency meaning for common words. 8-This can be collapsed into the second item: age level vocabulary. 9-"Commonly used words" seems dependent on the construct being measured, e.g., vocabulary instruction. Can this point be combined with the next one?

Should we re-word the above consideration? If so, what should it say?	N/A	N/A	
• Vocabulary appropriate for grade level	445555555555	**4.83**	1-I think both this and the next bullet could easily be combined with the first. 2-Depends on what you are measuring…
• Minimum use of unnecessary words	124445555555	**4.17**	
• Idioms avoided unless idiomatic speech is being measured	344555555555	**4.67**	
• Technical terms and abbreviations avoided (or defined) if not related to the constructs being measured	44455555555	**4.73**	
• Sentence complexity is appropriate for grade level	14455555555	**4.45**	
• Question to be answered is clearly identifiable	555555555555	**5**	1-Refer to the work and writing of Jamal Abedi at UCLA regarding language simplification as it affects testing of LEP students. Very critical work and he has many concrete suggestions about item writing and presentation that are consistent with our AERA, APA, NCME Standards for Educational and Psychological Testing.
Allow changes to its format without changing its meaning or difficulty (including visual or memory load)			1-The change in meaning is going to be speculative until the research is done, and maybe not even then.
• Allows for the use of braille or other tactile format	34455555555	**4.67**	1-Not all items can be brailled or made tactile in a meaningful way. Would this mean that such items cannot be used for any students? What's wrong with having equivalent items that can be so modified?
• Allows for signing to a student	33455555555	**4.55**	
• Allows for the use of oral presentation to a student	33345555555	**4.36**	
• Allows for the use of assistive technology	33445555555	**4.45**	

• Allows for translation into another language	12333355555	**3.64**	1-Validity may be compromised as bias is interjected. 2-Some critical and relevant words/concepts do not translate into every language. I think we might have to narrow to a few core languages. 3-This is a questionable and highly controversial issue, particularly when one realizes that such a standard is impossible to meet. About 72% of our LEP students are Spanish speakers, but the other 28% represent many diverse languages. How do we accommodate and what is the theoretical rationale and what is the technology for doing this? Is it possible? Is it beneficial? 4-Most likely invalidates the test scores for high stakes decision, should probably assess in an alternate fashion that more validly assesses students' ability, skill, knowledge, or learning. 5-A sticky wicket… 6-What about ELA tests? Also, what if the student prefers seeing the question in both their native language and the language of instruction? Should we deny them this opportunity? 7-There is the continual challenge of literal interpretation in translation, and limitations in expertise to translate to all needed languages. So, do students who speak Spanish receive interpreted tests, and those speaking a less well known language do not? 8-This issue will be one of the most difficult to overcome on a political and emotional level. Data analysis has shown these to be less than valid measures of student performance. Additionally, it has been suggested that it takes approximately 5 years for a student to become proficient in English.
Additional comments on this consideration?	N/A	N/A	
Does the test...			
Have an overall appearance that is clean and organized?			
• All images, pictures, and text provide information necessary to respond to the item	344445555555	**4.5**	1-Redundant with "pictures are needed to respond to item" and "minimum use of unnecessary words." 2-I still rate this as a 5!
• Information is organized in a manner that is consistent with an academic English framework with a left-right, top-bottom flow	444444445555	**4.33**	

In addition to the other considerations, a computer-based test should have these considerations:			
Layout and design			
• Sufficient contrast between background and text and graphics for easy readability	444455555555	**4.67**	1-"Sufficient luminance contrast between…"
• Color is not relied on to convey important information or distinctions	233334455555	**3.92**	1-"Color alone is not relied…"
• Font size and color scheme can be easily modified (through browser settings, style sheets, or on-screen options)	233344555555	**4.08**	
• Stimulus and response options are viewable on one screen when possible	344555555555	**4.67**	
• Page layout is consistent throughout the test	444555555555	**4.75**	
Navigation			
• Navigation is clear and intuitive; it makes sense and is easy to figure out	455555555555	**4.92**	
• Navigation and response selection is possible by mouse click or keyboard	344555555555	**4.67**	1-"Navigation and response selection is possible by mouse (or equivalent) or keyboard (or equivalent)."
• Option to return to items and return to place in test after breaks	3445555555	**4.6**	1-As before, returning to items will only work in non-adaptive tests.
Screen reader considerations			
• Item is intelligible when read by a text/screen reader	344455555555	**4.58**	
• Links make sense when read out of visual context ("go to the next question" rather than "click here")	444455555555	**4.67**	
• Non-text elements have a text equivalent or description	3334555555	**4.3**	

• Tables are only used to contain data, and make sense when read by screen reader	34444455555	**4.36**	
Test specific options			
• Access to other functions is restricted (e.g., e-mail, Internet, instant messaging)	33455555555	**4.55**	1-“Access to functions other than assistive technologies and supports is restricted (e.g., e-mail, Internet, instant messaging).”
• Pop up translations and definitions of key words/ phrases are available if appropriate to the test	333444445555	**4.08**	
Computer capabilities			
• Adjustable volume	334455555555	**4.5**	1-“Adjustable volume and rate of voice.”
• Speech recognition available (to convert user’s speech to text)	112334555555	**3.67**	
• Test is compatible with current screen reader software	334444455555	**4.25**	
Below are new considerations suggested by panelists on the first survey, please rank each consideration and comment if you wish (numeric rankings found below were provided by participants in last survey).			
Computer interfaces follow Section 508 guidelines (www.section508.gov)			
Students have received adequate training on use of test delivery system			

Students writing online can get feedback on length of writing on-demand in cases where there is a restriction on number of words			
Students are able to record their responses and read them back (or have them read-back using text-to-speech) as alternative to human scribe, but only if student has experience with this mode of expression and chooses it for the test			
Students are allowed to create persistent marks to the extent that they are already allowed to on paper-based booklets (e.g., marking items for review; eliminating multiple choice items, etc.)			
Alternate versions of computer interface provided that is amenable to use with screen readers (e.g., JAWS, Window-Eyes)			
When "modifying" items for use under a UD framework, there is convincing evidence that the item still measures the same or similar intended construct	4		Obviously, since I wrote these, I think they are all important. I didn't rate them as "5" because I think they are very hard to do, but I also think we need to be working in this direction.
A test constructed under a UD framework allows for the measurement of the same depth of knowledge levels as the "original" test	4		
A UD test is aligned to the standards to the same extent as the "original" test	4		

Test items are piloted, field tested, and normed on all subgroups for which the measure is designed	5		
Booklets/materials can be easily handled with limited motor coordination			
Response formats are easily correlated to question			
Computer-based option to mask items or text (e.g., split screen)	4		
Place for student to take notes (on the screen for CBT) or extra white space with paper-pencil	4		
Computer software for test delivery is designed to be amenable to all assistive technology	5		

Items on this form are based on information presented in Thompson, Johnstone, & Thurlow (2002, *Universal Design Applied to Large Scale Assessments*, Synthesis Report 44); Thompson & Thurlow 2002, *Universally Designed Assessments: Better Tests for Everyone!*, Policy Directions 14), and Kopriva (2002, *Ensuring Accuracy in Testing for English Language Learners*, CCSSO SCASS-LEP Consortium), as well from NCEO staff brainstorming and input received from participants in the Universal Design Pre-conference Clinic at the CCSSO Large Scale Assessment and Accountability Conference in San Antonio, Texas, June 2003 and input from a joint project/Delphi review with the Minnesota, Nevada, and South Carolina Departments of Education.

Appendix C

Original Considerations Plus All Expert Commentary

Delphi Review of Test Item Considerations

Rating scale for importance:

5=Extremely important to consider; 4=Very important to consider; 3=Important to consider; 2=Somewhat important to consider; 1=Not important to consider.

Scales adapted from Ziglio (1996).

Considerations when reviewing any test item:	Subject Responses	Mean	Please insert your comments here…
Does the item…			
Measure what it intends to measure			1-Ultimately this (is the) most important. Some of this can be accomplished within the concept of universal design…there may be content, knowledge, skills, and/or abilities that may not lend themselves well to UD in some instances.
• Reflects the intended content standards (reviewers have information about the content being measured)	555555555555	**5**	1-Agree

Should this be changed to read: "Reflects the intended construct (reviewers have information about the construct being measured)?	N/A	N/A	1-Yes. 2-Either way. 3-How about combining the 2 ideas? Reflects the intended construct that is aligned with representative (language proficiency or academic content) standards? 4-Construct is a formal term that theorists use. Content standards is what practitioners understand. 5-Yes. 6-Yes, Definitely. 7-No—but you might add something about the cognitive demand 8-This revision would alter the meaning. I think the construct is a sort of overarching concept (i.e., reading) whereas content standards are quite narrower (i.e., reproduces capital letters). It depends what you mean as to which one you should use. If the test is supposed to be a standards-based achievement test, then it must address standards. If no, then the item need only address the construct. 9-Probably content is the correct phrase…construct is typically Reading or Math, etc. The content is what reviewer typically know and evaluate. 10-No. I like better as it was. 11-Yes. That would fit better with the professional terminology. 12-Yes. 13-Yes, I believe it should be changed. Content is topical, constructs are conceptual. This difference In meaning is huge. Furthermore, constructs is a term used in APA standards and is deeper than content.

• Minimize skills required beyond those being measured	334444555555	**4.33**	1-While I think you could argue for a "5" for the 2nd statement here, I worry that if we try to separate skills too much, we'll wind up with tests that measure isolated, basic skills. 2-Both cognitive demand and specific content should be included in the review. Content to content match is insufficient (see Conserving Math Construct [CMC] template). 3-I would rather it said "minimizes skills required beyond those explicitly in the standard." 4-Very good observation. Relevant to higher level thinking. 5-Multiple response options should be available, perhaps as a pretest assessment. 6-Important but cannot override ability to measure all content areas to assessed (e.g., draw a graph of the results). 7-Crucial for test validity for deaf test takers. 8-This may be difficult with disabled students—but try. 9-Perhaps should be reworded as "Minimize skills required beyond those intended for measurement." Assure measurement of the intended construct. Assumptions about level of achievement possible/ abilities (language, sensory, motor, background knowledge, etc.) can interfere with the ability to accurately measure the intended construct. R1-How might the other skills be defined and targeted? It seems if the first part is clear and clearly defined (that is the construct is being measured), this should not be as much of an issue. My fear is that folks will take this beyond content and eliminate all attractive, yet non-distracting qualities of the assessment. R2-There may still be a limit in what can be made accessible and probably varies by which group is evaluated. R3-I think the issue is about "construct irrelevant variance" and furthermore, the way it is stated it implies that skills are being measured. Major changes reflect the comments of #9 above.

Based on these comments, would you change the wording of this consideration. If so, what should it say?	N/A	N/A	1-I could live with either statement 3 (if "standard" was made plural) or 9. 2-I'd probably make it a 4 instead of a 5. 3-Depends on the purpose of the test….a diagnostic test is more skills-driven; if standards-based, depends on how discrete the standards are; minimal skills can be embedded in more complex, contextualized items. Ultimately, it depends on what you are measuring. 4-I think the meaning of the phrase (as I interpret it) is test what you are supposed to test and not construct-irrelevant factors. 5-"Minimize skills required beyond those intended for measurement of a standard." 6- Yes. I would suggest changing the question/consideration to something like "does the item measure specific and carefully defined skills?" 7- The comments make me think that this consideration was interpreted very differently by the reviewers. Perhaps you should find out how this was interpreted. I interpreted this to mean: the items on the assessment should clearly stick to the domain being assessed. If that is the case, I give it a 5. 8-No change needed. 9-No. 10-I like reviewer comments 3 and 9 to get at the intended consideration. 11-I might say something like the following, "we understand that when measuring some skills, especially higher-order skills, it is difficult to focus the assessment items only on the explicitly targeted skills. However, please rate the importance of minimizing the skills required beyond those being measured." 12-I would change it to something referencing "access" skills that are limited to input-output sensory modalities: minimize interference from sensory input and output skills needed to respond."

Respect the diversity of the assessment population • Accessible to test takers (consider age, gender, ethnicity, and socio-economic level)	444555555555	**4.75**	
Should we add issues such as: regional differences? Students with disabilities? Language minority students? Others? **Note: This section relates to what bias review committees typically address. Does the word "bias" need to be included?**	N/A	N/A	1-Not necessarily, but change wording to "accessible to all test takers, regardless of such characteristics as…. Include a partial list. (Next to the Note: he wrote "No.") 2-Yes, bias should be included. I believe bias and accessib[ility] are separate issues from a review standpoint, though obviously related. For instance, "bias" in terms of the diverse sensitivity that is commonly reviewed in committees includes sensitivity towards terms that are sensitive to the deaf community–listen, hear, sounds, etc., and to the VI community–look, see, visualize, shadows, etc.. To review items from accessibility perspective requires a different perspective, beyond words or terms used in the item to the structure, layout, look, and feel of the item, first; and second, whether or not students have the intellectual capacity to access the items (opportunity to learn); and third, whether or not the test and items are able to access the information and measure the constructs that are intended. 3-Yes. I think there are two issues here: 1. bias and 2. subgroup representation. Every test is biased and yes, it needs to be recognized…UD hopefully serves to minimize its impact. English language learners are a subset of language minority students; different groups with different educational needs. 4-No. Regional differences are construct-irrelevant in our world. I think the term "sensitivity" is a better term here. Mike Zieky of ETS has written extensively on this issue. Jamal Abedi has written and researched the language issue. Bias is a different idea and shouldn't be used here. 5-Definitely add the disabilities and ELL issues with regard to accessibility. Bias in relation to language and ethnicity should be a part of the bias review. 6-Yes. We should include the word "bias" and phrase it something like "does the item unfairly affect or is the item biased against students with disabilities, ESL speakers, students from different regions, SES, etc." See also wording in the point that follows this one.

			7-While the effort is to make assessments more accessible, I am very afraid that trying to meet the needs of multiple audiences has the potential of stripping down the assessments too much. 8-Yes, the inclusion of the word "bias" would be helpful. 9-Yes on all. 10-Include students with disabilities and language minority students. 11-Yes, I would add SWD and LEP students. No, the word "bias" does not need to be included. 12-For sure, poverty should be explicitly addressed. I think regional differences may not be so important. Language differences are very important. Bias should be included and reference systematic variance that interferes with making a valid inference.
• Avoids content that might unfairly advantage or disadvantage any student subgroup	44445555555	**4.64**	
Have a clear format for text • Standard typeface	33334445555	**4**	1-What is meant by "standard?" If you mean the same throughout a form, then 1. If you mean an acceptable typeface for clarity, then 4. 2-"Standard typeface" does not communicate in a world where everyone has access to hundreds of fonts—specify fonts and criteria re: serifs. 3-To do otherwise would introduce construct irrelevant variance (CIV). 4-Tests may be different between but should be the same within. 5-This should be "selection of typeface," wherein the best proven typefaces are used. 6-General comment for all of these points: no one set of values will work for all students. All we can hope for is a reasonable compromise unless we embed flexibility, as per universal design principles. Perhaps explicitly state here that this consideration addresses print materials exclusively. 7-Serif recommended for print; sans-serif recommended for computer displays. For large-print booklets, consider specific fonts (see APH guidelines: http://sun1.aph.org/edresearch/lpguide.htm). 8-R to #5—I am not familiar with and would like to hear about the best "proven" typefaces . . . 9-R to #7—We must be careful about "recommendations" without some research to support them.

Instead of the word "standard" should we use the word "common?" "familiar?" "clear?" Please comment. **Would you change the wording entirely for the above consideration? If so, how?**	N/A	N/A	1-Now I realize that there are two issues: "standard" means both "consistent" and "common or clear." Both are needed. 2-In my experience, there are so many fonts and opinions about which to use, all with no research to support that one is necessarily better than the other for all types of readers. There may be 20 or so commonly used fonts and many will argue they are clear; and of course a 12 pt in one is not a 12 pt in the other. So, I'm not sure any one of these—common, standard, familiar, or clear—is any better than the other, unfortunately. You may need to provide examples for guidance . . .(This is a major source of problems in the publishing arena. There may be multiple fonts in one test booklet, all of which are considered "clear." We have at least tried to keep the font chosen consistent within a book. However, for captions and such where emphasis is required, we will often opt for a different font instead of going to a smaller size or the use of italics, often used approaches in text books.) 3-Why don't you include it in the next consideration…I don't think typeface and font size need to be 2 separate ones. 4-Standardized typeface. I think we all understand that a student should never be confused over type face. 5-While these terms may be more flexible, they are also somewhat subjective. The key is to have the typeface appropriate for the intended audience, and we must realize that the audience is very diverse. 6-Yes. There is no standard typeface, thus the myriad fonts used in various publisher's files, even within the same test or textbook. I suggest changing the wording to "does the item have a readable format and follow the guidelines for optimal readability according to current research." Then we must equip the reviewer with those guidelines. 7-As before, this stuff is all logical. Of course, keep the type face and format accessible. I would hate (to) see a recommendation for one type face et al. 8-I prefer "clear" to "standard," as I agree with comment 2 above. 9-Use clear. No. 10- I like comments 5 and 7 as guidance here. Changed wording as with comment 5. 11-The term doesn't matter as long as you define what you mean. Make sure you do that.

			12-This may be a problem as standardized is what you want to avoid; What about a limited range of variation that has common features of typeface (font type size of font, leading, kerning, and serif)? Clear and familiar beg the question (according to who?)
• Twelve (12) point minimum for all print, including captions, footnotes, and graphs (Type size appropriate for age group)	33444444555	**4.09**	
Additional comments on print size?	N/A	N/A	1-No. 2-Again, what is appropriate? And 12 pt. varies with font types. 3-Appropriateness of size contingent on disability, if applicable. 4-While general recommendations for print size are reasonable, the bottom line again is that some readers may be unable to see that print clearly. IF the design from the beginning is flexible then print outs could be made available to fit the needs of individuals. 5-Yes. There are guidelines which apply to readers with both regular vision and low vision. These can be found at www.aph.org. 6-N/A 7- Most important thing is that people can get an alternative font size. 8- Not by me. 9- No.
• Wide spacing between letters, words, and lines	22233333445	**3.09**	1-Dependent on age. 2-To do otherwise would introduce construct-irrelevant variance (CIV). 3-Wide spacing is not necessarily best; proper font selection is more important. 4-Need to define precisely. Also consider that wide spacing may have deleterious effect on non-visually impaired students.
Should we retain the above consideration?	N/A	N/A	1-No. 2-I just don't know. 3-No. 1. It's too subjective. 2. Content should be weighted over format. 4-Yes. 5-Isn't there an established standard that would transfer across font size? 6-No. Unless is it redefined. 7-Yes. 8-Yes. 9-I defer to other reviewers here. 10-Yes. 11-Not necessarily. Furthermore, this spacing may be developmental (e.g., more space for younger students).

• High contrast between color of text and background	33344445555	**4.09**	1-Also is one needed on color intensity? 2-And selection of paper/ink color. 3-Even with sufficient color contrast, color blind users may not be able to distinguish text and background. Suggest you further recommend high *print density* contrast. This would also avoid isoluminance effects for non-visually-impaired students.
Additional comments on color and background?			1-Agree with comment 3. 2-This will be different for computer screens. Some colors are better than others on screen. For instance, I think blue is not so good on screen. 3-No. 4-"Sufficient color contrast *and* print density contrast between text and background." 5-Again, equip reviewer with current research and examples. 6-N/A 7-None. 8-I defer to other reviewers here. 9-No. 10-Focus on maximum contrast, not color per se.
• Sufficient leading (blank space) between lines of text	23334444455	**2.82**	1-After having just had a conversation on this, I'll up my rating to 3. 2-"Standard typographic leading (blank spaces) between lines of text."
• Staggered right margins (no right justification)	22333334455	**3.36**	
Have clear pictures and graphics (when essential to item)			
• Pictures are needed to respond to item	335555555	**4.56**	1-Does this mean how important is it for there to be picture-based items? In some speaking tests, pictoral prompts are often the best way of getting students to produce language (without giving them directions and thus affecting their responses). 2-Deaf examinees tend to focus on pictures when provided, more than on text. A picture should be related to correct response and not be distraction to "trick" certain students (more than others). 3-Pictures, line art, etc. should be related to the item, should enhance understanding not required for understanding with the exception of data tables like in math and science tests. 4-Perhaps reword as "only construct-relevant visuals."

Should we change consideration to read: "When pictures are used, they should be redundant with the text as much as possible."?	N/A	N/A	1-No. However, this is one where a great deal of work is needed to arrive at universal criteria. I do not have the answer. 2-Often there are pictures used that are not redundant with the text but that are relevant to the item and to the construct. I'd be careful with the rewording. 3-Pictures, when used, should support text and provide an additional resource for students to construct meaning. 4-Providing redundancy of figures in the text might provide inappropriate clues and invalidate items (e.g., interpreting graphs). Still feel "only construct-relevant visuals" is best. 5-Yes. Specify what is meant by "clear." Does this refer to standards for well-designed "graphics?" The pictures should be value-added for students needing a visual cue. This also works for the person who will not benefit from pictures, (i.e., blind and perhaps low vision students.) 6-Pictures, diagrams, [etc.] should be included [if] they are integral to the question being asked. 7-No; I disagree with this re-write because it does not encompass the idea in comment 1 above. I like the suggested rewording in comment 4— "only construct-relevant visuals" because it seems to cover all the bases. 8-Yes, but change "they should be redundant" to "the content should be redundant." 9-Yes. See comments 2 and 4. 10-I wouldn't say "redundant," rather I'd say, be connected to the text as much as possible. 11-Yes, so the issue of multiple opportunities is clear. The pictures must also be consistent with the text.
• Pictures with clearly defined features	44444455555	**4.45**	1-As we noted previously, this seems like catering to the lowest common denominator. Might there be instances where grayscale and shading are most appropriate for providing construct-relevant information to the student? If so, then alternate representations would need to be considered for visually impaired students.
• Dark lines (minimum use of gray scale and shading)	33334444455	**3.82**	1-Pictures with "clearly defined features" covers this point This now seems redundant and could be removed.
• Sufficient contrast between colors	13333444555	**3.64**	1-Again, color may be a different and additional issue for computer based assessments. 2-"Color alone is not relied on to convey important information or distinctions."

• Color is not relied on to convey important information or distinctions	23334445555	**3.91**	1-This is perhaps too vague to be useful as a guideline.
• Pictures and graphs are labeled	33334444555	**3.91**	
Have concise and readable text			
• Commonly used words	12445555555	**4.18**	1-Unless it is a vocabulary test. 2-Commonly used words are not necessarily easy…they can carry multiple meanings. 3-What does this mean in a "content" assessment? 4-Again phrasing is inconsistent. This is not appropriate for vocabulary/language assessment. 5-If the construct is about unique words, then common words might make sense. 6-Important, but highly subjective. 7-That is, use common meaning, not low-frequency meaning for common words. 8-This can be collapsed into the second item: age level vocabulary. 9-"Commonly used words" seems dependent on the construct being measured, e.g., vocabulary instruction. Can this point be combined with the next one?
Should we re-word the above consideration? If so, what should it say?	N/A	N/A	1-Eliminate (agree with #8). 2-I think it should read something like "other than the terms relevant to the construct being measured, items should use simplified language and vocabulary (as they are different), and might even be one grade level below the grade being tested." 3-Use words appropriate for the age of the students and that reflect the purpose of the item. 4-"Age-appropriate language (e.g., vocabulary, sentence complexity) when language is not central to the intended construct." 5-Yes. I am not sure of the best wording, but word the consideration so that "concise and readable" are clearly defined and understood. 6-Please refer to the materials that I sent earlier. The issues about the meaning of "common words" in a content assessment need to be more directly addressed. 7-Merge with consideration below. 8-With the exception of subject-specific terminology, the text of an item should consist of commonly understood words or vocabulary consistent with each specific grade level.

			9-Drop this one. Covered in next item adequately. 10-See comments 2 and 7. Yes, reword the consideration to be consistent with those comments. 11-I think the bullets help explain the question just fine. 12-Commonly used words may be replaced by reference to words most frequently appearing in similar text or listed as the primary definition in the dictionary.
• Vocabulary appropriate for grade level	445555555555	**4.83**	1-I think both this and the next bullet could easily be combined with the first. 2-Depends on what you are measuring…
• Minimum use of unnecessary words	124445555555	**4.17**	
• Idioms avoided unless idiomatic speech is being measured	344555555555	**4.67**	
• Technical terms and abbreviations avoided (or defined) if not related to the content being measured	44455555555	**4.73**	
• Sentence complexity is appropriate for grade level	14455555555	**4.45**	
• Question to be answered is clearly identifiable	555555555555	**5**	1-Refer to the work and writing of Jamal Abedi at UCLA regarding language simplification as it affects testing of LEP students. Very critical work and he has many concrete suggestions about item writing and presentation that are consistent with our AERA, APA, NCME Standards for Educational and Psychological Testing.
Allow changes to its format without changing its meaning or difficulty (including visual or memory load)			1-The change in meaning is going to be speculative until the research is done, and maybe not even then.
• Allows for the use of braille or other tactile format	34455555555	**4.67**	1-Not all items can be brailled or made tactile in a meaningful way. Would this mean that such items cannot be used for any students? What's wrong with having equivalent items that can be so modified?
• Allows for signing to a student	33455555555	**4.55**	
• Allows for the use of oral presentation to a student	33345555555	**4.36**	
• Allows for the use of assistive technology	33445555555	**4.45**	

• Allows for translation into another language	12333355555	**3.64**	1-Validity may be compromised as bias is interjected. 2-Some critical and relevant words/concepts do not translate into every language. I think we might have to narrow to a few core languages. 3-This is a questionable and highly controversial issue, particularly when one realizes that such a standard is impossible to meet. About 72% of our LEP students are Spanish speakers, but the other 28% represent many diverse languages. How do we accommodate and what is the theoretical rationale and what is the technology for doing this? Is it possible? Is it beneficial? 4-Most likely invalidates the test scores for high stakes decision, should probably assess in an alternate fashion that more validly assess students ability, skill, knowledge, or learning. 5-A sticky wicket… 6-What about ELA tests? Also, what if the student prefers seeing the question in both their native language and the language of instruction? Should we deny them this opportunity? 7-There is the continual challenge of literal interpretation in translation, and limitations in expertise to translate to all needed languages. So, do students who speak Spanish receive interpreted tests, and those speaking a less well known language do not? 8-This issue will be one of the most difficult to overcome on a political and emotional level. Data analysis has shown these to be less than valid measures of student performance. Additionally, it has been suggested that it takes approximately 5 years for a student to become proficient in English.

Additional comments on this consideration?	N/A	N/A	1-There will be no easy answer to this one. Perhaps drop it and just not take a stand or list it later as an unresolved issue. 2-Seems to me if the terms used are as I've described above, there may be some lessening of the arguments here, particularly if accommodations such as customize dictionaries are allowed. Also, translations into Spanish, even, are not consistent. I know Jamal Abedi has run experiments where he has asked 2 separate translation companies to translate text from English into Spanish and back, finding the companies vary greatly in their translations to, supposedly, the same Spanish dialect. 3-Depends on what you are measuring… English language arts should not be subject to translation as it is idiosyncractic but universal content, anchored in standards, may be feasible if it undergoes content and bias reviews. 4-Don't all items inherently allow for this already, in theory? 5-Perhaps we should decide to exclude ESL speakers from this until we have more research? 6-1-If the original test is valid as a measure of the curriculum, the translation will share that validity. 6-2-Minnesota translates to Hmong and Somali. Only in these languages are their relevant words/concepts that do not translate easily into English. The other languages of state assessment (Spanish, Russian, Chinese, Korean, Haitian Creole) almost never pose a problem for translating words or concepts. Professional translators will tell you they can translate almost any word or idea, and if they encounter one they can't, they will tell you that too. 6-3-What harm is done by helping the 72% of LEP students who speak Spanish? We provide accommodations to others where possible, but some would propose that a translated test is harmful. Poppycock! 6-4-A translated test is always much less of a threat to validity and score comparability than an alternate assessment. 6-5-ELA tests should not be translated. Once they are translated, they are no longer a test of ELA. A bilingual booklet is an appropriate accommodation for all subjects except ELA.

			6-6-There is not a limitation in expertise to translate tests. The limitation is in money. Translation must be cost effective like everything else in education. You can't provide translated tests for very small numbers. The Lao decision and other civil rights decisions make it clear that numbers dictate expectations of school systems. A few psychometricians have maligned test translation unnecessarily. One psychometrician says something and most of the rest follow like sheep. Translation is no more a threat to validity than a change in option order or a change in font. Such changes might generate a miniscule change in item difficulty, but they don't affect validity. Neither does translation, which is the exact same test stated in a different language. 7-No. 8-None. 9-Translations need to be cognizant of dominant dialects.
Does the test...			
Have an overall appearance that is clean and organized?			
• All images, pictures, and text provide information necessary to respond to the item	344445555555	**4.5**	1-Redundant with "pictures are needed to respond to item" and "minimum use of unnecessary words." 2- I still rate this as a 5!
• Information is organized in a manner that is consistent with an academic English framework with a left-right, top-bottom flow	444444445555	**4.33**	

In addition to the other considerations, a computer-based test should have these considerations:			
Layout and design			
• Sufficient contrast between background and text and graphics for easy readability	444455555555	**4.67**	1-“Sufficient luminance contrast between…”
• Color is not relied on to convey important information or distinctions	233334455555	**3.92**	1-“Color alone is not relied…”
• Font size and color scheme can be easily modified (through browser settings, style sheets, or on-screen options)	233344555555	**4.08**	
• Stimulus and response options are viewable on one screen when possible	344555555555	**4.67**	
• Page layout is consistent throughout the test	444555555555	**4.75**	
Navigation			
• Navigation is clear and intuitive; it makes sense and is easy to figure out	455555555555	**4.92**	
• Navigation and response selection is possible by mouse click or keyboard	344555555555	**4.67**	1-“Navigation and response selection is possible by mouse (or equivalent) or keyboard (or equivalent).”
• Option to return to items and return to place in test after breaks	3445555555	**4.6**	1-As before, returning to items will only work in non-adaptive tests.
Screen reader considerations			
• Item is intelligible when read by a text/ screen reader	344455555555	**4.58**	
• Links make sense when read out of visual context. (“go to the next question” rather than “click here”)	444455555555	**4.67**	
• Non-text elements have a text equivalent or description	3334555555	**4.3**	
• Tables are only used to contain data, and make sense when read by screen reader	34444455555	**4.36**	

Test specific options • Access to other functions is restricted (e.g., e-mail, Internet, instant messaging) • Pop up translations and definitions of key words/phrases are available if appropriate to the test	33455555555 333444445555	**4.55** **4.08**	1-"Access to functions other than assistive technologies and supports is restricted (e.g., e-mail, Internet, instant messaging)."
Computer capabilities • Adjustable volume	334455555555	**4.5**	1-"Adjustable volume and rate of voice."
• Speech recognition available (to convert user's speech to text)	112334555555	**3.67**	
• Test is compatible with current screen reader software	334444455555	**4.25**	
Below are new considerations suggested by panelists on the first survey, please rank each consideration and comment if you wish (numeric rankings found below were provided by participants in last survey).			
Computer interfaces follow Section 508 guidelines (www.section508.gov)	??013445555		1-(No number was given—only 3 question marks.) 2-There are other more recent and specific computer accessibility guidelines available, such as those found at www.aph.org, Microsoft, WCAG, etc. This is important but not a consideration for the item review process. Include in general guidelines for CBT. 3-(No number was given—only 5 question marks.) 4-Consideration needs more elaboration. 5-Not sure this can be standardized across vendors, platforms, and new technological advances. 6-Won't be clear enough to most readers since so many considerations are in 508. Just make sure key issues are included in the other items. 7-I don't even know what these guidelines are, but if they are designed to promote maximum accessibility, I would think that would be important. 8-Unpack 508 guidelines

Students have received adequate training on use of test delivery system	0445555555555		1-Sticky consideration as you do not want students spending an over necessary amount of time on test prep. 2-Yes, this is important. 3-Absolutely essential. These devices should be tied to the learners' instruction and not unique to the assessment situation. 4-This is important but not a consideration for this item review process. Include in general guidelines for CBT. 5-Absolutely. 6-Students need to have critical keyboard and navigation skills.
Students writing online can get feedback on length of writing on-demand in cases where there is a restriction on number of words.	?012222333455		1-Writing on-line in and of itself is questionable for young students and those who have not been exposed to this technology. 2-Not sure about this. Is this too much guidance. Shouldn't they be able to follow directions. 3-(This went with the question mark.) Yes, if it is a tool that is familiar to the student from their instruction. 4-This is important but not a consideration for this item review process. Include in general guidelines for CBT. 5-If it is truly a writing test and students are taught about writing for particular purpose, length of composition should be part of that which is measured 6-Is this ever an issue? Are students ever penalized for too many words?
Students are able to record their responses and read them back (or have them read-back using text-to-speech) as alternative to human scribe, but only if student has experience with this mode of expression and chooses it for the test.	0233444445555		1-If feasible and part of the student's IEP. 2-Yes, good idea. 3-This is important but not a consideration for this item review process. Include in general guidelines for CBT. 4-I think this would be preferable to a human scribe anyhow. 5-This accommodation can lead to considerable confusion on the part of the student.
Students are allowed to create persistent marks to the extent that they are already allowed to on paper-based booklets (e.g., marking items for review; eliminating multiple choice items, etc.).	033455555555		1-As long as it doesn't impact scoring. 2-Good idea. 3-Is this a cross response platform statement? 4-This is important but not a consideration for this item review process. Include in general guidelines for CBT. 5-Until we do something with computer technology other than simply displaying a paper-and-pencil test item on screen we need to mimic that mode of administration. 6-If you mean "permanent marks" why not just say that? In any case, I think this is critical. 7-Why not?

Alternate versions of computer interface provided that is amenable to use with screen readers (e.g., JAWS, Window-Eyes).	?04444555		1-(Wrote 4 question marks.) 2-This is important but not a consideration for this item review process. Include in general guidelines for CBT. 3-As long as the different interfaces are judged equally fair. 4-Versions of interface (browsers) may limit this.
When "modifying" items for use under a UD framework, there is convincing evidence that the item still measures the same or similar intended construct.	44444445555		1-Obviously, since I wrote these, I think they are all important. I didn't rate them as "5" because I think they are very hard to do, but I also think we need to be working in this direction. 2-What is "convincing?" 3-(Response to "obviously…" comment: Yes. I Agree. 4-I agree. Very important, but we may need more research before including this as a consideration in the review process (i.e., publishers are currently researching issue of "equivalency" between paper and pencil and CBT. 5-If it is the same content, skill, or ability, then we should be fine… similar is not enough to maintain content validity and equivalence to prior test developed in a state. However, if we are redesigning a test then similar is most likely fine.
A test constructed under a UD framework allows for the measurement of the same depth of knowledge levels as the "original" test.	011444444455		1-Poorly stated. Addressed in more useful, generic terms elsewhere. 2-I don't understand why there would be 2 versions. Either the test is UD or it isn't. 3-This should be obvious. The mode of delivery and response should not interfere with the constructions being measured. 4-This is important but not a consideration for this item review process. Include in general guidelines for CBT. 5-I think this is still unknown at the moment but it is an empirical question that we could investigate.
A UD test is aligned to the standards to the same extent as the "original" test.	014444455555		1-Addressed in more useful, generic terms elsewhere. 2-See above. (Above is the statement: This should be obvious. The mode of delivery and response should not interfere with the constructions being measured.—not sure to what it's referring.) 3-This is important but not a consideration for this item review process. Include in general guidelines for accessible assessments. 4-Or to the construct… 5-Yes, but see above.

Test items are piloted, field tested, and normed on all subgroups for which the measure is designed.	033445555555		1-Agree, but this is very hard to do, due to N-count on many subgroups is too small. 2-Bravo! This is crucial but not a consideration for this item review process. Include in general guidelines for creating all assessments. 3-A valid test of the content is valid for all subgroups. The test is a measure of the content, not the subgroup. 4-There isn't a norming process in most state accountability tests. Including or not including various groups in the item analyses may or may not have an impact on those analyses depending on the sample size used for analysis (e.g., 1,000; 2,500; etc. students per item) and the proportion of the population each group represents. 5-A must! 6-Norms may not be important.
Booklets/materials can be easily handled with limited motor coordination.	04445555		1-Yes. 2-This is important but not a consideration for this item review process. Include in general guidelines for accessible assessments. 3-I'm not sure how this would work in practice. 4-Why not?
Response formats are easily correlated to question.	0144555		1-Explain. 2-Unclear. 3-Quite frankly, I don't understand what this one means. 4-Do you mean, "response formats are appropriate to the specific question?" If so, I think you should use my wording. 5-Not sure what this means.
Computer-based option to mask items or text (e.g., split screen).	0123444444		1- Too specific. 2-Don't understand. 3- This is important but not a consideration for this item review process. Include in general guidelines for CBT. 4- Yes, if it is important for students. 5- Why not?
Place for student to take notes (on the screen for CBT) or extra white space with paper-pencil.	03444444555		1-Good idea. 2-The CBT should as much as possible afford the strategies that one may use on a traditional looking format. 3-This is important but not a consideration for this item review process. Include in general guidelines for CBT. 4-Yes, very important. 5-All processing strategies should be made available.

Computer software for test delivery is designed to be amenable to all assistive technology.	02444455555		1-“Amenable” too open-ended. 2-Yes, but hard to do. 3-Ideal, may not be very realistic. 4-This is important but not a consideration for this item review process. Include in general guidelines for CBT. 5-This may be one of those moving targets that will be difficult to maintain due to technology changes. 6-That is, within reason for each disability. 7-Yes, otherwise it should not be used. 8-Very difficult to achieve and more difficult to maintain.

Items on this form are based on information presented in Thompson, Johnstone, & Thurlow (2002, *Universal Design Applied to Large Scale Assessments*, Synthesis Report 44); Thompson & Thurlow 2002, Universally *Designed Assessments: Better Tests for Everyone!*, Policy Directions 14), and Kopriva (2002, *Ensuring Accuracy in Testing for English Language Learners*, CCSSO SCASS-LEP Consortium), as well from NCEO staff brainstorming and input received from participants in the Universal Design Pre-conference Clinic at the CCSSO Large Scale Assessment and Accountability Conference in San Antonio, Texas, June 2003 and input from a joint project/Delphi review with the Minnesota, Nevada, and South Carolina Departments of Education.

Appendix D

Revised Considerations Based on Delphi Results

Considerations for Universally Designed Assessment Items

These guidelines contain suggestions that test developers, item reviewers, and others working on the development of assessments should consider at the beginning stages of designing an assessment that is accessible to the widest range of students possible. Unless stated otherwise, all considerations apply to both paper/pencil and computer-based assessments.

Considerations when reviewing any test item:
Does the item…
Measure what it intends to measure • Reflect the intended content standards (reviewers have information about the content being measured). • Minimize knowledge and skills required beyond what is intended for measurement. **Notes:** a. Content area assessments must be aligned to grade level state academic content standards (grade level expectations). b. Information about the content standard(s) assessed by each item must be supplied to reviewers. c. Careful consideration of the way content standards are phrased is important in determining what knowledge and skills involved in responding to an item are extraneous and which are relevant to what is being tested. d. When considering what is being measured there is somewhat of a "balancing act." In certain types of test items additional skills may be necessary. For example, responses to a listening test must be spoken or written, requiring skills in at least one modality in addition to listening. A similar issue is presented on math tests that require skills in reading. • While it is important to minimize knowledge and skills beyond what is intended for measurement, it cannot take precedence over the ability to measure all content areas to be assessed (e.g., drawing a graph of the results). • When measuring skills such as higher-order processing skills, it is difficult to focus the assessment items only on the explicitly targeted content areas.

Does the item…

Respect the diversity of the assessment population

- Sensitive to test taker characteristics and experiences (consider gender, age, ethnicity, socio-economic level, region, disability, and language)
- Avoid content that might unfairly advantage or disadvantage any student subgroup

Notes:

a. Avoid bias toward or against any group of students that may cause them to have difficulty responding to items or create emotional stress.
b. Carefully evaluate what assumptions items make about shared experiences.
c. Tests should strive to avoid content that negatively depicts any student subgroup and avoid content that potentially provokes a negative reaction in any student subgroup.
d. Gender, etc., should not be a barrier to understanding the task an item requires.
e. It is important to recognize that every test is biased, although universal design serves to minimize its impact. For example, English language learners are a subset of language minority students.
f. In an effort to make assessments more accessible, item writers and developers need to guard against stripping down assessments too much.

Does the item…

Have a clear format for text

- Standard typeface.
- Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (type size appropriate for age group), and adaptable font size for computers.
- High contrast between color of text and background.
- Sufficient blank space (leading) between lines of text.
- Staggered right margins (no right justification).

Notes:

a. Use clear, common, familiar, and consistent fonts (e.g., fonts with wide spacing between letters, words, and lines such as Times or Arial).
b. Avoid decorations and flourishes.
c. The term “blank space” rather than “white space” may be more accurate because the background is not always white.
d. A student should never be confused over type face.
e. Twelve-point varies with font types, as does spacing between letters.
f. Some readers may be unable to see more typical print sizes clearly (e.g., 12 point).
g. When selecting color in text or background, consider high print density contrast in order to avoid isoluminance (i.e., colors appearing equivalent for student with color blindness).

Does the item…

Have clear visuals (when essential to item)

- Visuals are needed to answer the question.
- Visuals with clearly defined features (minimum use of gray scale and shading).
- Sufficient contrast between colors.
- Color alone is not relied on to convey important information or distinctions
- Visuals are labeled.

Notes:

a. Pictures should have a purpose other than simply to be decorative.
b. Weigh whether the use of a visual (e.g., illustration) helps students or interferes with the content being tested. This is a judgment call, but should be carefully considered.

c. Labeling pictures is helpful, when possible. This is true even if the picture seems obvious.
d. Give additional clues besides color when possible (e.g., use text label “stop” and “go” next to stop light images).
e. Visuals should not distract or affect students who do not need the visual cue; rather they should add value for students who do need visual cues.
f. Pictures, when used, should support text and provide an additional resource for students to construct meaning.
g. There may be instances where grayscale and shading are appropriate for providing relevant information to students.

Does the item…

Have concise and readable text

- Commonly used words (except vocabulary being tested).
- Vocabulary appropriate for grade level.
- Minimum use of unnecessary words.
- Idioms avoided unless idiomatic speech is being measured.
- Technical terms and abbreviations avoided (or defined) if not related to the content being measured.
- Sentence complexity is appropriate for grade level.
- Question to be answered is clearly identifiable.

Notes:

a. The use of common words depends on the content assessed. For example, if vocabulary is being tested, difficult or uncommon words might be appropriate to include.
b. Use of commonly used words does not assume that the words are simple. They may carry multiple meanings.
c. Commonly used words may be replaced by reference words most frequently appearing in similar text or listed as the primary definition in the dictionary.
d. Some students may know less common words but may not know phrasal words. It is difficult to assume what is uncommon or difficult.
e. Other than the terms relevant to the construct being measured, items should use basic language and vocabulary (as they are different), and might even be one grade level below the grade being tested.
f. If reading is not the primary construct tested, keep reading level at or below grade level in order to minimize construct irrelevant variance.
g. With the exception of subject specific terminology, the text of an item should consist of commonly understood words or vocabulary consistent with each specific grade level.
h. Determination of complexity can include many factors such as use of clauses, use of the passive, number of syllables in a word, length of sentences, length of single passage, combined length of all reading passages, amount of extraneous text involved in non-reading problems, etc. Complexity needs to be considered on all tests.
i. When using authentic texts on reading passages, complexity may be difficult to control, but should at least be considered on test questions.

Does the item…

Allow changes to its format without changing its meaning or difficulty (including visual or memory load)

- Allows for the use of braille or other tactile format.
- Allows for signing to a student.
- Allows for the use of oral presentation to a student.
- Allows for the use of assistive technology.
- Allows for translation into another language.

Notes:

a. Not all items can be brailled or made tactile in a meaningful way. Under such circumstances, items may need to be modified so that students with visual impairments can answer equivalent items testing knowledge of the same content.
b. Validity may be compromised when critical and relevant words/concepts cannot be translated into different languages.

Overall test considerations:

Does the test…

Have an overall appearance that is clean and organized

- All visuals (e.g., images, pictures) and text provide information necessary to respond to the item.
- Information is organized in a manner consistent with an academic English framework with a left-right, top-bottom flow.
- Booklets/materials can be easily handled with limited motor coordination.
- Response formats are easily matched to question.
- Place for student to take notes (on the screen for CBT) or extra white space with paper-pencil.

Notes:

a. Images, pictures, and text that may not be necessary include sidebars, overlays, callout boxes, visual crowding, shading, and general busyness—anything that may distract a student.
b. Carefully consider whether students from some groups may misinterpret the flow of text and graphics based on characteristics of their native language or culture. Left-right and top-bottom flow is cultural. For example, in some languages, text may flow top to bottom.
c. When using “authentic visuals” (e.g., a map), carefully consider what is being tested in addition to the intended content.
d. If a test has a time limit, careful consideration should be given to why a time limit is necessary for the content being tested.
e. Readability indices are now found in major word processors, however, it is important to check to see they are working as intended.
f. Check to see which tests allow oral presentation.
g. Involve members of the major language groups in item review committees.
h. Consideration of these issues prior to administration of a test will also help with the administration of oral interpretations in the native language, if allowed on a content test. This issue is also relevant for sign language interpretations of tests, when appropriate.
i. There are many ways to translate content area assessments, such as side-by-side, or developing parallel forms. Carefully consider the plusses and minuses of each way prior to making a decision about your state test. There is no perfect solution.
j. Test items are piloted, field tested, and normed on all subgroups for which the measure is designed.
k. For computerized tests, students would need to have critical keyboard and navigation skills.

In addition to the other considerations, a computer-based test should have these considerations:

Layout and design

- Sufficient contrast between background and text and graphics for easy readability.
- Color alone is not relied on to convey important information or distinctions.
- Font size and color scheme can be easily modified (through browser settings, style sheets, or on-screen options).
- Stimulus and response options are viewable on one screen when possible.
- Page layout is consistent throughout the test.
- Computer interfaces follow Section 508 guidelines (www.section508.gov).

Notes:

a. The design of the color of the test needs to take into account test takers with color blindness including red/green distinctions.
b. The electronic format needs to be accessible through the specific assistive technology the student has experience using during the testing. The latest technology may not be what is used in the schools.
c. More recent and specific computer accessibility guidelines are available at www.aph.org, Microsoft, WCAG etc.

Navigation

- Students have received adequate training on use of test delivery system.
- Navigation is clear and intuitive; it makes sense and is easy to figure out.
- Navigation and response selection is possible by mouse click or keyboard.
- Option to return to items and return to place in test after breaks.

Notes:

a. How to navigate and navigation symbols should be intuitive and/or explained at the beginning of a test.
b. The screen resolution varies on different computers. Reviewers should check out items on different types of computers commonly used in schools.
c. Schools need reasonable minimum standards for computer and audio requirements for a test.
d. Test administration instructions should include standardized settings for the computer.
e. Students need practice opportunities before taking computer-based tests.
f. Some listening tests may want to limit the number of times a student can listen to a recording, depending on standards being tested.

Screen reader considerations

- Item is intelligible when read by a text/screen reader.
- Links make sense when read out of visual context ("go to the next question" rather than "click here").
- Non-text elements have a text equivalent or description.
- Tables are only used to contain data, and make sense when read by screen reader.

Notes:

a. Images and animations have text labels, *if* this does not supply the answer.
b. Captioning and transcripts of audio and video are available.
c. Provide titles and summaries for tables and graphs.
d. Header cells for columns and/or rows are designated.
e. Information in tables makes sense when linearized (i.e., read top left to bottom right cell).
f. Current screen reader technology might be difficult for ELLs to understand, real voice technology may be needed.

Test specific options

- Access to other functions is restricted (e.g., e-mail, Internet, instant messaging).
- Pop up translations and definitions of key words/phrases are available if appropriate to the test.
- Students writing online can get feedback on length of writing on-demand in cases where there is a restriction on number of words.
- Students are able to record their responses and read them back as an alternative to a human scribe.
- Students are allowed to create persistent marks to the extent that they are already allowed to on paper-based booklets (e.g., marking items for review; eliminating multiple choice items, etc.).

Notes:

a. Access to spell check might also be limited depending on the test.
b. Variable audio speed might be useful to some students if it does not interfere with the standard being tested.
c. The option for feedback on demand on the length of student writing would depend on the extent that keeping text length within some parameter is part of the construct being measured.

Computer capabilities

- Adjustable volume.
- Speech recognition available (to convert user's speech to text).
- Test is compatible with current screen reader software.
- Computer-based option to mask items or text (e.g., split screen).
- Computer software for test delivery is designed to be amenable to assistive technology.

Notes:

a. Alternate versions of computer interface provided that is amenable to use with screen readers (e.g., JAWS, Window-Eyes).

Appendix E

Supporting Statements by Researchers

Measures what it intends to measure

- Test development begins with a careful consideration of the skills proposed for measurement (Popham & Lindheim, 1980).
- Every item should reflect specified content and mental behaviors, as called for in test specifications (Haladyna, Downing, & Rodriguez, 2002).
- Removal of construct irrelevant variance increases tests scores for students with reading difficulties (Calhoun, Fuchs & Hamlett, 2000; Harker & Feldt, 1993; Koretz, 1997; Tindal, Heath, Hollenbeck, Almond & Harniss, 1998).
- Language in non-language arts assessments needs to be "transparent" enough to students to clearly determine construct being measured (Sharrocks-Taylor & Hargreaves, 1999).

Respects the diversity of the assessment population

- Items must be reviewed for bias that may exist against particular populations (National Research Council, 1999).
- Items that are designed from the start with equity and accessibility features are less likely to be biased against particular populations (Kopriva, 2000).
- Items must be free of content that makes a student's socioeconomic status or inherited academic aptitudes the dominant influence on how a student will respond to the item (Popham, 2001).
- Items must be free of content that may unfairly benefit or penalize students from diverse ethnic, socioeconomic, or linguistic backgrounds, or students with disabilities (Popham, 2001).
- Cultural norms, beliefs, and customs need to be respectfully reflected in illustrations (Schiffman, 1995).

Has a clear format for text

- The point sizes most often used are 10 and 12 point for documents to be read by people with excellent vision reading in good light (Gaster & Clark, 1995).
- Fourteen point type increases readability and can increase test scores for both students with and without disabilities, compared to 12-point type (Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch, 2000).
- Type size for captions, footnotes, keys, and legends needs to be at least 12 point (Arditi, 1999).
- Larger type sizes are most effective for young students who are learning to read and for students with visual difficulties (Hoener, Salend, & Kay, 1997).
- Large print is beneficial for reducing eye fatigue (Arditi, 1999).
- Shapes of letters and numbers should enable people to read text "quickly, effortlessly, and with understanding" (Schriver, 1997).
- The relationship between readability and point size is also dependent on the typeface used (Gaster & Clark, 1995; Worden, 1991).
- Letters that are too close together are difficult for partially sighted readers. Spacing needs to be wide between both letters and words (Gaster & Clark, 1995).
- Fixed-space fonts seem to be more legible for some readers than proportional-spaced fonts (Gaster & Clark, 1995).
- Leading should be 25–30 percent of the point (font) size for maximum readability (Arditi, 1999).
- Leading alone does not make a difference in readability as much as the interaction between point size, leading, and line length (Worden, 1991).

- Standard typeface, upper and lower case, is more readable than italic, slanted, small caps, or all caps (Tinker, 1963).
- Text printed completely in capital letters is less legible than text printed completely in lower-case, or normal mixed-case text (Carter, Dey, & Meggs, 1985)
- Italic is far less legible and is read considerably more slowly than regular lower case (Worden, 1991).
- Boldface is more visible than lower case if a change from the norm is needed (Hartley, 1985)
- Staggered right margins are easier to see and scan than uniform or block style right justified margins (Arditi, 1999; Grise, Beattie, & Algozzine, 1982; Menlove & Hammond, 1998).
- Justified text is more difficult to read than unjustified text—especially for poor readers (Gregory & Poulton, 1970; Zachrisson, 1965).
- Justified text is also more disruptive for good readers (Muncer, Gorman, Gorman, & Bibel, 1986).
- A flush left/ragged right margin is the most effective format for text memory. (Thompson, 1991).
- Unjustified text may be easier for poorer readers to understand because the uneven eye movements created in justified text can interrupt reading (Gregory & Poulton, 1970; Hartley, 1985; Muncer, Gorman, Gorman, & Bibel, 1986; Schriver, 1997).
- Justified lines require the distances between words to be varied. In very narrow columns, not only are there extra wide spaces between words, but also between letters within the words (Gregory & Poulton, 1970).
- Longer lines, in general, require larger type and more leading (Schriver, 1997).
- Optimal length is 24 picas—about 4 inches (Worden, 1991).
- Lines that are too long make readers weary and may also cause difficulty in locating the beginning of the next line, causing readers to lose their place (Schriver, 1997; Tinker, 1963).
- Lines of text should be about 40–70 characters, or roughly eight to twelve words per line (Heines, 1984; Osborne, 2001; Schriver, 1997).
- Blank space anchors text on the paper and helps increase legibility (Menlove & Hammond, 1998; Smith & McCombs, 1971).
- A general rule is to allow text to occupy only about half of a page. Too many test items per page can make items difficult to read (Tinker, 1963).

Has clear pictures and graphics (when essential to item)

- Graphics with a clear sense of unity, a clear focal point, and balance reduce the cognitive load of perceiving information and computer-based tests should allow students to change the size of the font (see computer specific considerations below) and thus increase speed with which the user can access graphic material (Szabo and Kanuka, 1998).
- If illustrations are present they are at best essential information, good if they support the information, and unnecessary if they are unrelated to the construct or item (Sharrocks-Taylor & Hargreaves, 1999).
- Illustrations should be placed directly next to the information for which they refer (Silver, 1994; West, 1997).
- Placing labels directly on plot lines of graphs reduces the load on short-term memory (Gregory & Poulton, 1970).
- Quantitative displays should be structured so that readers can easily construct appropriate inferences about the data (Schriver, 1997).
- Illustrations should be placed directly next to the information for which they refer (Silver, 1994; West, 1997).
- Graphs, illustrations, and other graphic aids can facilitate comprehension (Rakow & Gee, 1987)

Has concise and readable text

- General readability principles such as fewer words per sentence and the removal of irrelevant difficult words increases comprehension of items (Popham & Lindheim, 1980; Rakow & Gee, 1987).
- Flow of sentences is also an important feature. Caution should be taken when reducing reading load so that sentences do not become disjointed or incomprehensible (Anderson, Hiebert, Scott, & Wilkinson, 1985).
- Compound sentences can be written in two separate sentences (if sentences are still comprehensible) (Gaster & Clarke, 1995).
- Most important ideas should be stated first in a sentence (Gaster & Clarke, 1995).
- Noun-pronoun relationships should be clear (Gaster & Clarke, 1995).
- Illustrations should be placed close to the text they support (Gaster & Clarke, 1995), or removed if they do not support text.
- Readability increases when students have likely had experiences or prior knowledge relating to items (Rakow & Gee, 1987).
- Content within items is clearly organized (Rakow & Gee, 1987)
- The content of every item should be independent from content of other items on the test (Haladyna et al., 2002)
- Questions are clearly framed (Rakow & Gee, 1987)
- Limit the number of words, difficulty of words (Popham & Lindheim, 1980), and grammatical complexity of test materials (Popham & Lindheim, 1980)
- Keep vocabulary simple for the group of students being tested (Haladyna et al., 2000).
- Minimize the amount of reading in each item (Haladyna et al., 2002).
- Avoid window dressing (excessive verbiage; Haladyna et al., 2002).
- Simple, clear, commonly used words should be used whenever possible (Gaster & Clarke, 1995).
- Technical terms should be defined (Gaster & Clarke, 1995).
- One idea, fact, or process should be introduced at a time, then ideas developed logically (Gaster & Clarke, 1995).
- If time and setting are important to the sentence, they should be placed at the beginning of the sentence (Gaster & Clarke, 1995).
- Sequence steps of instructions in the exact order that they will be needed (Gaster & Clarke, 1995).
- Vocabulary should be grade-level appropriate (Rakow & Gee, 1987).
- Sentence complexity must be appropriate for grade level (Rakow & Gee, 1987).
- Definitions and examples must be clear and understandable (Rakow & Gee, 1987).
- Required reading skills are appropriate for students' cognitive level (Rakow & Gee, 1987).
- Use of plain language: "text-based language that is straightforward, concise, and uses everyday words to convey meaning. The goal of plain language editing strategies is to improve the comprehensibility of written text while preserving the essence of its message." (Hanson, Hayes, Schriver, LeMahieu, & Brown, 1998, p.2).
- Reduce the verbal and organizational complexity of test items while preserving their essential content (i.e., the skills and concepts they were intended to measure.) (Hanson et al, 1998, p.2).
- Reduce excessive length; reduce wordiness and remove irrelevant material (Brown, 1999).
- Eliminate unusual or low frequency words and replace with common words (e.g., replace "utilize" with "use") (Brown, 1999).
- Avoid ambiguous words (e.g., crane) (Brown, 1999).
- Avoid irregularly spelled words (e.g., trough, feign) (Brown, 1999).
- Avoid proper names, replace with common names or no names at all (Brown, 1999).

Allows changes to its format without changing its meaning or difficulty (including visual or memory load) • Construct irrelevant graphs, vertical text, untranslatable material, and decorative graphics all create situations where accommodating students who use braille, American Sign Language, or non-English languages is difficult.
Additional considerations for computer-based assessments • Students reported difficulties with computers including excessive need for forward and back buttons, unfamiliarity with response mechanisms, and an inability to see entire problems on screens (Trotter, 2001). • Students may not be familiar with skills like scrolling or using text on multiple screens (Cole, Tindal, & Glasgow, 2000). • Some students have had little access to computers and calculators prior to testing (Bridgeman, Harvey, & Braswell, 1995; MacArthur & Graham, 1987).

Appendix F

Item Review Checklist

Considerations for Universally Designed Assessment Items

Subject _____ Item Number _____ Reviewer initials _____

Grade _____ Test form _____

Star (★) areas of strength and Check (✓) areas of concern for each Does the item…	Item #									
Measure what it intends to measure • Reflect the intended content standards (reviewers have information about the content being measured) • Minimize knowledge and skills required beyond what is intended for measurement										
Respect the diversity of the assessment population • Sensitive to test taker characteristics and experiences (consider gender, age, ethnicity, socio-economic level, region, disability and language) • Avoid content that might unfairly advantage or disadvantage any student subgroup										
Have a clear format for text • Standard typeface • Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (Type size appropriate for age group) • High contrast between color of text and background • Sufficient blank space (leading) between lines of text • Staggered right margins (no right justification)										

Have clear visuals (when essential to item) • Visuals are needed to answer the question • Visuals with clearly defined features (minimum use of gray scale and shading) • Sufficient contrast between colors • Color alone is not relied on to convey important information or distinctions • Visuals are labeled										
Does the item…	**Item #**									
Have concise and readable text • Commonly used words (except vocabulary being tested) • Vocabulary appropriate for grade level • Minimum use of unnecessary words • Idioms avoided unless idiomatic speech is being measured • Technical terms and abbreviations avoided (or defined) if not related to the content being measured • Sentence complexity is appropriate for grade level • Question to be answered is clearly identifiable										
Allow changes to its format without changing its meaning or difficulty (including visual or memory load) • Allows for the use of braille or other tactile format • Allows for signing to a student • Allows for the use of oral presentation to a student • Allows for the use of assistive technology • Allows for translation into another language										
Have an overall appearance that is clean and organized • All visuals (e.g., images, pictures) and text provide information necessary to respond to the item • Information is organized in a manner consistent with an academic English framework with a left-right, top-bottom flow • Booklets/materials can be easily handled with limited motor coordination • Response formats are easily matched to question • Place for student to take notes (on the screen for CBT) or extra white space with paper-pencil										

Appendix G

Item Review Comments Form

Considerations for Universally Designed Assessment Items

Subject _____ Item Number _____ Reviewer initials _____

Grade _____ Test form _____

Does the item…	Areas of Strength	Areas of Concern	Suggestions for Improvement
Measure what it intends to measure • Reflect the intended content standards (reviewers have information about the content being measured) • Minimize knowledge and skills required beyond what is intended for measurement			
Respect the diversity of the assessment population • Sensitive to test taker characteristics and experiences (consider age, gender, ethnicity, socio-economic level, region, disability and language) • Avoid content that might unfairly advantage or disadvantage any student subgroup			
Have clear format for text • Standard typeface • Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (type size appropriate for age group) • High contrast between color of text and background • Sufficient blank space (leading) between lines of text • Staggered right margins (no right justification)			

Have clear visuals (when essential to item) • Visuals are needed to answer the question • Visuals with clearly defined features (minimum use of gray scale and shading) • Sufficient contrast between colors • Color alone is not relied on to convey important information or distinctions • Visuals are labeled			
Have concise and readable text • Commonly used words (except vocabulary being tested) • Vocabulary appropriate for grade level • Minimum use of unnecessary words • Idioms avoided unless idiomatic speech is being measured • Technical terms and abbreviations avoided (or defined) if not related to the content being measured • Sentence complexity is appropriate for grade level • Question to be answered is clearly identifiable			
Allow changes to its format without changing its meaning or difficulty (including visual or memory load) • Allows for the use of braille or other tactile format • Allows for signing to a student • Allows for the use of oral presentation to a student • Allows for the use of assistive technology • Allows for translation into another language			
Does the test…	**Areas of Strength**	**Areas of Concern**	**Suggestions for Improvement**
Have an overall appearance that is clean and organized • All visuals (e.g., images, pictures) and text provide information necessary to respond to the item • Information is organized in a manner consistent with an academic English framework with a left-right, top-bottom flow • Booklets/materials can be easily handled with limited motor coordination • Response formats are easily matched to question • Place for student to take notes (on the screen for CBT) or extra white space with paper-pencil			