Universal Design in Educational Assessments

William D. Schafer & Min Liu

University of Maryland

More than 20 years ago, the term "Universal design" was proposed by Ron Mace, an architect, who suggested architectural design should be designed to the greatest extent possible for most people (Mace, 1998). The Center for Universal Design at North Carolina State University (Center for Universal Design., 1997) has defined this concept as "the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design." The term was quickly extended into other fields, including environment, recreation and health care. Together with the conceptualization, seven general principles have emerged to guide engineers, architects and environmental designers in their work (e.g., see the NCSU web site, a popular resource for universal design in general). These principles along with associated guidelines, taken from the http://www.design.ncsu.edu/cud/ web site (follow the path from quick links to UD Principles), are:

1. Equitable Use. The design is useful and marketable to people with diverse abilities.

Guidelines:

- Provide the same means of use for all users; identical whenever possible; equivalent when not.
- Avoid segregating or stigmatizing any users.
- Make provisions for privacy, security, and safety equally available to all users.
- Make the design appealing to all users.

2. Flexibility in Use. The design accommodates a wide range of individual preferences and abilities.

Guidelines:

- Provide choice in methods of use.
- Accommodate right- or left-handed access and use.
- Facilitate the user’s accuracy and precision.
- Provide adaptability to the user’s pace.

3. Simple and Intuitive. Use of the design is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level.

Guidelines:

- Eliminate unnecessary complexity.
- Be consistent with user expectations and intuition.
- Accommodate a wide range of literacy and language skills.
- Provide effective prompts and feed-back during and after task completion.

4. Perceptible Information. The design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities.

Guidelines:

- Use different modes (pictorial, verbal, tactile) for redundant presentation of essential information.
- Maximize "legibility" of essential information.
- Differentiate elements in ways that can be described (i.e., make it easy to give instructions or directions).
- Provide compatibility with a variety of techniques or devices used by people with sensory limitations.

5. Tolerance for Error. The design minimizes hazards and the adverse consequences of accidental or unintended actions.

Guidelines:

- Arrange elements to minimize hazards and errors: most used elements, most accessible; hazardous elements eliminated, isolated, or shielded.
- Provide warnings of hazards and errors.
- Provide fail safe features.
- Discourage unconscious action in tasks that require vigilance.

6. Low Physical Effort. The design can be used efficiently and comfortably, and with a minimum of fatigue.

Guidelines:

- Allow user to maintain a neutral body position.
- Use reasonable operating forces.
- Minimize repetitive actions.
- Minimize sustained physical effort.

7. Size and Space for Approach and Use. Appropriate size and space is provided for approach, reach, manipulation, and use regardless of the user's body size, posture, or mobility.

Guidelines:

- Provide a clear line of sight to important elements for any seated or standing user.
- Make reach to all components comfortable for any seated or standing user.
- Accommodate variations in hand and grip size.
- Provide adequate space for use of assistive devices or personal assistance.

Universal design has been used in various educational areas. The Council for Exceptional Children (CEC, 1999) first introduced it into the general education curriculum. Pisha and Coyne (2001) described a project which incorporated universal design element in learning U.S. history. It is also being introduced into instruction in several other educational settings, (Silver, Bourke, & Strehorn, 1998; Scott, McGuire, & Shaw, 2003). The concept has also been applied to educational assessment.

Universal Design and Assessment

As with universal design in architecture, universally designed assessments seek to make assessments accessible to diverse users. Indeed, NCLB of 2001 defined “universally designed assessments” as “designed from the beginning to be valid and accessible with respect to the widest possible range of students, including students with disabilities and students with limited English proficiency”[1] and requires that all state assessments use it. The National Center for Educational Outcomes, NCEO (Thompson, Johnstone, & Thurlow, 2002) has developed seven elements for universally designed assessments:

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

[1] See webpage: http://education.umn.edu/nceo/TopicAreas/UnivDesign/UnivDesign_FAQ.htm

According to Thompson, Thurlow & Malouf (2004), universal design makes it easier to define a good test item and can also make assessment more compatible with accommodations. In support, Johnstone (2003) studied sixth grade students and found that showed that a universally designed test significantly improved students' performance compared with a traditionally designed test while holding constructs constant.

Besides these advantages of using universal design, three U.S. laws have forced large-scale assessments to become more inclusive. In addition to the NCLB requirement mentioned above, the Assistive Technology Act of 2004 (ATA) and the Individuals with Disabilities Education Act of 2004 (IDEA) both require ”designing and delivering products and services that are usable by people with the widest possible range of functional capabilities”[2].

Many states, school districts, and test companies consider universal design as the best way to realize this goal. The NCEO's State Special Education Outcomes reports in 2003 (Thompson & Thurlow, 2003) and 2005 (Thompson, Johnstone, Thurlow, & Altman, 2005) document that universal design has been given much more attention at the state level. For example, the 2005 report indicates that 45 states address universal design in some way (though in different phases of test development, e.g.

[2] See webpage: http://education.umn.edu/nceo/TopicAreas/UnivDesign/UnivDesign_FAQ.htm

item development, item review, requests for proposals).

Applying Universal Design to Assessments

Although the seven elements of universal design noted above could guide states and test companies in designing large-scale assessments, not all principles apply in obvious ways. There is still a need for specific information in test development to make universal design more concrete as they are applied to assessments. Six phases have been differentiated and several steps within each phase have been proposed[3]. Here, each step is cited (in bold; taken directly from the web site), and followed by a discussion of possible implications for MSDE.

Phase A. Test conceptualization:

1. **Define the construct(s) to be measured precisely and explicitly so the test can be designed to measure the construct while minimizing the effects of irrelevant factors.** For MSDE, the definition of the construct to be assessed using each test is the statement of core learning goals and within those expectations as elaborated by assessment limits. They need to be judged as to whether they are indeed a precise and explicit definition of the construct. This is an important step since the validity of each

[3] Retrieved from http://education.umn.edu/nceo/TopicAreas/UnivDesign/UnivDesign_FAQ.htm; with comments

form of the assessment derives from its match to the statement of the construct and if the construct is not clearly defined, the ability to reach a judgment of match, and therefore of validity, is compromised. Once determined, the definition of the construct is static and different forms are constructed according to the test map to assess it. Therefore, the judgment about clarity of construct definition needs to be reached only once.

2. **Include the full range of students in the definition of the target population.** For MSDE's tests, the target population needs to be defined as broadly as possible. But what dimensions need to be addressed in defining the range of students. For the purpose of universal design, dimensions that may be relevant include physical and intellectual limitations, and cultural and other experiential backgrounds. These apply to all tests. It would be helpful for MSDE to have an explicit statement of the range of students that constitute its school population. The purpose of the statement would be to define clearly the scope of differences in students that needs to be addressed in test construction and administration. The statement would be independent of any one test.

Phase B. Test construction:

1. **Develop items that minimize the effects of extraneous factors and that can be used with accommodations as appropriate, (e.g., avoid unnecessary use of graphics that cannot be presented in Braille, use font size and white space appropriate for clarity and focus, avoid unnecessary linguistic complexity when it is not being assessed). It is the construct that must be held constant, not the design features; there are times, for example, when linguistic complexity is appropriate and necessary.** Items should be written to measure the aspect of the construct called for in the test map and nothing else. This principle is fundamental to good item writing practice since the validity of the item depends on its insensitivity to irrelevant student characteristics. In practice, it can be implemented in item writer training. Two elements seem crucial: (1) to make sure item writers focus on the statement of the construct and the assessment limits, and (2) to make sure item writers understand how the statement of the target population can help them determine whether their work can be used appropriately with the full range of students who are the intended examinees. Training for both these elements can be facilitated by the materials developed following the recommendations in Phase A.

2. **Provide for a full range of test performance to avoid ceiling or floor effects.** This principle seems more applicable to tests that are intended to measure well throughout their range than tests that are designed to classify students into achievement levels as accurately as possible. MSDE's tests seem more like the latter. Nevertheless, some purposes of the scores may depend on accuracy throughout the range of ability, such as selection by employers or postsecondary educational agencies. To the extent that these purposes exist, providing a broad range of score accuracy may be desired. In order to reach a judgment about the need to broaden the range of accuracy, a statement about the purposes of MSDE's tests would be helpful. The statement should focus on uses made of the scores (e.g., "to determine the achievement levels of students") as opposed to the development of the scores (e.g., to measure the achievement of students").
3. **Undergo a review of items using tools such as NCEO's "Considerations for Universally Designed Assessments" (in press). By promoting a structured review of items, test companies can determine the design strengths and weaknesses of items before field testing. Determining well-designed items and items that need minor adjustments**

may save time and money over unstructured item reviews that simply eliminate potentially problematic items. The tool was unavailable at the time of this writing. However, the importance of an item review before field testing is well understood by MSDE. The review will be most effective if it includes professionals who are capable of evaluating the possible effects of the full range of the target population on item validity and who are also capable of recommending modifications that will overcome sources of invalidity while maintaining assessment of the aspect of the construct the item is intended for. The statements suggested in Phase A would be helpful in selecting the panel.

Phase C. Test tryout (field testing)

1. **Include a full range of students in the tryout sample (e.g., students with disabilities, students with limited English proficiency, other students with special needs). Because there may be constraints in sampling due to the low numbers of students with specific characteristics, there may be a need to identify over-sampling strategies (e.g., select groups of items for which additional sampling will occur).** MSDE's field testing is done using random subsamples of the entire student

population. This ensures that the full range of students is represented.

2. **Include the use of accommodations during the test tryout.** MSDE routinely includes all accommodations in its tryouts since they are part of the regular tests.

Phase D. Item analysis:

1. **Analyze item characteristics to determine which items can be used with the full range of students and with accommodations.** The next step seems to cover the current one since qualitative analysis has already been accomplished in Phase B, Step 3.
2. **Use a wide range of statistical tests to determine if items are functioning differently for particular populations. Populations of students with particular disabilities or primary language are often small in number, so using multiple analysis techniques will help test designers to see patterns of items to "flag" for further investigation. Examples of statistical techniques can be found in Analyzing Results of Large-scale Assessments to Ensure Universal Design.** The citation suggests that differential item functioning (DIF) tests could be applied to evaluate items for focal groups of

students with disabilities by disability category (e.g., learning disabilities, visual impairment, hearing impairment, emotional disturbance), accommodation group (e.g., read aloud, extended time, large print), or combinations of disability category and accommodation group (e.g., students with learning disabilities who receive read aloud accommodations), noting that the latter could result in sample sizes too small to be of much use. Four types of analyses were exemplified. The first was item rank difference (items were ranked by P values for focal and reference groups and differences of five or more ranks were flagged). The second was item-test correlation (the differences between range-restriction-corrected correlations for the focal and reference groups were compared using an effect size measure and its standard error). The third was DIF by contingency table analysis using raw scores (Mantel-Haenszel or log odds ratio or, when focal group sizes were particularly small, a procedure based on a focal-group size weighted average of the differences in the P values across all raw scores). The fourth was DIF using IRT (e.g., using a program that compares ICC heights for focal and reference groups for 100 scale score values). The procedure used by MSDE for DIF is Mantel-Haenszel (with extensions) and this approach could easily be adapted to study DIF for disability categories or for accommodations. It is recommended that MSDE undertake a study of the feasibility of these analyses as a routine addition to its item-screening process when items are field tested.

Phase E. Analysis of "flagged" items

Conduct cognitive labs (think aloud studies) with a small number of actual students who will take the test. Student data can be used to determine if the design of items set forth by test designers is comprehensible to students. Involving a wide range of students is helpful in gathering particular perspectives.

1. **Information on cognitive labs can be found in forthcoming NCEO reports.** While cognitive labs might be helpful in informing participants in item reviews, this is probably not a cost-effective step for MSDE at this time since it is a relatively expensive method of data collection. Instead, the possibility of using cognitive labs should be held open while representatives of content and disability groups review items that have been flagged. If they feel they need more understanding about why items demonstrate quantitative differences, then some cognitive lab data might be collected for certain of these items.

Phase F. Test revision:

1. **Eliminate items with evidence of disability bias.** As with all DIF analyses, the routine elimination of items based on a statistical procedure is not recommended. Instead, consistent with current practice, a committee study of individual items flagged by any DIF analysis should result in item elimination from the pool only if the committee can explain the differential performance to its satisfaction.
2. **Include the full range of students and the use of accommodations in the test administration.** MSDE already administers its tests to all students and its accommodations policy provides the full range of available accommodations that were used during item field testing.

Discussion

There appear to be two fundamental goals of universal design. First, test items should be stripped of all sources of difficulty that are not elements of the construct the test is intended to assess. In other words, practitioners are asked to avoid invalid sources of difficulty.

Second, tests and their items should be designed so that they are amenable to the full range of accommodations that will be available. A term such as “accommodation-friendly” might be coined for this goal.

Both these goals seem consistent with good assessment practice, in general. Both will remove invalid variation among student scores and allow more straightforward interpretation of those scores as representing academic achievement over the assessments' respective domains.

However, areas of possible conflict exist between assessment of the construct described in the domain of a test and universal design principles. This may arise when portions of the domain are not applicable to the full range of students. For example, it may be unreasonable for deaf students to be asked to learn content that depends on the sounds of words, such as rhyming and alliteration. It is also possible that an accommodation actually alters the construct, as in the use of a reading accommodation for deaf students. Also, it may be difficult to accommodate when certain modes of administration are used, such as when visual simulations need to be accommodated for blind students (e.g., verbal descriptions may or may not be equivalent to the demonstrations). Areas such as these need further exploration before changes in assessment policy to address them can be recommended.

It may be inevitable that some accommodations change the construct, such as the read-aloud accommodation for phonics. It would be possible not to administer those items and to turn them off in scoring, as MSDE does now for the read-aloud accommodation. Proficiency levels can still be determined using the same cuts as with the full population. In score reports, some system is needed to remind users that the IEP accommodations have changed the construct in known ways.

In item writing, it would be possible to tag items or item sets when they have contexts that cannot be administered with certain accommodations. Special paper-and-pencil forms or computer-administered item or item set selection rules could be developed that would avoid administering those items to students who receive the relevant accommodations. Indeed, it may be possible to represent the construct faithfully in assessments using those rules.

An Appendix contains some work that may be helpful in implementation.

Action Recommendations

Universal design is a desirable goal. Its use in most instances improves test validity. Several recommendations have been made in this paper that could guide aspects of future assessment work at MSDE. Comment and recommendations from content, special needs, and psychometric groups within MSDE should be sought before implementation.

Note: This paper was developed by the authors for the Maryland State Department of Education (MSDE), who funded the project through the Maryland Assessment Research Center for Educational Success (MARCES) at the University of Maryland. The opinions are solely those of the authors and do not necessarily reflect those of MSDE or of MARCES.

References

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

Center for Universal Design. (1997). *What is Universal Design?* North Carolina State University. http://www.design.ncsu.edu/cud/

Council for Exceptional Children. (1999). Universal design: Research connections.

Johnstone, C. J. (2003). *Improving Validity of Large-scale Tests: Universal Design and Student Performance (Technical Report 37).* Minneapolis, MN: University of Minnesota

Johnstone, C. J., Thompson, S. J., Moen, R. E., Bolt, S., & Kato, K. (2005). *Analyzing Results of Large-scale Assessments to Ensure Universal Design (Technical Report 41).* Minneapolis, MN: University of Minnesota

Mace, R. (1998). *A perspective on universal design.* Paper presented at the Designing for the 21st Century: An International Conference on Universal Design.

National Center on Educational Outcomes (2005). http://education.umn.edu/NCEO/OnlinePubs/2005StateReport.htm/

.

Pisha, B., & Coyne, P. (2001). Smart from the start: The promise of universal design for learning. *Remedial and Special Education, 22*(4), 197-203.

Scott, S., McGuire, J., & Shaw, S. (2003). Universal Design for instruction: A new paradigm for adult instruction in postsecondary education. *Remedial and Special Education, 24*(6), 369-379.

Silver, P., Bourke, A., & Strehorn, K. C. (1998). Universal instructional design in higher education: An approach for inclusion. *Equity & Excellence in Education, 31*(2), 47-51.

Thompson, S., Thurlow, M., & Malouf, D. B. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied testing technology* http://www.testpublishers.org/Documents/Creating_Better_Tests%20Final%20Revision%205.15.04.pdf.

Thompson, S. J., Johnstone, C. J., Thurlow, M., & Altman, J. (2005). *2005 State Special Education Outcomes: Steps Forward in a Decade of Change*. Minneapolis, MN: University of Minnesota

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design Applied to Large-Scale Assessment (Synthesis Report 44).* Minneapolis, MN: University of Minnesota

Thompson, S. J., & Thurlow, M. (2003). *2003 State Special Education*

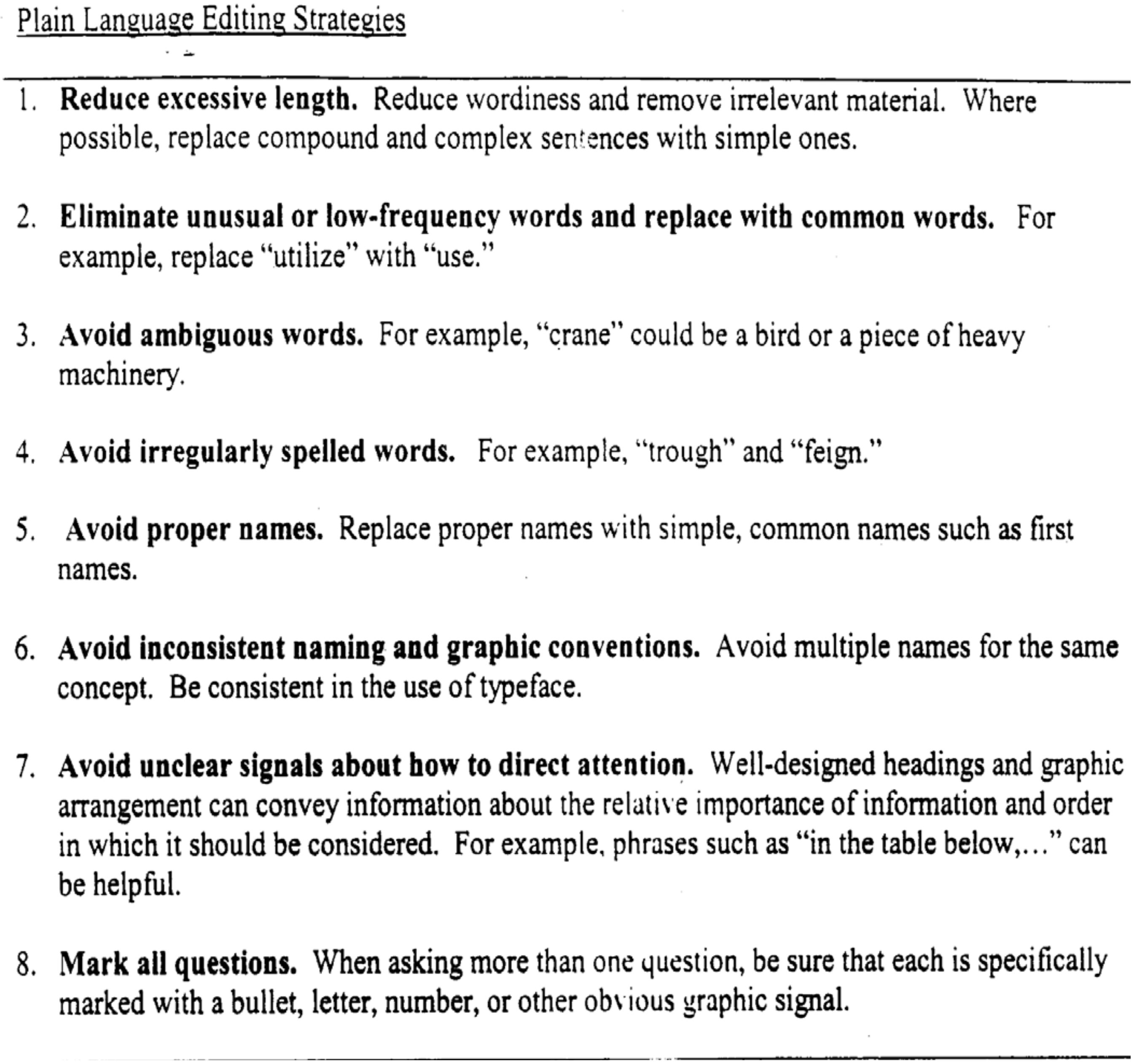
Outcomes:Marching On. Minneapolis, MN: University of Minnesota

Appendix:

Table 1. Plain Language Editing Strategies

Strategy	**Description**
Reduce excessive length	Reduce wordiness and remove irrelevant material.
Use common words	Eliminate unusual or low frequency words and replace with common words (e.g., replace “utilize” with “use”).
Avoid ambiguous words	For example, “crane” should be avoided because it could be a bird or a piece of heavy machinery.
Avoid irregularly spelled words	Examples of irregularly spelled words are “trough” and “feign.”
Avoid proper names	Replace proper names with simple common names such as first names.
Avoid inconsistent naming and graphic conventions	Avoid multiple names for the same concept. Be consistent in the use of typeface.
Avoid unclear signals about how to direct attention	Well-designed heading and graphic arrangement can convey information about the relative importance of information and order in which it should be considered.
Mark all questions	Give an obvious graphic signal (e.g., bullet, letter, number) to indicate separate questions.

Source: Brown (1999).



Plain Language Editing Strategies

1. **Reduce excessive length.** Reduce wordiness and remove irrelevant material. Where possible, replace compound and complex sentences with simple ones.

2. **Eliminate unusual or low-frequency words and replace with common words.** For example, replace "utilize" with "use."

3. **Avoid ambiguous words.** For example, "crane" could be a bird or a piece of heavy machinery.

4. **Avoid irregularly spelled words.** For example, "trough" and "feign."

5. **Avoid proper names.** Replace proper names with simple, common names such as first names.

6. **Avoid inconsistent naming and graphic conventions.** Avoid multiple names for the same concept. Be consistent in the use of typeface.

7. **Avoid unclear signals about how to direct attention.** Well-designed headings and graphic arrangement can convey information about the relative importance of information and order in which it should be considered. For example, phrases such as "in the table below,…" can be helpful.

8. **Mark all questions.** When asking more than one question, be sure that each is specifically marked with a bullet, letter, number, or other obvious graphic signal.

Table 2. Dimensions of Legibility and Characteristics of Maximum Legibility

Dimension	**Maximum Legibility Characteristics**
Contrast	Black type on matte pastel or off-white paper is most favorable for both legibility and eye strain.
Type Size	Large type sizes are most effective for young students who are learning to read, students with visual difficulties, and individuals with eye fatigue issues. The legal size for large print text is 14 point.
Spacing	The amount of space between each character can affect legibility. Spacing needs to be wide between both letters and words. Fixed-space fonts seem to be more legible for some readers than proportional-spaced fonts.
Leading	Leading, the amount of vertical space between lines of type, must be enough to avoid type that looks blurry and has a muddy look. The amount needed varies with type size (for example, 14-point type needs 3-6 points of leading).
Typeface	Standard typeface, using upper and lower case, is more readable than italic, slanted, small caps, or all caps.
Justification	Unjustified text (with staggered right margin) is easier to see and scan than justified text especially for poor readers.

Line Length	Optimal length is about 4 inches or 8 to 10 words per line. This length avoids reader fatigue and difficulty locating the beginning of the next line, which causes readers to lose their place.
Blank Space	A general rule is to allow text to occupy only about half of a page. Blank space anchors text on the paper and increases legibility.
Graphs and Tables	Symbols used on graphs need to be highly discriminable. Labels should be placed directly next to plot lines so that information can be found quickly and not require short-term memory.
Illustrations	When used, an illustration should be directly next to the question for which it is needed. Because illustrations create numerous visual and distraction challenges, and may interfere with the use of some accommodations (such as magnifiers), they should be used only when they contain information being assessed.
Response Formats	Response options should include larger circles (for bubble response tests), as well as multiple other forms of response.

Source: Thompson, Thurlow, and Malouf (2004).